Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)

Christophe Servan, Anne Vilnat

18e Conférence en Recherche d'Information et Applications, 16e Rencontres Jeunes Chercheurs en RI, 30e Conférence sur le Traitement Automatique des Langues Naturelles, 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (CORIA-TALN) [1]

Actes de CORIA-TALN 2023.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN),
volume 4 : articles déjà soumis ou acceptés en conférence internationale

Christophe Servan, Anne Vilnat (Éds.)

Paris, France, 5 au 9 juin 2023

1. https://coria-taln-2023.sciencesconf.org/

Avec le soutien de

Préface

Organisée conjointement par les laboratoires franciliens sous l'égide de l'Association francophone de Recherche d'Information et Applications (ARIA) et l'Association pour le Traitement Automatique des Langues (ATALA), la conférence CORIA-TALN-RJCRI-RECITAL 2023 regroupe :

— la 18ème Conférence en Recherche d'Information et Applications (CORIA)
— la 30ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) ;

ainsi que les deux conférences associées, destinées aux jeunes chercheuses et chercheurs :

— Les 16ème Rencontres Jeunes Chercheurs en RI (RJCRI)
— la 25ème Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)

La conférence TALN (Traitement Automatique des Langues Naturelles) est un rendez-vous annuel qui offre, depuis 1994, le plus important forum d'échange international francophone aux acteurs universitaires et industriels des technologies de la langue. Cet événement, qui accueille habituellement près de 250 participants, couvre toutes les avancées récentes en matière de communication écrite et parlée et de traitement informatique de la langue notamment la recherche et l'extraction d'information, la fouille de textes, le dialogue homme-machine, la fouille d'opinions, la traduction automatique, les systèmes de questions-réponses, le résumé automatique...

Cette année, ont été soumis 51 articles longs et 12 articles courts pour la conférence principale, dont respectivement 29 ont été acceptés pour une présentation orale (dont 2 prises de position) et 9 pour une présentation sous forme de posters. 19 présentations courtes, sous forme de posters, d'articles déjà publiés lors de conférences internationales complètent le programme de la conférence, ainsi que des démonstrations et des présentations de projets en cours. L'alternance de sessions communes entre TALN, CORIA et RJC et de sessions plus spécifiques devraient permettre de susciter des échanges fructueux.

En complément de la conférence principale, se tiennent les ateliers "Défi Fouille de Texte" (DEFT), "Atelier sur l'analyse et la recherche de textes scientifiques" (ARTS), "Humain ou pas humain ? : les nouveaux défis pour les humains" (hOUPSh) et le tutoriel "Apprentissage Profond pour le TAL français pour les debutants" (TutoriAL). Ces ateliers et tutoriel illustrent à la fois des tendances nouvelles présentes dans la communauté et des activités récurrentes.

Un grand merci à toutes celles et tous ceux qui ont soumis leurs travaux, ainsi qu'aux membres du comité de programme et aux relectrices et relecteurs pour le travail qu'ils ont accompli. Ce sont eux qui font vivre la conférence. Merci au comité d'organisation réparti sur la région parisienne, et aux sponsors qui nous ont permis d'organiser cet événement.

Christophe Servan et Anne Vilnat, co-présidents de TALN

Comités

Comité de programme

Présidents

— Christophe SERVAN
— Anne VILNAT

Membres

— Rachel BAWDEN
— Caroline BRUN
— Marie CANDITO
— Rémi CARDON
— Pascal DENIS
— Yannick ESTEVE
— Benoît FAVRE
— Amel FRAISSE
— Thomas GERALD
— Natalia GRABAR
— Lydia-Mai HO-DAC
— José MORENO
— Vassilina NIKOULINA
— Yannick PARMENTIER
— Sylvain POGODALLA
— Solène QUINIOU
— Didier SCHWAB
— Iris TARAVELLA-ESHKOL

Comité d'organisation

— Marie CANDITO
— Thomas GERALD
— José MORENO
— Benjamin PIWOWARSKI
— Christophe SERVAN
— Laure SOULIER
— Anne VILNAT

Table des matières

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Questionner pour expliquer : construction de liens explicites entre documents par la génération automatique de questions

Elie Antoine[1], Hyun Jung Kang[2], Ismaël Rousseau[2], Ghislaine Azémard[3], Frederic Bechet[1], Géraldine Damnati[2]
(1) Aix Marseille Univ, CNRS, LIS, France {first.last}@lis-lab.fr
(2) Orange Innovation, DATA&AI, Lannion {first.last}@orange.com
(3) FMSH/Univ Paris 8 Chaire UNESCO ITEN azemard@msh-paris.fr

RÉSUMÉ

Cette article présente une méthode d'exploration de documents basée sur la création d'un ensemble synthétique de questions et de réponses qui est ensuite utilisé pour établir des liens explicables entre les documents. Nous menons une évaluation quantitative et qualitative des questions automatiquement générées en termes de leur forme et de leur pertinence pour l'exploration de la collection. De plus, nous présentons une étude quantitative des liens obtenus grâce à notre méthode sur une collection de document provenant d'archives numérisés.

ABSTRACT

Expliciting links between documents through automatic question generation

This paper presents a method for document mining based on the creation of a synthetic set of questions and answers. This set is used to establish explainable links between documents. We conduct a quantitative and qualitative evaluation of the automatically generated questions in terms of their form and their relevance to the exploration of the collection. In addition, we present a quantitative study of the links obtained with our method on a collection of documents from a digitized archive.

MOTS-CLÉS : Génération de questions ; Compréhension de documents ; Explicabilité.

KEYWORDS: natural language understanding ; question generation ; explainability.

1 Introduction

Cette étude porte sur l'exploration de collections de documents par des méthodes d'appariements permettant de créer des liens entre des textes de la collection. De nombreuses méthodes basées sur des mesures de similarité permettent de créer de tels liens, cependant la justification des liens pour les utilisateurs est bien souvent réduite à la présence de mots clés en commun. Nous proposons dans cette étude d'utiliser le paradigme des questions/réponses sur des textes afin de créer des liens explicables : deux paragraphes sont liés si les questions que posent ces paragraphes sont également liés.

Cet article est structuré comme suit : la section 2 présente notre méthodologie pour générer des liens explicables entre les documents basés sur des modèles de génération de questions ; la section 3 présente notre méthode de génération et de filtrage de questions ; la section 4 décrit comment les questions générées peuvent être utilisées pour créer des liens explicables entre les documents ; enfin,

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 1–9.

les sections6 et 7 présentent une étude expérimentale sur notre corpus d'archives avec des évaluations quantitatives, descriptives et qualitatives de la méthode proposée.

2 Exploration par la génération de questions

Lors de l'exploration d'une collection d'archives thématiques, des liens peuvent être établis entre des documents ou des parties de documents en fonction de différents critères tels que la co-occurrence d'entités (personnes, lieux, organisations, dates, ...), de mots clés liés à une base de connaissances ou à un thésaurus (Tsatsaronis *et al.*, 2014), ou directement par une mesure de similarité statistique entre des documents ou des parties de documents tels que des phrases (Wang *et al.*, 2016) ou des paragraphes (Dai *et al.*, 2015). La structure en graphe ainsi obtenue peut être utilisée pour concevoir des interfaces de navigation telles que des cartographies ou directement en insérant des liens hypertextes.

Générer des liens basés sur les mots clés/entités pose plusieurs problèmes : d'une part la grande quantité de liens générés si de grands ensembles de mots clés ou d'entités sont considérés et d'autre part le fait que la simple occurrence de termes pertinents ne signifie pas que leurs contextes d'occurrence sont pertinents ou intéressants pour les utilisateurs.

Les liens basés sur la similarité permettent de corriger ce dernier point en prenant en compte les mots dans leur contexte, mais l'utilisation de mesures de similarité statistique rend souvent les liens difficiles à interpréter, obligeant les utilisateurs à les vérifier un par un pour évaluer leur pertinence, ce qui peut être très chronophage.

Récemment, les avancées dans les modèles de Question-Réponse (QR) à partir de texte ont permis l'utilisation de questions directes en langage naturel afin d'accéder à des documents électroniques. Des résultats impressionnants ont été obtenus avec des modèles de langage pré-entraînés tels que BERT sur des corpus de référence comme SQuAD (Rajpurkar *et al.*, 2016). Cependant il a été montré que le type de questions que ces modèles gèrent le mieux sont les questions littérales simples pour lesquelles une réponse factuelle peut être trouvée dans le texte en une seule phrase, et que les performances diminuent lorsque l'on traite de questions plus abstraites ou nécessitant un contexte plus large qu'une phrase pour être abordées. De plus, la plupart de ces études ont été appliquées uniquement au texte de Wikipedia.

Dans notre étude nous n'allons pas utiliser le mode *question→réponse* pour accéder au texte, mais plutôt le mode *réponse→question* : en sélectionnant des zones potentiellement *intéressantes* dans nos corpus nous allons générer des questions à partir du contenu de ces zones (les *réponses*), puis chercher des similarités entre les questions générées pour proposer des liens aux utilisateurs qui soient motivés par la paire de questions mises en relation et qui constitue l'*explication* du lien pour l'utilisateur.

3 Génération de questions

Pour la génération de question, nous utilisons une variation d'un des modèles proposés par (Bechet *et al.*, 2022), où une annotation sémantique (SRL) suivant le formalisme PropBank (Palmer *et al.*, 2005) est effectuée afin de sélectionner des *réponses* potentielles pour générer des questions. Cette sélection par un analyseur sémantique nous permet de ne garder que des contextes potentiellement riches

en terme de sens qui peuvent servir de supports à l'expression d'un lien avec d'autres documents.

La génération de question est vue comme une tâche de génération de texte de type *séquence-à-séquence* avec le modèle *BARThez* (Kamal Eddine *et al.*, 2021) entraîné sur le corpus FQuAD (d'Hoffschmidt *et al.*, 2020) de questions/réponses en français. La séquence en entrée contient la *réponse* (ANS), l'unité lexicale qui déclenche la relation sémantique LU et le contexte CTX, comme dans l'exemple suivant provenant de notre jeu d'entrainement *FQuAD* :
source : `[ANS:ARG2] Héra [LU] appelée[CTX] Cérès fut également appelée Héra en Allemagne pendant une brève période.`
cible : `Quel nom Cérès a-t-elle porté pendant une brève période en Allemagne ?`

Nous appliquons une série de filtres pour améliorer la qualité et réduire la quantité d'exemples générés. La première étape (**F1**) consiste à restreindre l'analyse SRL pour n'inclure que les cadres ayant un déclencheur strictement verbal (en rejetant les verbes auxiliaires), car ceux-ci sont considérés comme étant de meilleure qualité en raison de leur facilité de détection.

Pour améliorer encore la qualité des exemples générés, nous appliquons un filtre (**F2**) sur les requêtes afin d'éliminer celles dont les réponses ou les contextes ne sont pas informatifs. Cela inclut les réponses de moins de 5 caractères, ou appartenant à la liste NLTK (Bird *et al.*, 2009) des mots d'arrêt, afin d'éliminer les réponses contenant uniquement des coréférences pronominales. Les requêtes avec un contexte de moins de 5 mots ou une réponse qui n'est pas située dans la phrase du déclencheur sont également éliminées.

Les questions générées sont également soumises à un filtre (**F3**) basé sur la méthodologie "roundtrip consistency" proposée par (Alberti *et al.*, 2019). Ce filtre consiste à ne conserver que les exemples synthétiques pour lesquels un modèle de question-réponse [1] est capable d'extraire une partie de la réponse cible de la question générée. Nous considérons que le modèle a réussi à récupérer la réponse s'il y a un chevauchement minimum de 30 % entre la réponse prédite et la réponse de la requête.

Enfin, nous appliquons un dernier filtre (**F4**) pour éliminer les questions dupliquées, qui sont un phénomène fréquent dû à de légères variations dans certaines requêtes, résultant souvent en des questions très similaires ou identiques.

4 Générer des liens explicables

L'originalité principale de notre approche est l'utilisation de nos questions/réponses synthétiques pour établir des liens entre les documents de notre corpus. Alors que les méthodes traditionnelles consistent à calculer la similarité via les plongements de documents à un niveau de granularité choisi (phrase, paragraphe ou bloc de texte, page), notre approche consiste à calculer une mesure de similarité sur des plongements obtenus par la concaténation de questions et de réponses produites par notre méthode décrite dans la section 3.

Nous obtenons des structures "`<question> | <réponse>`". Par exemple, voici la structure obtenue sur l'exemple de la question générée donnée dans la section précédente : **`Quel nom Cérès a-t-elle porté pendant une brève période en Allemagne ? | Héra`**

Notre plongement pour chaque paire question-réponse utilise la bibliothèque SentenceTransformer

1. Dans notre cas *CamemBert-large* (Martin *et al.*, 2020) entrainé sur *FQuAD*

(Reimers & Gurevych, 2019) [2]. Une mesure de similarité cosinus est ensuite utilisée entre toutes les paires de ces projections, ce qui donne lieu au calcul d'une matrice de similarité.

Les sections suivantes présentent une étude expérimentale sur un corpus d'archives numérisées provenant du domaine des sciences sociales.

5 Thèmes des questions

Nous définissons les thèmes de la question générée par rapport à un thésaurus spécifique qui a été créé pour le domaine de l'autogestion. À partir de connaissances préalables dans le domaine, une première liste de notions a été construite. Elle a ensuite été enrichie par une liste de mots-clés et de phrases-clés extraites des articles de la revue "Autogestion". Ces termes sont principalement des phrases nominales extraites grâce à une analyse morphosyntaxique des documents. Lorsque des variantes flexionnelles d'une locution sont rencontrées, la forme ayant le plus grand nombre d'occurrences est choisie (forme majoritaire). Parmi toutes les phrases-clés extraites, les experts ont sélectionné une liste d'entrées supplémentaires pour le thésaurus en choisissant des termes qui font référence à des notions générales pouvant être pertinentes pour indexer des documents.

Le thésaurus est ensuite trié hiérarchiquement pour former une structure en arbre de profondeur maximale de quatre niveaux. L'arbre a 437 feuilles et est organisé en huit notions générales à la racine de l'arbre (*Organisations, Classes sociales, Développement économique, Exercice du pouvoir, Juridique, Modèles politiques, Psycho-sociologie, Valeurs sociales*).

$\lvert w \in T \rvert$	0	1	2	3	4	5
$\lvert Q + R \rvert$	43693	25159	8472	2129	355	58

TABLE 1 – Répartition du nombre de mots (w) appartenant au thésaurus T parmi les questions+réponses ($Q + R$)

terme	**occurrences**
travailleurs	3247
travail	2668
pouvoir	2214
société	1947
révolution	1834
production	1742
contrôle	1470
système	1426
ouvrier	1387
mouvement	1362

TABLE 2 – Les dix termes les plus fréquents du thésaurus

Nous avons analysé notre corpus de paires de questions/réponses synthétiques pour étudier l'utilisation des entrées du thésaurus. Nos résultats montrent que **30,6%** des questions générées et **25,1%** des

2. Nous utilisons le modèle multilingue *distiluse-base-multilingual-cased-v1* (Reimers & Gurevych, 2020)

réponses contiennent au moins un terme du thésaurus. De plus, nous avons constaté que **45,3%** des paires question/réponse incluent au moins un mot du thésaurus dans la question ou dans la réponse, et **10,4%** contiennent un mot du thésaurus dans les deux. Une description plus détaillée de la répartition du nombre d'entrées détectées dans les questions et les réponses peut être trouvée dans le Tableau 1 et les 10 entrées les plus fréquentes dans le Tableau 2.

La paire avec le plus de termes du thésaurus est la suivante : `Q : Qu'est ce qui rendra possible le` **`développement`** `de la` **`participation`** `des` **`travailleurs`** `et de leurs` **`organisations`** `à la direction et à la gestion des entreprises nationales ? A : le` **`changement`** `-- en` **`droit`** `et dans les faits -- des formes de la` **`propriété`**

Cette analyse suggère qu'en plus de permettre la création de liens pour explorer la collection, les questions générées pourraient également être un moyen d'illustrer les principales notions abordées dans le journal. Des interfaces dédiées pourraient être développées à cette fin dans le cadre de travaux futurs.

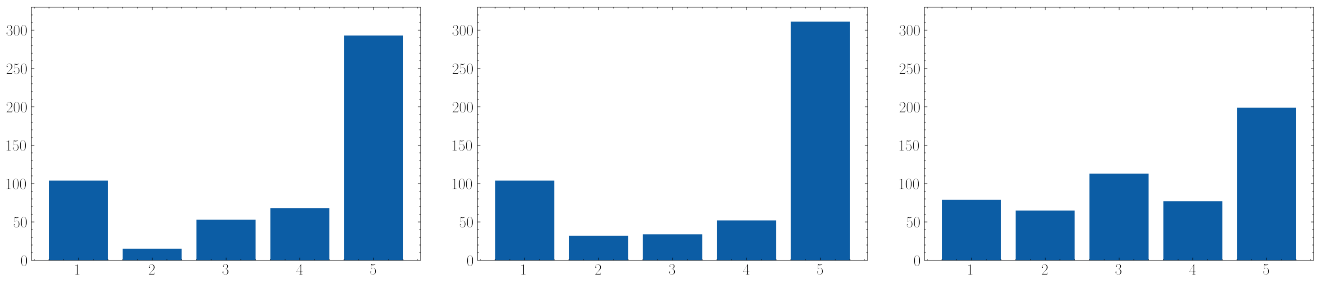
6 Évaluation qualitative des questions générées

Filtre	F1	F2	F3	F4
Nb. questions	247 907	193 685	129 119	79 869

TABLE 3 – Résumé des différents filtres : F2 (suppression des réponses non informatives), F3 (*round-trip consistency*), F4 (suppression des doublons).

Nous avons appliqué notre méthode de génération de questions à un corpus de 24 numéros de la collection *Autogestion* provenant d'un fond d'archives en sciences sociales allant de 1966 à 1979. Chaque numéro contient plusieurs articles courts ou longs pour un total de 448 articles. La version électronique de ce corpus étant obtenue par OCR, nous disposons de deux niveaux de segmentation supplémentaires : *page* (correspondant à l'OCR de chaque image d'une page donnée de la collection) et *bloc de texte* (l'unité minimale de texte cohérent produite par le système OCR). Nous considérons ici un sous-ensemble du corpus entier, contenant 4786 pages, 33551 blocs de texte pour un total de 1,5 million de tokens. Initialement, le processus d'étiquetage des rôles sémantiques produit 143 317 détections de cadres, qui sont réduites à 124 925 lorsque l'on se concentre sur les verbes non auxiliaires à partir du processus de filtrage **F1**. Chaque détection de cadre sémantique produit en moyenne 1,7 élément de trame, ce qui signifie que le premier ensemble est composé de 247 907 questions. Le tableau 3 indique le nombre de questions générées à la suite des processus de filtrage décrits dans la Sextion 3.

En plus de l'évaluation quantitative et descriptive des questions générées sur le corpus de *self-management*, nous avons également effectué une première évaluation qualitative sur un sous-ensemble de la collection selon deux dimensions. La première dimension se concentre sur la qualité de la forme de la question, avec des questions catégorisées comme "*Valide*", "*Question incohérente*" ou "*Question non grammaticale*". Dans la deuxième dimension, nous évaluons la pertinence de la question une fois qu'elle a été validée dans la dimension précédente. Cette évaluation implique trois échelles de Likert à 5 points :



(a) Likert (1) (b) Likert (2) (c) Likert (3)

FIGURE 1 – Évaluation de la pertinence des questions générées

1. "Le segment mis en évidence correspond bien à une réponse à la question"
2. "La question est pertinente dans le contexte de la phrase"
3. "La question est pertinente dans le contexte général de la lecture"

Un annotateur professionnel a été embauché pour cette tâche et a annoté un total de 582 questions. Environ 92% des questions ont été validées sur leur forme, ce qui confirme la qualité syntaxique de notre système de génération de questions.

Pour les annotations de pertinence, les résultats sont également prometteurs. En termes d'adéquation de réponse, la majorité des questions (67%) ont reçu une note indiquant un niveau élevé d'adéquation [3] (Figure 1a). Les deux échelles de Likert mesurant la pertinence de la question sont plus subjectives, mais une grande proportion de questions (plus de 68%) ont été évaluées comme pertinentes dans le contexte local (Figure 1b). Dans le contexte global de la lecture, le pourcentage de questions évaluées comme pertinentes diminue, avec un peu plus de la moitié des questions répondant au même critère de score (Figure 1c).

Pour vérifier l'accord entre les annotateurs, un sous-ensemble de 129 questions a été annoté par un second annotateur. Pour la première dimension (forme de surface), nous avons remarqué seulement 11 désaccords entre les deux annotateurs. Pour la seconde, concernant le premier Likert avec une évaluation simplifiée en 3 catégories en regroupant les choix 1 et 2 et les choix 4 et 5, nous avons mesuré 25 désaccords sur 115 annotations. Avec le même regroupement, nous avons obtenu 43 désaccords sur 115 annotations pour le Likert 2 et 65 désaccords sur 115 annotations pour le Likert 3. Ces nombres plus élevés de désaccords étaient attendus, car cette dernière évaluation est hautement subjective.

7 Évaluation des liens entre documents

L'évaluation des liens produits grâce aux questions nécessite une évaluation auprès d'utilisateurs dans un contexte réaliste étant donné la forte subjectivité de cette évaluation. Une évaluation de ce type est en cours. Cependant il est également possible d'évaluer ces liens par des mesures complémentaires automatiques. Ainsi nous avons évalué notre approche d'appariement basée sur les questions par rapport à des méthodes d'appariement uniquement basées sur la similarité textuelle. Pour quantifier la

3. Dans ce paragraphe, la notion de haut niveau d'adéquation correspond aux scores Likert > 3

	Pourcentage d'intersection		
Ensembles de similarités	(ALL)	(OUT_ART)	(OUT_NUM)
[QA] // [SENTENCE]	21 %	17 %	19 %
[QA] // [TEXTBLOCK]	23 %	12 %	20 %

TABLE 4 – % d'intersection entre les ensembles de similarités

différence, nous considérons l'ensemble de liens produits par une question comme une entité unique et calculons l'intersection avec l'ensemble de liens générés par d'autres méthodes de similarité.

Pour cela, nous avons réalisé le même appariement que celui présenté dans la section 4 (noté **QA** dans la suite), mais en utilisant la mesure de similarité uniquement sur le texte selon deux granularités : phrase (**SENTENCE**) et bloc de texte de l'OCR (**TEXTBLOCK**). Nous créons un lien entre deux phrases ou deux TextBlocks si la similarité entre leurs embeddings est inférieure à un seuil. Comme pour la similarité sur les questions, nous ne conservons que les N meilleurs liens. Dans cette expérience, N était fixé à 49.

Pour nous assurer que les liens sont comparés au même niveau de granularité, nous considérons que les liens sont identiques s'ils pointent vers la même page. Le pourcentage de chevauchement est calculé comme l'intersection entre l'ensemble des liens produits par les questions sur une phrase ou un TextBlock et ceux produits par la similarité sur le texte.
Bien que cette évaluation (Table 4) ne permette pas à elle seule de mesurer la qualité de nos liens, elle nous montre que notre méthode produit des liens originaux avec près de 80% des 49 pages les plus similaires qui sont différentes de celles produites en utilisant des méthodes de similarité directement sur les segments de texte. Une évaluation subjective qui vérifiera le retour des lecteurs professionnels aux liens et explications proposés par notre méthode sera bientôt réalisée.

8 Conclusion

Cet article propose une nouvelle approche pour construire des liens entre des documents en se basant sur la génération de questions. Nos expériences montrent la qualité de nos questions générées automatiquement, leur pertinence dans un contexte local, ainsi que l'originalité des liens produits par l'appariement de ces questions. Des expériences restent à mener pour étudier de manière plus qualitative les liens générés, ainsi que pour enrichir et filtrer de manière plus fine la grande quantité de questions sur le corpus.

Remerciements

Ces travaux ont été partiellement financés par l'Agence Nationale pour la Recherche (ANR) à travers le projet ANR-19-CE38-0011 (ARCHIVAL).
Ces travaux ont bénéficié d'un accès aux ressources en HPC/IA de l'IDRIS au travers de l'allocation de ressources 2022-AD011012688R1 attribuée par GENCI.

Références

ALBERTI C., ANDOR D., PITLER E., DEVLIN J. & COLLINS M. (2019). Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6168–6173, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1620.

BECHET F., ANTOINE E., AUGUSTE J. & DAMNATI G. (2022). Question generation and answering for exploring digital humanities collections. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4561–4568, Marseille, France : European Language Resources Association.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural language processing with Python : analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

DAI A. M., OLAH C. & LE Q. V. (2015). Document embedding with paragraph vectors. *CoRR*, **abs/1507.07998**.

D'HOFFSCHMIDT M., BELBLIDIA W., HEINRICH Q., BRENDLÉ T. & VIDAL M. (2020). FQuAD : French question answering dataset. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 1193–1208, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.107.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : a skilled pretrained French sequence-to-sequence model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9369–9390, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank : An annotated corpus of semantic roles. *Comput. Linguist.*, **31**(1), 71–106. DOI : 10.1162/0891201053630264.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 2383–2392 : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

REIMERS N. & GUREVYCH I. (2019). Sentence-bert : Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* : Association for Computational Linguistics.

REIMERS N. & GUREVYCH I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4512–4525, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.365.

TSATSARONIS G., VARLAMIS I. & VAZIRGIANNIS M. (2014). Text relatedness based on a word thesaurus. *CoRR*, **abs/1401.5699**.

WANG Z., MI H. & ITTYCHERIAH A. (2016). Sentence similarity learning by lexical decomposition and composition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1340–1349, Osaka, Japan : The COLING 2016 Organizing Committee.

Annexes

A Exemples de questions par annotations en likert

Le score donné à la question pour l'affirmation est indiqué entre parenthèses.
"La question est pertinente dans le contexte général de la lecture" :

- Qu'est ce qui ne sautent pas aux yeux ? (1)
- Comment simplifient-t-on le problème ? (1)
- Qui a rejeté la solution préconisée ? (1)
- Qu'est ce qui n'existe pas ? (1)
- Qu'exige le jeu du marché ? (3)
- Qu'a permis ce système dans la période en question ? (3)
- Qui est à l'origine de la participation des travailleurs ? (3)
- Qu'exige l'autogestion ? (5)
- Qui agit en tant que force idéologique d'avant-garde dans la réalisation des intérêts fondamentaux des travailleurs ? (5)
- Contre quoi les travailleurs ont-ils à lutter ? (5)
- Quel est le moyen le plus simple d'exercer un contrôle sur les travailleurs ? (5)

"La question est pertinente dans le contexte de la phrase" :

- Quelle est la conséquence de ces déclarations ? (1)
- Qu'accompagne l'autogestion ? (1)
- Qu'est-ce qui est à l'origine de l'économie auto-organisée et de ses intérêts ? (1)
- Quel syndicat s'oppose à ce mouvement ? (1)
- Qui favorise le développement de ce mouvement ? (3)
- Qui tend à atomiser l'économie ? (3)
- Combien de travailleurs sont capables de gérer l'usine ? (3)
- Qu'exige le jeu du marché ? (5)
- Quelle proportion des travailleurs se retrouve salariée ? (5)
- Que permet le statut de coopérative ? (5)
- Quel est le moyen le plus simple d'exercer un contrôle sur les travailleurs ? (5)

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

HATS : Un jeu de données intégrant la perception humaine appliquée à l'évaluation des métriques de transcription de la parole

Thibault Bañeras-Roux[1,2] Jane Wottawa[3] Michael Rouvier[2] Teva Merlin[2]
Richard Dufour[1]
(1) Laboratoire des Sciences du Numérique de Nantes (LS2N), France
(2) Laboratoire Informatique d'Avignon (LIA), France
(3) Laboratoire d'Informatique de l'Université du Mans (LIUM), France
thibault.roux@univ-nantes.fr, jane.wottawa@univ-lemans.fr,
michael.rouvier@univ-avignon.fr, teva.merlin@univ-avignon.fr,
richard.dufour@univ-nantes.fr

RÉSUMÉ

Traditionnellement, les systèmes de reconnaissance automatique de la parole (RAP) sont évalués sur leur capacité à reconnaître correctement chaque mot contenu dans un signal vocal. Dans ce contexte, la mesure du taux d'erreur-mot est la référence pour évaluer les transcriptions vocales. Plusieurs études ont montré que cette mesure est trop limitée pour évaluer correctement un système de RAP, ce qui a conduit à la proposition d'autres variantes et d'autres métriques. Cependant, toutes ces métriques restent orientées "système" alors même que les transcriptions sont destinées à des humains. Dans cet article, nous proposons un jeu de données original annoté manuellement en termes de perception humaine des erreurs de transcription produites par divers systèmes de RAP. 143 humains ont été invités à choisir la meilleure transcription automatique entre deux hypothèses. Nous étudions la relation entre les préférences humaines et diverses mesures d'évaluation pour les systèmes de RAP, y compris les mesures lexicales et celles fondées sur les plongements de mots.

ABSTRACT

HATS : An open dataset integrating human perception applied to the evaluation of Automatic Speech Recognition metrics

Traditionally, Automatic Speech Recognition (ASR) systems are evaluated on their ability to correctly recognize each word contained in a speech signal. In this context, the Word Error Rate metric is the reference for evaluating speech transcripts. Several studies have shown that this measure is too limited to correctly evaluate an ASR system, which has led to the proposal of other variants and other metrics. However, all these metrics remain system-oriented, even when transcripts are intended for humans. In this paper, we describe an original manually annotated dataset in terms of human perception of transcription errors produced by various ASR systems. 143 humans were asked to choose the best automatic transcription between two hypotheses. We investigate the relationship between human preferences and various evaluation metrics for ASR systems, including lexical and embedding-based metrics.

MOTS-CLÉS : reconnaissance de la parole, jeu de données, perception, métrique, corpus.

KEYWORDS: speech recognition, dataset, perception, metric, corpus.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 10–18.

1 Introduction

La Reconnaissance Automatique de la Parole (RAP) consiste à transcrire de la parole en texte. Depuis l'utilisation des systèmes de RAP fondés sur les Modèles de Markov cachés (Juang & Rabiner, 1991), le domaine a connu d'importants progrès avec l'utilisation des réseaux de neurones profonds et des méthodes auto-supervisées telles que wav2vec (Baevski *et al.*, 2020) et HuBERT (Hsu *et al.*, 2021) qui permettent d'extraire des informations de la parole sans données étiquetées. Ces transcriptions automatiques peuvent être utilisées par les humains dans le cas par exemple du sous-titrage, de la rédaction de messages ou par des systèmes tiers tels que les assistants personnels virtuels.

Face à des erreurs dans un flux de parole ou des des textes, un humain est capable de les intégrer et éventuellement de les corriger si elles n'impactent pas fondamentalement le sens de la séquence (Cutler, 2012). Les erreurs dans les transcriptions automatiques proviennent de divers facteurs, tels que le bruit dans le signal vocal, les accents des locuteurs ou les limitations techniques. La question est de savoir quelles erreurs sont acceptables et lesquelles entraînent des difficultés de compréhension chez l'humain. Par conséquent, il apparaît souhaitable que les métriques d'évaluation des systèmes de RAP se rapprochent de la perception humaine.

Les métriques automatiques les plus couramment utilisées pour évaluer les systèmes de RAP sont le taux d'erreur-mot (WER pour *Word Error Rate* en anglais), qui mesure le nombre de mots incorrectement transcrits, et le taux d'erreurs-caractères (CER pour *Character Error Rate* en anglais) qui évalue le nombre de caractères qui diffèrent par rapport à la référence. Cependant, de nombreux travaux (Wang *et al.*, 2003; Favre *et al.*, 2013; Itoh *et al.*, 2015; Kafle & Huenerfauth, 2017) ont soulevé des problèmes liés à ces mesures tels que l'absence de pondération des erreurs ou encore le manque d'informations linguistiques et de connaissances sémantiques. En réponse à ces problèmes, le développement de nouvelles métriques a suscité un intérêt croissant dans la communauté. Des métriques alternatives se focalisant sur la qualité et l'efficacité des transcriptions automatiques ont été proposées (Nam & Fels, 2019; Gordeeva *et al.*, 2021; Kim *et al.*, 2021; Bañeras-Roux *et al.*, 2022).

Des évaluation humaines de systèmes de RAP ont par le passé été réalisées, notamment à l'aide d'une expérience *côte-à-côte* (Gordeeva *et al.*, 2021; Kafle & Huenerfauth, 2017; Kim *et al.*, 2022), consistant à demander à des sujets humains de choisir la meilleure transcription parmi deux proposées. Ces études ont également permis d'évaluer des métriques automatiques. La présente étude s'inspire de ce protocole expérimental mais, au lieu d'altérer le signal vocal ou d'utiliser différentes sorties du même système de RAP afin d'obtenir deux hypothèses différentes, notre étude met en compétition les sorties de 10 systèmes intégrant des architectures différentes. En plus, les hypothèses ont été appariées selon un ensemble de critères métriques bien définis.

Cette expérience nous permet de distribuer librement un nouveau jeu de données ouvert, nommé appelé HATS (Human Assessed Transcription Side-by-side), intégrant des préférences humaines sur des transcriptions automatiques. Comme seconde contribution, une étude originale est menée à partir de HATS sur l'évaluation des métriques automatiques de systèmes de RAP en étudiant leur accord avec les évaluations humaines. Notre objectif est de mettre en évidence les métriques qui corrèlent le mieux avec la perception humaine.

2 Systèmes de transcription et métriques

Dans la section 2.1, nous présentons les systèmes de reconnaissance automatique de la parole (RAP), y compris le protocole expérimental, utilisés pour construire le jeu de données HATS. Ensuite, nous présentons dans la section 2.2 les différentes mesures d'évaluation utilisées dans notre étude.

2.1 Systèmes de reconnaissance automatique de la parole (RAP)

Dans cette étude, nous avons implémenté huit systèmes de bout-en-bout (*end-to-end*) en utilisant la boîte à outils Speechbrain (Ravanelli *et al.*, 2021), et deux systèmes *pipeline* fondés sur des réseaux de neurones artificiels profonds et des Modèles de Markov caché (DNN-HMM) à l'état de l'art [1] en utilisant la boîte à outils Kaldi (Povey *et al.*, 2011). Concernant les systèmes de bout-en-bout, chacun a été entraîné en utilisant un modèle acoustique auto-supervisé différent (sept systèmes utilisent des variantes des modèles wav2vec2 appris sur le français et un utilise le modèle XLS-R-300m). Pour les systèmes pipeline, l'un des systèmes contient une étape supplémentaire de ré-évaluation en utilisant un modèle de langage neuronal.

Tous les systèmes ont été entraînés sur du français avec les corpus ESTER 1 et 2 (Galliano *et al.*, 2006, 2009), EPAC (Esteve *et al.*, 2010), ETAPE (Gravier *et al.*, 2012), REPERE (Giraudel *et al.*, 2012) et des données internes. Ces corpus représentent environ 940 heures d'audio composées de données de diffusion radio et télévision. Les transcriptions permettant de construire notre corpus HATS sont extraites du corpus de test de REPERE, qui représente environ 10 heures de données audio.

2.2 Métriques pour la RAP

En plus des métriques lexicales classiques telles que le taux d'erreur-mot (WER) et le taux d'erreur-caractère (CER), nous étudierons trois métriques sémantiques fondées sur les plongements de mots. La première, le **taux d'erreur-plongement** ou **EmBER** (pour *Embedding Error Rate*) (Bañeras-Roux *et al.*, 2022), est un WER où les erreurs de substitution sont pondérées en fonction de la distance cosinus entre le plongement lexical d'un mot de référence et le plongement du mot substituant. Les plongements lexicaux sont obtenus à partir de fastText (Grave *et al.*, 2018; Bojanowski *et al.*, 2017). La deuxième métrique, **SemDist** (Kim *et al.*, 2021), consiste à calculer une similarité cosinus entre la référence et l'hypothèse en utilisant des plongements obtenus au niveau de la phrase. Différentes méthodes sont utilisées afin d'évaluer leur impact sur la métrique : utiliser le plongement du premier token des modèles CamemBERT (Martin *et al.*, 2020) ou FlauBERT (Le *et al.*, 2020) ou d'un modèle de plongement de phrase (SentenceBERT (Reimers & Gurevych, 2019)). La dernière métrique sémantique est le **BERTScore** (Zhang* *et al.*, 2020), qui calcule un score de similarité pour chaque mot de la phrase candidate avec chaque mot de la phrase de référence en utilisant des plongements contextuels. Dans notre étude, nous utilisons un modèle multilingue BERT et le modèle CamemBERT. Alors que les transcriptions textuelles sont issues de la parole, nous considérons finalement une métrique **taux d'erreur-phonème** ou **PER** (pour *Phoneme Error Rate*) qui consiste à calculer une distance de Levenshtein entre les séquences de phonèmes de référence et d'hypothèse obtenues avec l'aide d'un convertisseur texte-à-phonème [2].

3 Évaluation humaine

La collecte du corpus HATS est décrite dans cette partie. La section 3.1 résume la mise en place de l'expérience perceptive tandis que la section 3.2 décrit le protocole permettant la sélection des transcriptions automatiques en vue de leur évaluation humaine.

1. https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/
2. https://github.com/Remiphilius/PoemesProfonds

3.1 Expérience perceptive

Dans notre étude, l'expérience côte-à-côte consiste à présenter d'une part, au sujet, une référence textuelle transcrite manuellement à partir d'un court extrait de parole, et d'autre part deux transcriptions automatiques, chacune réalisée par un système de RAP différent. Les transcriptions automatiques présentaient toujours des déviations par rapport à la référence (*i.e.* contenaient des erreurs de transcription). À l'aide de la souris, les participants devaient choisir, selon eux, la meilleure hypothèse en fonction de la référence. L'étude a utilisé un protocole d'instruction minimal, permettant aux participants de déterminer eux-mêmes les critères qui étaient importants pour déterminer la qualité d'une transcription. La référence était uniquement sous forme écrite afin que les sujets et méthodes automatique aient accès aux mêmes informations (Vasilescu *et al.*, 2012).

Pour l'étude, 143 participants se sont portés volontaires en ligne. Avant de débuter l'évaluation, les participants ont rempli un questionnaire afin de renseigner leur âge, le nombre de langues parlées, leur langue maternelle, ainsi que leur niveau d'éducation. Chaque participant a évalué 50 triplets de transcription dans un ordre aléatoire, avec un temps total moyen de 15 minutes par participant.

3.2 Protocole de sélection des transcriptions automatiques

Les triplets de transcriptions, issues du corpus de test REPERE, ont été sélectionnées en respectant les 3 critères suivants : (1) les deux hypothèses sont différentes et doivent avoir au moins un caractère de différence avec la référence, (2) chaque système doit avoir au moins quelques hypothèses face à des hypothèses de chaque systèmes et (3) la sélection des paires d'hypothèses s'appuie sur le respect de critères basées sur les scores des métriques.

Categorie	**Critères de métriques**	**Référence**	**Hypothèse A**	**Hypothèse B**
(A)	WER =	et on découvre les spectateurs	ϵ on découvre les spectateurs	et on découvre les spectacles
(A)	WER =	et on découvre les spectateurs	ϵ on découvre les spectateurs	et on découvre les spectacles
(A)	CER >	sur la vie politique	ϵ la vie politique	c' la vie politique
(A)	SemDist >>	c' est à paris	ϵ est à paris	c' est appau ϵ
(B)	WER = ; SemDist >	encore du rock	corps du rock	encore du rok
(C)	WER $\neq$ BERTscore	où les passions sont si vives	ϵ les patients sont si vive	où les patients sont si vifs

TABLE 1 – Détails de quelques critères de choix des stimuli avec des exemples.

Le point (3) peut être divisé en trois catégories différentes : (A) chaque métrique a été comparée à elle-même en présentant soit le même score, soit un score légèrement différent, soit un score très différent entre les deux hypothèses, (B) dans les deux hypothèses, le WER ou le CER étaient égaux mais le WER ou le CER, EmbER, SemDist, BERTScore étaient différents, (C) les métriques indiquaient des prédictions opposées sur quelle est la meilleure hypothèse (*e.g.* $\text{WER}_{(\text{hypA})} > \text{WER}_{(\text{hypB})}$ mais $\text{CER}_{(\text{hypA})} < \text{CER}_{(\text{hypB})}$). Le tableau 1 illustre la manière dont les hypothèses ont été confrontés avec des exemples concrets utilisés dans la tâche d'évaluation humaine.

4 Le jeu de données HATS

4.1 Description du corpus

Le corpus HATS comprend 1 000 références avec respectivement deux transcription automatiques provenant de systèmes de RAP différents. 143 humains ont chacun évalué 50 triplets référence-hypothèses, ce qui a conduit au final à 7 150 annotations. Notons que tous les triplets de transcription ont été évaluées par au moins 7 participants.

4.2 Méthodologie d'évaluation des métriques

Nous ne conservons que les annotations de transcription ayant obtenu un niveau de consensus suffisant entre les annotateurs. Nous calculons leur accord de la façon suivante : Soit A le nombre de sujets choisissant la première transcription automatique et B ceux ayant choisi l'autre, alors l'accord humain se calcule en prenant le maximum entre A et B, divisé par la somme de A et B. Les choix humains sont alors considérés selon trois pourcentages d'accord : **100%** (seulement les triplets où tous les sujets ont choisi la même hypothèse), **70%**, ou **0%** (pas de filtre) ; ce qui correspond respectivement à 371, 819 et 1000 triplets. Le seuil de 70 % a été choisi afin d'avoir un accord cohérent des annotateurs même si tous les participants ne répondent pas de la même manière (Nowak & Rüger, 2010).

Pour l'évaluation des métriques automatiques, nous cherchons à savoir si celles-ci corrèlent avec la perception humaine (*i.e.* la métrique automatique désigne la même que celle choisit par les humains). Nous obtenons donc au final, pour chaque métrique, son taux de couverture par rapport à l'annotation humaine. Le code ainsi que les données de HATS sont mises disponibles publiquement [3].

5 Évaluation des métriques

Le tableau 2 résume les résultats obtenus par chaque métrique automatique selon leur accord avec la perception humaine. Sans surprise, plus l'accord humain est élévée, plus les performances des métriques sont élevées. En contradiction avec les résultats d'études antérieures (Kim *et al.*, 2022), notre étude montre que le CER corrèle mieux avec la perception humaine que le WER. Cette divergence peut être attribuée à l'utilisation d'un texte écrit comme référence dans notre expérience perceptive, plutôt qu'un texte audio, ou à des variations linguistiques intrinsèques entre le français et l'anglais (l'orthographe du français comporte beaucoup de lettres muettes).

Il est intéressant de noter qu'au niveau des phonèmes, le PER donne de bons résultats, meilleurs que ceux du WER et du CER, malgré le fait que les humains aient fait leurs choix sur la base du texte uniquement. Cela montre que les humains semblent tenir compte de la façon dont les phrases sont phonétisées, même pendant la lecture. Ceci est particulièrement vrai si les phrases sont contrastées avec une référence.

Bien que les hypothèses choisies selon BERTScore avec les plongements BERT-base-multilingue soient 8 % meilleures que celles choisies selon SemDist utilisant des plongements de phrases multilingues, il serait précipité de conclure que la stratégie BERTScore est meilleure pour déterminer la qualité des transcriptions car les deux métriques utilisent différents plongements. En comparant ces métriques avec les mêmes plongements, SemDist est meilleure que BERTScore quand il s'agit des

3. https://github.com/thibault-roux/metric-evaluator

Accord	100%	70%	0% (Full)
Taux d'erreur mot	63% (23%)	53% (28%)	49% (28%)
Taux d'erreur caractère	77% (17%)	64% (21%)	60% (22%)
Taux d'erreur plongement	73% (12%)	62% (16%)	57% (17%)
BERTScore BERT-base multilingue	84% (0%)	75% (1%)	70% (1%)
BERTScore CamemBERT-base	81% (0%)	72% (0%)	68% (0%)
BERTScore CamemBERT-large	80% (0%)	68% (0%)	65% (0%)
SemDist CamemBERT-base	86% (0%)	74% (0%)	70% (0%)
SemDist CamemBERT-large	80% (0%)	71% (0%)	67% (0%)
SemDist Phrases CamemBERT-base	86% (0%)	75% (0%)	71% (0%)
SemDist Phrases CamemBERT-large	90% (0%)	78% (0%)	73% (0%)
SemDist Phrases multilingual	76% (0%)	66% (0%)	62% (0%)
SemDist FlauBERT-base	65% (0%)	62% (0%)	59% (0%)
Taux d'erreur phoneme	80% (14%)	69% (16%)	64% (17%)

TABLE 2 – Performance de chaque métrique en fonction de l'accord hymain. **Full** signifie qu'aucun filtre sur l'accord n'a été appliqué à l'ensemble des données. Le nombre entre parenthèses indique le pourcentage de fois où la mesure a donné la même note aux deux hypothèses.

plongements de CamemBERT-base, et SemDist a des performances similaires à BERTScore quand il s'agit des plongements de CamemBERT-large.

Sur les accords à 70 % et 0 %, le taux d'erreur mot a des performances proche d'un choix aléatoire. Cela est dû au fait que dans notre jeu de données, de nombreux cas présentent des hypothèses avec le même WER, des prédictions égales étant considérées comme un échec de la métrique puisque les humains sont capables de sélectionner une hypothèse. De plus, nous pouvons observer que SemDist utilisant les plongements de FlauBERT a de moins bonnes performances que le CER. Cela met en évidence la nécessité de choisir soigneusement les plongements et de les évaluer sur un jeu de données tel que HATS avant de tirer des conclusions sur les systèmes au niveau sémantique. Enfin, selon notre corpus orienté vers l'humain, la meilleure métrique est SemDist utilisant les plongements de phrase de CamemBERT-large, ce qui peut s'expliquer par le fait que cette métrique s'appuie sur des plongements spécifiquement entraînés pour maximiser la similarité entre des phrases ayant un sens similaire. Il est important de noter qu'une grande quantité de données annotées est nécessaire pour utiliser ces métriques fondées sur les plongements.

6 Conclusion et perspectives

Dans cette étude, des métriques automatiques appliquées à différents systèmes de RAP ont été comparées à l'évaluation humaine de différentes hypothèses erronées selon une référence écrite.

Nos résultats montrent que SemDist avec les plongements de phrases de BERT évaluent les transcriptions d'une manière qui semble acceptable pour les évaluateurs humains. Dans le cas de plongements de phrases, BERTScore semble être la deuxième meilleure option. Cette métrique est plus stable que SemDist sur les plongements de BERT. Néanmoins, si possible, les métriques devraient être évaluées sur des ensembles de données comprenant également des annotations humaines, comme HATS.

Bien que ces nouvelles méthodes d'évaluation soient intéressantes en RAP, l'avantage des métriques WER et CER est leur faible coût de calcul et l'interprétabilité du score. Par conséquent, la prochaine

étape pourrait consister à développer des métriques en corrélation avec la perception humaine tout en restant interprétables, ce qui n'est pour l'instant pas le cas de la métrique SemDist par exemple.

Dans le cadre d'un travail futur, une étude supplémentaire pourrait être menée en reproduisant l'expérience actuelle en utilisant une référence audio au lieu d'une référence textuelle, de sorte que les sujets ne disposent pas d'informations sur les caractères. Cette approche nous permettrait d'examiner les variations éventuelles et de déterminer si la métrique CER est toujours considérée comme meilleure que le WER dans un contexte multimodal.

Références

BAEVSKI A., ZHOU Y., MOHAMED A. & AULI M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, **33**, 12449–12460.

BAÑERAS-ROUX T., ROUVIER M., WOTTAWA J. & DUFOUR R. (2022). Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. In *Interspeech 2022*.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, **5**, 135–146.

CUTLER A. (2012). *Native listening : Language experience and the recognition of spoken words*. Mit Press.

ESTEVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. In *International Conference on Language Resources and Evaluation (LREC)*.

FAVRE B., CHEUNG K., KAZEMIAN S., LEE A., LIU Y., MUNTEANU C., NENKOVA A., OCHEI D., PENN G., TRATZ S. *et al.* (2013). Automatic human utility evaluation of ASR systems : Does WER really predict performance ? In *INTERSPEECH*, p. 3463–3467.

GALLIANO S., GEOFFROIS E., GRAVIER G., BONASTRE J.-F., MOSTEFA D. & CHOUKRI K. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *International Conference on Language Resources and Evaluation (LREC)*, p. 139–142.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The repere corpus : a multimodal corpus for person recognition. In *International Conference on Language Resources and Evaluation (LREC)*, p. 1102–1107.

GORDEEVA L., ERSHOV V., GULYAEV O. & KURALENOK I. (2021). Meaning Error Rate : ASR domain-specific metric framework. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, p. 458–466.

GRAVE É., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

GRAVIER G., ADDA G., PAULSSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *International Conference on Language Resources and Evaluation (LREC)*, p. 114–118.

HSU W.-N., BOLTE B., TSAI Y.-H. H., LAKHOTIA K., SALAKHUTDINOV R. & MOHAMED A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 3451–3460.

ITOH N., KURATA G., TACHIBANA R. & NISHIMURA M. (2015). A metric for evaluating speech recognizer output based on human-perception model. In *Sixteenth Annual Conference of the International Speech Communication Association*.

JUANG B. H. & RABINER L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, **33**(3), 251–272.

KAFLE S. & HUENERFAUTH M. (2017). Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 165–174.

KIM S., ARORA A., LE D., YEH C.-F., FUEGEN C., KALINLI O. & SELTZER M. L. (2021). Semantic Distance : A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. In *Proc. Interspeech 2021*, p. 1977–1981. DOI : 10.21437/Interspeech.2021-1929.

KIM S., LE D., ZHENG W., SINGH T., ARORA A., ZHAI X., FUEGEN C., KALINLI O. & SELTZER M. (2022). Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In *Proc. Interspeech 2022*, p. 3978–3982. DOI : 10.21437/Interspeech.2022-11144.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 2479–2490.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219.

NAM S. & FELS D. (2019). Simulation of Subjective Closed Captioning Quality Assessment Using Prediction Models. *International Journal of Semantic Computing*, **13**(01), 45–65.

NOWAK S. & RÜGER S. (2010). How reliable are annotations via crowdsourcing : a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, p. 557–566.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, volume CONF : IEEE Signal Processing Society.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. arXiv :2106.04624.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992.

VASILESCU I., ADDA-DECKER M. & LAMEL L. (2012). Cross-lingual studies of ASR errors : paradigms for perceptual evaluations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, p. 3511–3518.

WANG Y.-Y., ACERO A. & CHELBA C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, p. 577–582 : IEEE.

ZHANG* T., KISHORE* V., WU* F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with bert. In *International Conference on Learning Representations*.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Résumé automatique multi-documents guidé par une base de résumés similaires

Florian Baud[1,2] Alex Aussem[1]
(1) LIRIS UMR 5205 CNRS, 25 Avenue Pierre de Coubertin, 69622 Villeurbanne, France
(2) Visiativ, 26 Rue Benoit Bennier, 69260 Charbonnières-les-Bains, France
florian.baud@liris.cnrs.fr, alexandre.aussem@liris.cnrs.fr

RÉSUMÉ

Le résumé multi-documents est une tâche difficile en traitement automatique du langage, ayant pour objectif de résumer les informations de plusieurs documents. Cependant, les documents sources sont souvent insuffisants pour obtenir un résumé qualitatif. Nous proposons un modèle guidé par un système de recherche d'informations combiné avec une mémoire non paramétrique pour la génération de résumés. Ce modèle récupère des candidats pertinents dans une base de données, puis génère le résumé en prenant en compte les candidats avec un mécanisme de copie et les documents sources. Cette mémoire non paramétrique est implémentée avec la recherche approximative des plus proches voisins afin de faire des recherches dans de grandes bases de données. Notre méthode est évalué sur le jeu de données *MultiXScience* qui regroupe des articles scientifiques. Enfin, nous discutons de nos résultats et des orientations possibles pour de futurs travaux.

ABSTRACT

Non-Parametric Memory Guidance for Multi-Document Summarization

Multi-document summarization is a difficult task in natural language processing, aiming to summarize information from several documents. However, the source documents are often insufficient to obtain a qualitative summary. We propose a retriever-guided model combined with non-parametric memory for summary generation. This model retrieves relevant candidates from a database and then generates the summary considering the candidates with a copy mechanism and the source documents. The retriever is implemented with Approximate Nearest Neighbor Search (ANN) to search large databases. Our method is evaluated on the *MultiXScience* dataset which includes scientific articles. Finally, we discuss our results and possible directions for future work.

MOTS-CLÉS : Résumé multi-document, Augmentée par recherche, Guidage.

KEYWORDS: Multi-document summarization, Retrieval augmented, Guidance.

1 Introduction

Le résumé multi-documents automatique s'effectue à l'aide de deux méthodes : extractive (Wang *et al.*, 2020; Liu *et al.*, 2021) ou par abstraction (Jin *et al.*, 2020; Xiao *et al.*, 2022). Les méthodes dites extractives classent les phrases des documents sources afin d'obtenir un résumé. Ces méthodes réutilisent bien les informations importantes pour construire un résumé de qualité mais manquent de cohérence entre les phrases. Pour surmonter ce problème, les méthodes par abstraction sont étudiées pour rendre les résumés obtenus avec une meilleure cohérence. Les modèles générant des résumés

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 19–27.

par abstraction montrent d'excellentes performances sur le style d'écriture, mais oublient souvent des informations clés pour obtenir un résultat de qualité.

Pour que les modèles par abstraction tiennent compte des informations essentielles, (Dou *et al.*, 2021) guide leur modèle avec des informations supplémentaires comme un ensemble de mots-clés, des triplets de graphes, des phrases importantes des documents sources ou des résumés similaires récupérés depuis une base de connaissances. Avec comme mesure de similarité, la similarité cosinus. Leur méthode, qui utilise toutes les formes d'informations mentionnées précédemment, améliore la qualité et la contrôlabilité du résumé par rapport aux modèles non guidés. Cependant, le guidage nécessite des données d'entraînement spécifiques, notamment pour les mots-clés, les triplets de graphes et les phrases surlignées.

Notre proposition est qu'en guidant avec des résumés préexistants, le modèle puisse s'inspirer du résumé dans sa globalité mais aussi de pouvoir extraire des mots-clés et des phrases en utilisant un mécanisme de copie (Gu *et al.*, 2016; See *et al.*, 2017). Par conséquent, ce travail se concentre sur le guidage par des résumés similaires extraits d'une base de connaissances en utilisant la similarité cosinus. Le modèle, inspiré de RAG (Lewis *et al.*, 2020), est entièrement différentiable. En outre, le générateur du modèle utilise un mécanisme de copie sur les résumés remontés de la base de connaissances inspiré de (Cai *et al.*, 2021). Les conclusions de ces deux travaux ont motivé l'élaboration de notre modèle pour la tâche de résumé de texte multi-documents.

Nous démontrons le potentiel de notre méthode sur la base *MultiXScience* (Lu *et al.*, 2020) regroupant des articles scientifiques. Dans le cas d'articles scientifiques, les documents sources sont souvent insuffisants pour générer la partie *"related work"*. Des connaissances externes sont nécessaires pour rédiger un tel paragraphe. Notre objectif est de générer la partie *"related work"* avec notre méthode en intégrant des informations externes, dans notre cas des résumés préexistant et qui sont proches du résumé cible.

Dans cet article, nous étudions un modèle séquence à séquence guidé par une mémoire non paramétrique de résumés similaires. Notre contribution est double : premièrement, nous intégrons une mémoire non paramétrique comme définis dans (Lewis *et al.*, 2020) afin de récupérer les candidats à la génération du résumé, et deuxièmement, nous utilisons un mécanisme de copie pour intégrer ces candidats dans la procédure de génération. Le code de notre travail est disponible sur github [1].

2 Travaux similaires

Nous commençons par une revue des travaux similaires, tout d'abord les travaux sur les résumés d'articles scientifiques. En effet, (Cohan *et al.*, 2018) proposent de résumer des articles scientifiques provenant de *Arvix* et *Pubmed*, ils capturent la structure du document pour mieux représenter l'information du document source. Cependant cela concerne le résumé automatique d'un seul document. Dans le même but, (Cohan & Goharian, 2018; Yasunaga *et al.*, 2019) proposent de résumer un article avec les articles qui le citent. L'inconvénient de cette méthode est qu'elle ne peut pas être utilisée lors de la rédaction d'un article. Dans ce travail, nous utilisons les méthodes de résumé de texte multi-documents avec les références et non les articles qui citent les documents à résumer contrairement aux deux travaux précédents.

Les modèles utilisant des guidages sont proches de notre travail. (Cao *et al.*, 2018; Dou *et al.*, 2021)

1. https://github.com/florianbaud/retrieval-augmented-mds

extraient des résumés similaires d'une base de connaissances pour aider la génération du résumé cible. Cependant, ils utilisent des systèmes de recherche d'information tels que *ElasticSearch* pour trouver des candidats à la génération du résumé. De même, (An *et al.*, 2021) introduisent un système de recherche avec des vecteurs denses pour le résumé de texte, mais ils n'entraînent pas le système de recherche avec le reste du modèle. Dans notre cas, la recherche est effectuée avec des vecteurs denses et est entraînable pour trouver les candidats les plus pertinents à la génération du résumé.

Le domaine des modèles augmentés avec un système de recherche d'information différentiable partagent des points communs avec notre travail. Rag, (Lewis *et al.*, 2020) qui a introduit ce type de modèle, est utilisé pour la tâche de questions-réponses, où un contexte est donné pour répondre à la question. Le modèle récupère plusieurs contextes avec un système de recherche d'information entraînable end-to-end, puis répond à la question en utilisant chacun des candidats récupérés. Ces types de modèles sont également utilisés dans la tâche de traduction, où (Cai *et al.*, 2021) traduit une phrase avec une base de traduction préétablie. Leur modèle recherche dans cette base des traductions possibles de la phrase à traduire, puis les intègre dans la génération de la traduction par un mécanisme de copie. L'architecture que nous proposons repose sur la même idée : elle est basée sur un système de recherche d'information différentiable qui intègre la mémoire au moyen d'un mécanisme de copie. Il est intéressant de déterminer si le mécanisme de copie employé avec succès en traduction apporte également un gain au résumé multi-documents.

3 Méthode proposée

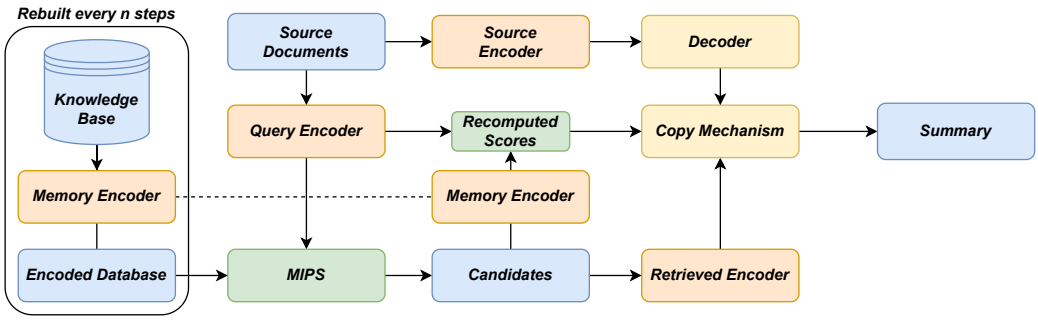


FIGURE 1 – Dans un premier temps, tous les éléments de la base de connaissances sont encodés avec un encodeur dédié : le *"Memory Encoder"*. Les documents sources sont transformés avec deux encodeurs distincts : le *"Query Encoder"* se charge d'encoder les documents sources afin de rechercher dans la base de connaissances et le *"Source Encoder"* se charge de représenter les documents sources pour la génération du résumé. Après avoir récupéré le top-k de la recherche, les candidats sont encodés avec l'encodeur *"Retrieved Encoder"* et avec le *"Memory Encoder"* pour recalculer le score de pertinence afin de propager le signal lors de la rétro-propagation du gradient. Finalement, le décodeur prend en entrée les documents sources et candidats pour la génération du résumé.

Inspiré par (Cai *et al.*, 2021), nous proposons un modèle composé d'un système de recherche d'information différentiable end-to-end et d'un générateur augmenté avec un mécanisme de copie. Nous commençons par présenter le système de recherche d'information différentiable puis le décodeur augmenté avec le mécanisme de copie. La figure 1 illustre l'architecture du modèle dans sa globalité.

3.1 Encodeur à mémoire non paramétrique

Les systèmes de recherche d'informations consistent en une requête et une base de recherche, l'objectif étant de retrouver les éléments pertinents de la base à l'aide de la requête. Dans notre cas les documents sources et les résumés cibles correspondent aux requêtes et à la base de recherche ou mémoire, respectivement noté Q et M. Ces documents sont encodés avec un modèle de type *Longformer* (Beltagy *et al.*, 2020). *LongFormer* possède une architecture *Transformer* (Vaswani *et al.*, 2017) qui peut traiter de longues séquences d'entrée avec une attention fenêtrée sur tous les tokens et une attention globale sur quelques tokens. Nous encodons les documents sources et les résumés avec deux encodeurs pré-entraînés distincts, un pour les documents sources et l'autre pour les résumés :

$$h^q = LED^q_{enc}(q) \quad q \in Q$$
$$h^m = LED^m_{enc}(m) \quad m \in M$$

où l'encodeur *LongFormer* est désigné par LED_{enc}. Tous les résumés de la base de connaissances sont encodés et stockés en amont de l'entraînement. Lors de la recherche dans la base, nous prenons le token *[CLS]* correspondant à un token spécial censé représenter le sens global du document encodé. Ce token est normalisé afin de calculer un score avec une fonction de pertinence :

$$h^q_{cls} = norm(h^q_{cls})$$
$$h^m_{cls} = norm(h^m_{cls})$$
$$score(x, y) = x^\top \cdot y$$

Le score représente la similarité cosinus entre les documents sources q et les documents de la mémoire m qui tombent dans l'intervalle $[-1, 1]$. Pour une recherche rapide, nous retrouvons les k documents les plus pertinents $m_{topk} = (m_1, \ldots, m_k)$ de la mémoire en utilisant la bibliothèque FAISS (Johnson *et al.*, 2021). À chaque propagation en avant de l'entraînement, les vecteurs des documents candidats $\{h^m_{cls,i}\}^k_{i=1}$ sont recalculés ainsi que les scores de pertinence $\{s_i = score(h^m_{cls,i}, h^q_{cls})\}^k_{i=1}$ afin de propager le signal jusqu'à l'encodeur de la mémoire comme dans (Cai *et al.*, 2021; Lewis *et al.*, 2020). De ce fait, les scores recalculés viennent biaiser le mécanisme de copie du décodeur permettant la propagation du signal jusqu'aux encodeurs.

L'encodeur de la mémoire ne ré-encode pas toute la base de connaissances à chaque étape de l'entraînement car cela serait un calcul coûteux. Au lieu de cela, la base de connaissances et l'index FAISS sont mis à jour à intervalles réguliers. D'autre part, nous encodons les candidats les plus recherchés (top-k) et les documents sources avec deux encodeurs distincts, LED^r_{enc} et LED^s_{enc}, comme indiqué ci-dessous :

$$h^r_{topk} = LED^r_{enc}(m_{topk})$$
$$h^s = LED^s_{enc}(q)$$

Le décodeur prend en compte ces deux résultats avec de l'attention croisée.

3.2 Le décodeur avec mécanisme de copie

Dans la partie décodeur de notre modèle, nous utilisons le décodeur de *LongFormer* et nous appliquons un mécanisme de copie aux candidats précédemment récupérés. Ainsi, nous avons :

$$h^d = LED_{dec}(y, h^s)$$

où LED_{dec} correspond à la partie décodeur du modèle *LongFormer*, et y est le résumé ciblé. Le décodeur prend en compte les documents sources h^s et les tokens précédents $y_{1:t-1}$, produisant un état caché h^d_t à chaque pas de la génération t. La probabilité du token suivant est calculée avec une fonction $softmax$:

$$P_{dec}(y_t) = softmax(W_d \cdot h^d_t + b_d) \tag{1}$$

où W_d est une matrice $hiddens_{size} \times vocab_{size}$ et b_d un vecteur de biais ; ces deux paramètres sont entraînés. Ensuite, nous incorporons les candidats du top-k m_{topk} avec un mécanisme de copie en calculant une attention croisée entre h^d_t et h^r_{topk}. Pour cela, nous réutilisons la partie attention croisée de *LongFormer* pour l'ajouter après son décodeur original. Cette nouvelle couche n'a qu'une seule tête d'attention afin d'utiliser les poids d'attention comme probabilité de copier un mot parmi les candidats du top-k. Étant donné k documents encodés dans h^r_{topk}, nous pouvons construire un ensemble de plongements (embeddings) de mots $\{r_{i,j}\}_{j=1}^{L_i}$ où $i \in [1, k]$, $j \in [1, L_i]$ et L_i est la longueur du document i. Concrètement, le poids d'attention du $j$ème token dans le $i$ème document pertinent est exprimé comme suit,

$$\alpha_{ij} = \frac{\exp(h^{d\top}_t W_a r_{i,j} + \beta s_i)}{\sum_{i=1}^{k}\sum_{j=1}^{L_i}\exp(h^{d\top}_t W_a r_{i,j} + \beta s_i)}$$

$$c_t = W_c \sum_{i=1}^{k}\sum_{j=1}^{L_i} \alpha_{ij} r_{i,j}$$

où W_a et W_c sont des paramètres entraînable, c_t est une représentation pondérée des k meilleurs candidats et β est un scalaire entraînable qui contrôle le score de pertinence entre les candidats récupérés et l'état caché du décodeur, permettant le flux de gradient vers les encodeurs comme dans (Cai *et al.*, 2021; Lewis *et al.*, 2020). L'équation 1 peut être réécrite pour inclure la mémoire :

$$P_{dec}(y_t) = softmax(W_d \cdot (h^d_t + c_t) + b_d) \tag{2}$$

Ainsi, la probabilité du prochain token prend en compte les poids d'attention des k candidats les plus importants. La probabilité finale du prochain token est donnée par :

$$P(y_t) = (1 - \lambda_t)P_{dec}(y_t) + \lambda_t \sum_{i=1}^{k}\sum_{j=1}^{L_i} \alpha_{ij} \mathbb{1}_{r_{ij}=y_t}$$

où λ_t est un scalaire agissant comme une porte logique calculé par un réseau feed-forward $\lambda_t = g(h^d, c_t)$. Le modèle est entraîné avec la fonction de perte de la log-vraisemblance $\mathcal{L} = -\log P(y^*)$ où y^* est le résumé cible.

3.3 Détails de l'entraînement

Notre modèle est composé de plusieurs encodeurs et décodeurs basés sur le *LongFormer* (Beltagy *et al.*, 2020). La taille de notre modèle est de 1.9 milliard de paramètres entraînables. Pour entraîner le modèle, nous avons utilisé la librairie *DeepSpeed* (Rasley *et al.*, 2020).

L'entraînement du modèle utilise les données *MultiXScience* comprenant 30 369 articles scientifiques pour l'entraînement, 5 066 articles de validation et 5 093 articles de test. L'objectif est de générer la partie *"related work"* en utilisant le résumé de l'article et les résumés des articles cités. Il s'agit d'un ensemble de données intéressant à expérimenter car l'écriture de la partie *"related work"* nécessite des connaissances extérieures aux documents sources.

Au début de l'apprentissage, les poids sont initialisés de façon aléatoire et l'encodeur sélectionne de "mauvais" candidats. Pour surmonter ce problème, nous pré-entraînons le système de recherche d'informations sur les données *MultiXScience* afin de commencer l'entraînement avec des résultats de bonne qualité. L'objectif est de maximiser la similarité entre le résumé et la section *"related work"*. Ces deux sections sont encodées avec les deux encodeurs précédents afin de calculer la similarité cosinus. Ainsi, pour une taille de données d'entraînement égale à N, nous avons N sections "abstract" encodées avec $A = \{LED^q_{enc}(a_i)\}^N_{i=1}$ et N sections "related work" encodées avec $B = \{LED^m_{enc}(b_j)\}^N_{j=1}$, le but est d'obtenir une similarité cosinus égal à 1 lorsque $j = i$ correspondant aux exemples positifs et -1 sinon pour les exemples négatifs. Nous calculons pour chaque élément de A, les erreurs suivantes :

$$\mathcal{L}_i(A, B) = -\log \frac{\exp\left(score(A_i, B_i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(score(A_i, B_j)/\tau\right)}$$

où τ est un paramètre de température choisi arbitrairement. L'erreur finale est $\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_i$ rétro-propagée dans les deux encodeurs.

4 Expérimentation

Dans cette section, nous présentons les expériences réalisées sur le jeu de données *MultiXScience* pour évaluer notre modèle. L'entraînement du modèle complet est plus difficile en raison de sa taille mais aussi du problème du démarrage à froid du système d'informations. Ce dernier correspond au fait que les résumés similaires récupérés ne sont pas suffisamment pertinents pour aider le modèle à générer des résumés de qualité.

Pour réduire la charge de calcul, nous avons utilisé un modèle réduit où la base de connaissances n'est pas du tout reconstruite. En outre, les paramètres de l'encodeur de mémoire ont été gelés afin de réduire la complexité de l'apprentissage. Ces deux modifications ont permis de réduire considérablement le temps d'apprentissage. Le modèle réduit a moins de paramètres entraînables (1.4 milliard). Le modèle a été entraîné pendant deux jours sur quatre GPU v100 avec l'optimiseur Adam et un taux d'apprentissage de $1e-4$, une taille de données d'entraînement de 32, un top-k de 5 pour le récupérateur et avec 2k étapes de *warmup* et une décroissance linéaire jusqu'à 20k étapes. Les scores rouges (Lin, 2004) sur le jeu de données *MultiXScience* sont reportés dans le tableau 1.

Method	**R-1**	**R-2**	**R-L**
Notre modèle	28.9	6.2	17.6
(Xiao *et al.*, 2022)*	31.9	**7.4**	18.0
(Lu *et al.*, 2020)*	**33.9**	6.8	**18.2**

TABLE 1 – Le score ROUGE (R-1/R-2/R-L) de nos résultats sur le jeu de données de test MultiXScience. Le symbole * signifie que les résultats ont été empruntés à (Xiao *et al.*, 2022).

Malgré sa réduction de taille, nous observons que le modèle est compétitif par rapport à l'état de l'art avec des méthodes n'utilisant pas de mécanisme de copie couplé avec un système de recherche d'informations sur des résumés similaires.

5 Conclusion

Cet article présente une architecture pour le résumé de texte multi-documents inspirée par des modèles augmentés par un système de recherche d'informations. Cette architecture comprend un système de recherche d'informations qui cherche dans une base de connaissances des résumés similaires pour la génération d'un résumé. Ces documents sont intégrés dans la génération au moyen d'un mécanisme de copie. Une version réduite du modèle a été évaluée sur le jeu de données *MultiXScience*. Les résultats sont compétitifs par rapport à l'état de l'art, mais nous espérons améliorer encore nos résultats, d'une part en corrigeant correctement le problème du démarrage à froid, et d'autre part en utilisant le modèle complet. Pour la suite des travaux, nous prévoyons également d'augmenter la taille de la base de connaissances avec de nouvelles données, et d'appliquer notre méthode à d'autres jeux de données pour la tâche de résumé multi-documents.

Remerciements

Nous remercions le Centre de Calcul CNRS/IN2P3 (Lyon - France) pour la mise à disposition des ressources informatiques nécessaires à ce travail. De plus, ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011013300 attribuée par GENCI. Enfin, nous remercions Roch Auburtin de la société Visiativ pour les conseils fournis sur nos travaux.

Références

AN C., ZHONG M., GENG Z., YANG J. & QIU X. (2021). Retrievalsum : A retrieval enhanced framework for abstractive summarization.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The long-document transformer. *arXiv :2004.05150*.

CAI D., WANG Y., LI H., LAM W. & LIU L. (2021). Neural machine translation with monolingual translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 7307–7318, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.567.

CAO Z., LI W., LI S. & WEI F. (2018). Retrieve, rerank and rewrite : Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 152–161, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1015.

COHAN A., DERNONCOURT F., KIM D. S., BUI T., KIM S., CHANG W. & GOHARIAN N. (2018). A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings*

of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers), p. 615–621, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2097.

COHAN A. & GOHARIAN N. (2018). Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, **19**(2), 287–303. DOI : 10.1007/s00799-017-0216-8.

DOU Z.-Y., LIU P., HAYASHI H., JIANG Z. & NEUBIG G. (2021). GSum : A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4830–4842, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.384.

GU J., LU Z., LI H. & LI V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1631–1640, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1154.

JIN H., WANG T. & WAN X. (2020). Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6244–6254, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.556.

JOHNSON J., DOUZE M. & JÉGOU H. (2021). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, **7**(3), 535–547. DOI : 10.1109/TBDATA.2019.2921572.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA : Curran Associates Inc.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LIU Y., ZHANG J., WAN Y., XIA C., HE L. & YU P. (2021). HETFORMER : Heterogeneous transformer with sparse attention for long-text extractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 146–154, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.13.

LU Y., DONG Y. & CHARLIN L. (2020). Multi-XScience : A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 8068–8074, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.648.

RASLEY J., RAJBHANDARI S., RUWASE O. & HE Y. (2020). Deepspeed : System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, p. 3505–3506, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3394486.3406703.

SEE A., LIU P. J. & MANNING C. D. (2017). Get to the point : Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1073–1083, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1099.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

WANG D., LIU P., ZHENG Y., QIU X. & HUANG X. (2020). Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6209–6219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.553.

XIAO W., BELTAGY I., CARENINI G. & COHAN A. (2022). PRIMERA : Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 5245–5263, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.360.

YASUNAGA M., KASAI J., ZHANG R., FABBRI A. R., LI I., FRIEDMAN D. & RADEV D. R. (2019). Scisummnet : A large annotated corpus and content-impact models for scientific paper summarization with citation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**(01), 7386–7393. DOI : 10.1609/aaai.v33i01.33017386.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Traduction à base d'exemples du texte vers une représentation hiérarchique de la langue des signes

Élise Bertin-Lemée[1] Annelies Braffort[2] Camille Challant[2]
Claire Danet[2] Michael Filhol[2]
(1) SYSTRAN, 5 rue Feydeau, Paris, France
(2) LISN, Univ. Paris-Saclay, bât. 507, rue du Belvédère, 91405 Orsay, France
elise.bertinlemee@systrangroup.com, {annelies.braffort, camille.challant, claire.danet, michael.filhol}@lisn.upsaclay.fr

RÉSUMÉ

Cet article présente une expérimentation de traduction automatique de texte vers la langue des signes (LS). Comme nous ne disposons pas de corpus aligné de grande taille, nous avons exploré une approche à base d'exemples, utilisant AZee, une représentation intermédiaire du discours en LS sous la forme d'expressions hiérarchisées.

ABSTRACT

Example-Based Machine Translation from Text to a Hierarchical Representation of Sign Language

This paper presents an experiment in automatic translation from text to sign language (SL). As we do not have a large aligned corpus, we have explored an example-based approach, using AZee, an intermediate representation of the discourse in SL in the form of hierarchical expressions.

MOTS-CLÉS : Langue des signes, traduction à base d'exemples, représentation intermédiaire.

KEYWORDS: Sign Language, example-based translation, intermediate representation.

1 Introduction

Le travail présenté ici a été réalisé dans le cadre d'un projet qui visait à étudier des solutions d'accessibilité pour les contenus audiovisuels pour les personnes sourdes à l'aide de sous-titrage et de traduction en langue des signes (LS). Pour ce dernier objectif, les trois principales contributions ont été la constitution d'un corpus aligné de texte et de LS (Bertin-Lemée *et al.*, 2022), un système de traduction automatique (TA) du texte en une représentation formelle de la LS (Bertin-Lemée *et al.*, 2023), et un système permettant de générer des animations d'avatars signants à partir de cette représentation (Dauriac *et al.*, 2022). Après un aperçu des enjeux et des travaux récents dans le domaine, nous expliquons la méthode et les choix de conception et décrivons l'implémentation du système de traduction. Enfin, nous donnons des résultats préliminaires et discutons des questions soulevées pour l'évaluation.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 28–34.

2 Traduction du texte vers la langue des signes

Les LS sont des langues naturelles pratiquées au sein des communautés de Sourds et la Langue des Signes Française (LSF) est celle utilisée en France. Dans la suite, nous parlerons des LS en général quand les aspects abordés les concernent toutes et de la LSF lorsque cela concerne uniquement cette langue. Ce sont des langues visuo-gestuelles : une personne s'exprime en LS en utilisant de nombreuses composantes corporelles (mains et bras, mais aussi torse, épaules, tête, regard et expressions faciales) et son interlocuteur perçoit le message par le canal visuel. Le système linguistique des LS exploite ces canaux spécifiques : de nombreuses informations sont exprimées simultanément et s'organisent dans l'espace, et l'iconicité joue un rôle central et structurant à tous les niveaux (du sub-lexical au discours). À ce jour, les LS n'ont pas de système d'écriture et les rares systèmes graphiques pour la transcription, tels que HamNoSys (Hanke, 2004) ou SignWriting (Bianchini, 2014), ne décrivent que l'aspect lexical. Ceux-ci ne peuvent représenter pleinement le niveau discursif, la multilinéarité, l'utilisation de l'espace et les structures illustratives. Tous ces aspects contribuent aux défis de la TA depuis ou vers les LS.

La TA d'un texte vers une LS est un sujet de recherche assez récent et encore très peu exploré. L'approche dominante en TA est l'approche neuronale, qui s'appuie sur la disponibilité de grands volumes de données alignées (de l'ordre de plusieurs millions de phrases), non disponibles dans cette quantité pour les LS. Si des tentatives existent avec ces méthodes, elles ne donnent pas encore de résultats satisfaisants, comme l'ont montré les évaluations humaines lors du premier défi portant sur la traduction automatique de LS suisse-allemande (DSGS) vers l'allemand (Müller *et al.*, 2022).

La traduction automatique basée sur des exemples (EBMT [1]) est une autre approche fondée sur l'analogie qui utilise un corpus bilingue contenant des textes et leurs traductions (Nagao, 1984). Étant donné un texte à traduire, on sélectionne dans ce corpus des segments contenant des éléments similaires. Ces éléments sont ensuite utilisés pour traduire les éléments du texte original dans la langue cible, et ces phrases sont recombinées pour former une traduction complète. L'approche EBMT peut être mise en œuvre sur des corpus plus petits et donc envisagée dans notre cas. Les capacités de traduction restent liées à la taille du corpus, mais dans un domaine ciblé, on peut espérer obtenir des résultats de meilleure qualité que ceux obtenus actuellement par les approches neuronales.

Comme les LS n'ont pas de forme écrite, une approche courante est de procéder en deux étapes : une première étape consiste à traduire le texte en une représentation intermédiaire, et une seconde étape utilise cette représentation comme entrée d'un système de synthèse pour contrôler l'animation d'un personnage virtuel afin d'afficher le contenu en LS sous forme de vidéo.

Après une première génération d'études basées principalement sur des approches à base de règles (Veale *et al.*, 1998; Zhao *et al.*, 2000; Marshall & Safar, 2004), d'autres ont exploré des approches à base d'exemples (Morrissey & Way, 2005). Elles ont parfois été combinées avec des approches statistiques, comme par exemple De Martino *et al.* (2017) qui a développé un système qui traduit automatiquement quelques textes en portugais brésilien vers la LS brésilienne (LIBRAS), en fonction du contexte et de la fréquence d'apparition dans les traductions précédentes. À notre connaissance, ces projets n'ont donné lieu à aucune suite.

La grande majorité des projets utilisant une représentation intermédiaire de la LS, y compris les plus récents (Egea Gómez *et al.*, 2021), utilisent des séquences de gloses, chaque glose [2] représentant

1. En anglais : *example-based machine translation*.
2. Une glose est une étiquette textuelle, généralement un seul mot, qui reflète la signification du signe qu'elle représente.

une unité lexicale. Les systèmes de traduction traitent alors une séquence de tokens. Cependant, avec ce type de représentation, il est très difficile, voire impossible, de traiter les phénomènes courants de la LS, tels que l'activité non manuelle, les relations spatiales, les structures iconiques ou le rythme des mouvements. Il en résulte des animations de très mauvaise qualité, incomplètes, voire incompréhensibles. Pour cette raison, il parait indispensable d'envisager une représentation intermédiaire plus riche que de simples concaténations de gloses.

Dans certaines approches récentes de bout-en-bout (Stoll *et al.*, 2020), l'utilisation d'une représentation intermédiaire n'est pas présente. Cette approche neuronale, encore très expérimentale, génère directement des vidéos de contenus signés photoréalistes à partir d'entrées textuelles. En plus de nécessiter des corpus de très grande taille, elle n'offre pas les mêmes avantages que les avatars (anonymat dans le rendu, apparence modifiable) que nous avons choisi de privilégier.

3 Approche à base d'exemples et représentation hiérarchique

Comme nous ne disposons pas de corpus aligné de grande taille, nous avons choisi d'explorer l'approche à base d'exemples. Par ailleurs, nous avons retenu AZee comme représentation intermédiaire, une approche formelle de représentation du discours en LS (Filhol *et al.*, 2014). Celle-ci permet de définir des *règles de production*, qui associent des formes à articuler (par exemple, hausser les sourcils) à un sens identifié (par exemple, l'expression d'un doute). En les combinant, on peut construire des *expressions discursives* hiérarchiquement structurées représentant des énoncés complets, déterminant les formes à produire de manière suffisament détaillée pour permettre ensuite de contrôler l'animation d'un avatar (Challant & Filhol, 2022).

Par exemple, considérons les six règles de production suivantes identifiées pour la LSF : les trois sans argument que sont `ministre`, `environnement` et `parler`, ainsi que les trois ci-dessous comportant des arguments (en italique) :

- `info-about`(*topic*, *info*) : *info*, qui est ciblée, est donnée sur un *topic* ;
- `side-info`(*focus*, *info*) : *focus* avec une information supplémentaire (non focalisée) *info* à son sujet ;
- `nerveusement`(*sig*) : *sig* d'une manière nerveuse.

Ces règles peuvent être combinées hiérarchiquement dans l'expression suivante pour former la structure d'un énoncé signifiant "*le ministre de l'écologie parle nerveusement*" et respectant la grammaire de la LSF :

```
:info-about
  'topic
  :side-info          (*)
    'focus
    :ministre
    'info
    :environnement
  'info
  :nerveusement
    'sig
    :parler
```

Afin d'explorer l'utilisation d'une approche à base d'exemples, nous avons utilisé une banque [3] de près de 2000 alignements entre des segments de texte en français et des expressions de ce type, créée à partir du corpus parallèle français-LSF du projet.

L'approche à base d'exemples s'appuie sur l'analogie avec des exemples existants pour traduire de nouveaux contenus. Cela signifie que nous pouvons tenter de traduire une nouvelle phrase en trouvant des exemples suffisamment proches, et en remplaçant ce qui est différent. Par exemple, pour traduire la phrase "*la présidente parle nerveusement*" qui ne figure pas dans la base d'exemples, on peut partir de l'exemple "*le ministre de l'écologie parle nerveusement*" qui, lui, est présent dans la base, et substituer une traduction de "*la présidente*" à celle de "*le ministre de l'écologie*" dans le segment aligné. Dans cet exemple, "*présidente*" n'a pas de correspondance dans la phrase et est nommé "anti-match", "*ministre de l'écologie*" est nommé sa "correction". L'hypothèse est que si nous trouvons les parties correspondant à chaque anti-match dans la traduction alignée, nous pouvons tenter de les remplacer par les traductions de leurs corrections.

4 Implémentation

Pour une phrase donnée à traduire, on va chercher dans le corpus des alignements dans lesquels le segment de texte est exactement identique et on récupère les expressions alignées correspondantes. En cas d'échec, on considère tous les alignements de texte qui sont "proches" et dont les différences sont les "anti-matchs". La structure globale de la traduction dans la représentation intermédiaire est ainsi conservée, dans laquelle on va pouvoir faire les substitutions.

Dans l'exemple précédent, on peut considérer que le segment "*le ministre de l'écologie parle nerveusement*" est proche de "*la présidente parle nerveusement*". L'anti-match unique $\bar{m}_1$ est "*le ministre de l'écologie*" et sa correction c_1 est "*la présidente*". Notre approche est alors : (1) d'identifier la sous-expression marquée `(*)` ci-dessus comme le nœud correspondant à la traduction de $\bar{m}_1$; et (2) d'y substituer une traduction de c_1. Pour ces deux tâches, on utilise récursivement le même algorithme. Pour (1) on génère les traductions possibles de $\bar{m}_1$ en vue d'y trouver `(*)`, et pour (2) on génère directement celles de c_1, qui pourront être substituées à `(*)`.

Un des problèmes de cette approche est celui de l'échec de la traduction, qui est d'autant plus susceptible de se produire que le corpus d'alignements d'exemples est petit. Dans de tels cas, nous avons recours à une solution de repli où nous décomposons la requête en une partition de plus petits morceaux de texte, que nous traduirons séparément et concaténerons dans le résultat avec pour seule règle le suivi de l'ordre en source. Cette stratégie de repli produit une LS de moins bonne qualité, et équivaut en fait à une traduction littérale (mot à mot) si elle est utilisée systématiquement. Mais elle permet de juxtaposer des morceaux de LS plus importants et donc plus complets et plus fluides sans recourir à la simple concaténation d'unités uniquement lexicales.

L'implémentation pratique de l'algorithme s'appuie sur plusieurs modules de traitement de texte pour trouver les meilleures correspondances dans le corpus existant.
Pour permettre la mise en correspondance, l'anti-match et les partitions sous-phrastiques, la tokenisation au niveau du mot est d'abord effectuée par Open-NMT Tokenizer [4] et une certaine flexibilité est permise lors de la recherche de segments correspondants avec la ponctuation et les articles.

3. L'ensenble du corpus est disponible ici : https://www.ortolang.fr/market/corpora/rosetta-lsf
4. https://github.com/OpenNMT/Tokenizer

Ensuite, le défi principal est de définir quel type de "similarité" dans la langue source peut produire les meilleurs candidats pour la génération de la langue cible. La sémantique et la syntaxe entrent en jeu pour déterminer les éléments similaires à remplacer ou à traduire séparément. En pratique, nous nous appuyons sur deux types d'analyse de texte à différentes étapes de l'algorithme.

Pour trouver les meilleurs anti-matches dans la base de données actuelle et les remplacer par des corrections, nous utilisons l'appariement de chaîne de caractères et considérons comme candidats tous les alignements qui ont des tokens en commun avec le texte soumis. Les meilleures correspondances ont été définies empiriquement comme étant celles qui ont le maximum de tokens en commun, ainsi que la longueur minimale en nombre de tokens ou la meilleure proportion de tokens similaires dans l'ensemble des tokens. Pour l'instant, ce choix reste arbitraire et mériterait une étude comparative.

Lorsque les approches d'appariement et d'anti-match échouent, nous avons recours à des partitions déterminées en naviguant dans l'arbre de dépendance syntaxique obtenu à l'aide de spaCy [5], une bibliothèque avec des modèles prêts à l'emploi et des chaînes de traitement optimisées pour les langues naturelles. L'analyse syntaxique par dépendance n'explore pas toutes les partitions possibles d'une phrase mais limite l'exploration à des morceaux syntaxiquement valides.

5 Test et discussion

Pour tester notre système, nous avons construit un jeu de test en créant des phrases mélangeant des segments de différentes phrases de notre corpus, pour étudier les résultats produits. Notre jeu de test est composé de 15 phrases. Par exemple, la phrase "*Recul de l'âge légal à la retraite : c'est ce que proposent les retraités pour leurs enfants*" a été créée à partir des phrases suivantes du corpus :

— "***Recul de l'âge légal à la retraite :*** *"Il ne faut pas prendre les Français pour des canards sauvages", lance Valérie Pécresse.*"
— "*Des routes nationales bientôt privatisées ?* ***C'est ce que proposent*** *les sociétés d'autoroutes dans une note interne.*"
— "*Solidarité : une ancienne abbaye accueille* ***des retraités***"
— "*Au Japon, des dizaines de pères français se battent désespérément* ***pour*** *voir* ***leurs enfants.***"

Grâce à ce jeu de tests, nous avons pu faire valider, par le biais de focus groups avec des locuteurs de LSF, que notre approche est préférable à une approche basée sur une simple concaténation de signes lexicaux. En effet, les structures spécifiques à la LSF peuvent être trouvées dans les traductions finales, ce qui n'est pas le cas lorsque la langue est réduite à une séquence de gloses.

En outre, l'approche produit des résultats présentant une certaine forme de créativité. En LSF, les paraphrases ou les ajouts sont couramment utilisés, et font d'ailleurs partie de notre corpus tel qu'il a été livré initialement par le traducteur au moment de la création du corpus vidéo. Ces éléments ont été alignés par la suite comme exemples, et apparaissent donc fréquemment dans les traductions générées, même si ce n'est pas toujours strictement nécessaire. Par exemple "*Alsace*" peut être signé par un seul signe, mais il est aussi exprimé dans notre corpus par une expression bien plus complexe impliquant des référencements spatialisés (zone à l'Est de la France placée sur un plan vertical), typique de la LSF quand aucun contexte n'existe encore.

De plus, la sortie de l'algorithme est un ensemble de traductions (construites à partir des différentes

5. https://spacy.io

combinaisons de substitution), et pas nécessairement une expression unique. Cela rend compte, d'une certaine manière, de la réalité de la tâche de traduction. Dans notre jeu de test, le nombre de traductions proposées pour une requête varie de 1 à 12 (moyenne : 4).

6 Conclusion et perspectives

Nous avons présenté une nouvelle idée de système de TA du texte vers la LSF, utilisant une approche à base d'exemples et une représentation hiérarchique de la LS, ainsi qu'un algorithme d'appariement, substitution et concaténation. Le corpus utilisé contient des alignements de textes français et leurs traductions en LSF décrites selon cette représentation. Un prototype a été réalisé et testé sur quelques exemples, fournissant ainsi une preuve de concept. Les capacités de ce système et la taille du corpus doivent encore être étendues avant de pouvoir effectuer de véritables évaluations. Mais nous pouvons d'ores et déjà souligner que l'évaluation d'un tel système ne sera pas facile, puisqu'il propose une traduction d'une langue vers une représentation d'une autre langue, non lisible directement.

Les métriques habituellement utilisées, qu'elles soient quantitatives ou qualitatives, sont adaptées aux cas où les langues source et cible sont textuelles. Dans notre cas, la cible n'est pas directement la LSF, mais une représentation formelle. En outre, cette représentation est utilisée pour générer des animations qui, certes, sont directement "lisibles" par les locuteurs de LSF, mais qui nécessitent de leur côté des phases d'évaluations qui ne sont pas liées aux aspects linguistiques mais plutôt à l'aspect de l'avatar, aux mouvements et à leur degré de bio-réalisme. La mise en place d'un protocole d'évaluation robuste et complet est clairement un sujet d'étude à part entière, qui devra être abordé dans un avenir proche.

Remerciements

Ce travail a été financé par le projet d'investissement Bpifrance "Grands défis du numérique", dans le cadre du projet ROSETTA (RObot for Subtitling and intElligent adapTed TranslAtion). Nous remercions Noémie Churlet, Raphaël Bouton et Media'Pi ! pour leur engagement dans ce projet, qui n'aurait pas eu la même validité et le même impact sans eux.

Références

BERTIN-LEMÉE E., BRAFFORT A., CHALLANT C., DANET C., DAURIAC B., FILHOL M., MARTINOD E. & SEGOUAT J. (2022). Rosetta-LSF : an Aligned Corpus of French Sign Language and French for Text-to-Sign Translation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.

BERTIN-LEMÉE E., BRAFFORT A., CHALLANT C., DANET C. & FILHOL M. (2023). Example-Based Machine Translation from Text to a Hierarchical Representation of Sign Language. In *The 24th Annual Conference of The European Association for Machine Translation*.

BIANCHINI C. S. (2014). *Analyse métalinguistique de l'émergence d'un système d'écriture des langues des signes : SignWriting et son application à la langue des signes italienne (LIS)*. Thèse de doctorat. Thèse de doctorat dirigée en Sciences du langage de l'Université Paris 8.

CHALLANT C. & FILHOL M. (2022). A First Corpus of AZee Discourse Expressions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, p. 1560-1565, Marseille, France.

DAURIAC B., BRAFFORT A. & BERTIN-LEMÉE E. (2022). Example-based Multilinear Sign Language Generation from a Hierarchical Representation. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology : The Junction of the Visual and the Textual : Challenges and Perspectives*, p. 21–28, Marseille, France : European Language Resources Association.

DE MARTINO J. M., SILVA I. R., BOLOGNINI C. Z., COSTA P. D. P., KUMADA K. M. O., CORADINE L. C., DA SILVA BRITO P. H., DO AMARAL W. M., BENETTI Â. B., POETA E. T. *et al.* (2017). Signing Avatars : Making Education More Inclusive. *Universal access in the information society*, **16**(3), 793–808.

EGEA GÓMEZ S., MCGILL E. & SAGGION H. (2021). Syntax-aware Transformers for Neural Machine Translation : The Case of Text to Sign Gloss Translation. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, p. 18–27 : INCOMA Ltd.

FILHOL M., HADJADJ M. & CHOISIER A. (2014). Non-Manual Features : The Right to Indifference. In *International Conference on Language Resources and Evaluation*, p. 49–54, Reykjavik, Iceland.

HANKE T. (2004). Hamnosys - representing sign language data in language resources and language processing contexts. In O. STREITER & C. VETTORI, Éds., *LREC 2004, Workshop proceedings : Representation and processing of sign languages.* : Paris : ELRA.

MARSHALL I. & SAFAR E. (2004). Sign language generation in an ALE HPSG. In S. MÜLLER, Éd., *Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar, Center for Computational Linguistics, Katholieke Universiteit Leuven*, p. 189–201, Stanford, CA : CSLI Publications. DOI : 10.21248/hpsg.2004.11.

MORRISSEY S. & WAY A. (2005). An Example-Based Approach to Translating Sign Language. In *Workshop on Example-Based Machine Translation, MT SUMMIT*, p. 109–116, Phuket, Thailand : Asia-Pacific Association for Machine Translation, Tokyo.

MÜLLER M., EBLING S., AVRAMIDIS E., BATTISTI A., BERGER M., BOWDEN R., BRAFFORT A., CIHAN CAMGÖZ N., ESPAÑA-BONET C., GRUNDKIEWICZ R., JIANG Z., KOLLER O., MORYOSSEF A., PERROLLAZ R., REINHARD S., RIOS A., SHTERIONOV D., SIDLER-MISEREZ S. & TISSI K. (2022). Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, p. 744–772, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

NAGAO M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, p. 173–180. USA : Elsevier North-Holland, Inc.

STOLL S., CAMGOZ N. C., HADFIELD S. & BOWDEN R. (2020). Text2sign : Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, **128**(4), 891–908.

VEALE T., CONWAY A. & COLLINS B. (1998). The Challenges of Cross-modal Translation : English-to-Sign-Language Translation in the Zardoz System. *Machine Translation*, **13**(1), 81–106.

ZHAO L., KIPPER K., SCHULER W., VOGLER C., BADLER N. & PALMER M. (2000). A Machine Translation System from English to American Sign Language. In *Conference of the Association for Machine Translation in the Americas*, p. 54–67 : Springer.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Annotation Linguistique pour l'Évaluation de la Simplification Automatique de Textes*

Rémi Cardon, Adrien Bibal, Rodrigo Wilkens, David Alfter,
Magali Norré, Adeline Müller, Patrick Watrin, Thomas François
CENTAL (IL&C), UCLouvain, Place Montesquieu 3, 1348 Louvain-la-Neuve, Belgique
prenom.nom@uclouvain.be

RÉSUMÉ

L'évaluation des systèmes de simplification automatique de textes (SAT) est une tâche difficile, accomplie à l'aide de métriques automatiques et du jugement humain. Cependant, d'un point de vue linguistique, savoir ce qui est concrètement évalué n'est pas clair. Nous proposons d'annoter un des corpus de référence pour la SAT, ASSET, que nous utilisons pour éclaircir cette question. En plus de la contribution que constitue la ressource annotée, nous montrons comment elle peut être utilisée pour analyser le comportement de SARI, la mesure d'évaluation la plus populaire en SAT. Nous présentons nos conclusions comme une étape pour améliorer les protocoles d'évaluation en SAT à l'avenir.

ABSTRACT

Linguistic Corpus Annotation for Automatic Text Simplification Evaluation

Evaluating automatic text simplification (ATS) systems is a difficult task that is either performed by automatic metrics or user-based evaluations. However, from a linguistic point-of-view, it is not always clear on what bases these evaluations operate. We propose to annotate the ASSET corpus to shed more light on ATS evaluation. In addition to contributing with this resource, we show how it can be used to analyze SARI's relation to linguistic operations. We present our insights as a step to improve ATS evaluation protocols in the future.

MOTS-CLÉS : evaluation, resource, automatic text simplification, annotation.

KEYWORDS: évaluation, ressource, simplification automatique de textes, annotation.

1 Introduction

La simplification automatique de textes (SAT) consiste à rendre des textes plus accessibles pour un public donné. Plusieurs états de l'art sur le sujet ont été publiés ces dernières années (Saggion, 2017; Al-Thanyyan & Azmi, 2021; Štajner, 2021). La SAT est principalement étudiée avec des approches par apprentissage profond (Nisioi *et al.*, 2017; Alva-Manchego *et al.*, 2020b; Cooper & Shardlow, 2020), mais d'autres travaux poursuivent les recherches par le biais de systèmes à base de règles (Evans & Orasan, 2019; Wilkens *et al.*, 2020). Un des verrous majeurs du domaine est l'évaluation des systèmes de SAT (Grabar & Saggion, 2022). Il existe deux approches courantes : le jugement humain et l'évaluation automatique. Dans la première approche, il est demandé à des personnes de

∗. Cet article est une adaptation d'un article précédemment publié (Cardon *et al.*, 2022)

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 35–48.

noter la sortie d'un système selon trois critères : grammaticalité, préservation du sens, et simplicité. Pour l'évaluation automatique, les métriques les plus courantes sont : BLEU (Papineni *et al.*, 2002), empruntée à la traduction automatique ; SARI (Xu *et al.*, 2016a), spécifiquement proposée pour la SAT ; et la formule de lisibilité Flesch-Kincaid (Kincaid *et al.*, 1975). Bien que ces métriques ne soient pas idéales (Sulem *et al.*, 2018a; Alva-Manchego *et al.*, 2021; Tanprasert & Kauchak, 2021), leur facilité d'utilisation rend leur application répandue. BLEU et SARI nécessitent des simplifications de référence et sont décrites comme étant plus fiables à mesure que le nombre de références augmente. Comme différents publics ont différents besoins en termes de simplification (Rennes *et al.*, 2022), il est crucial de s'assurer qu'un jeu de référence reflète ces besoins. D'autres métriques sont moins fréquemment utilisées, tel que BertScore (Zhang *et al.*, 2020) – initialement conçue pour la génération automatique – et SAMSA (Sulem *et al.*, 2018b) – conçue pour la SAT, mais difficile à utiliser car elle nécessite une annotation sémantique.

En anglais, trois corpus sont couramment utilisés pour l'évaluation : TurkCorpus (Xu *et al.*, 2016b), ASSET (Alva-Manchego *et al.*, 2020a) et Newsela (Xu *et al.*, 2015). Les deux premiers ont les mêmes phrases sources, avec différentes simplifications obtenues par production participative (*crowdsourcing*). Pour les autres langues, qui reçoivent moins d'attention de la communauté, les systèmes sont évalués avec des corpus *adhoc* (Kodaira *et al.*, 2016; Cardon & Grabar, 2020; Anees & Abdul Rauf, 2021; Spring *et al.*, 2021; Todirascu *et al.*, 2022). La manière dont les métriques automatiques sont liées aux opérations de simplification n'est pas connue. Des travaux récents (Vásquez-Rodríguez *et al.*, 2021) ont exploré la relation de ces métriques avec les opérations computationnelles basiques comme l'ajout et la suppression, mais aucun travail n'a étudié cette relation avec les opérations de simplification présentes dans des typologies linguistiques (Brunato *et al.*, 2022; Gala *et al.*, 2020; Amancio & Specia, 2014). Dans cet article, nous souhaitons étudier l'impact du type d'opérations linguistiques sur les métriques automatiques d'évaluation de la SAT. Pour cela, nous avons annoté le corpus ASSET avec les opérations linguistiques qu'il contient.

Nous présentons un état de l'art sur les typologies de transformations pour la simplification (Section 2). Ensuite nous décrivons le processus d'annotation et la ressource produite (Section 3) puis des analyses menées sur SARI avec notre ressource (Section 4). Enfin nous concluons (Section 5) par une synhèse des enseignements obtenus suite à nos expériences.

2 État de l'art

Avant que les méthodes neuronales soient au centre de la recherche en SAT, les typologies d'opérations étaient nécessaires au développement des systèmes, car elles représentaient la base conceptuelle des approches à base de règles. Comme présenté par Siddharthan (2014), les premiers systèmes de SAT s'occupaient de syntaxe (Chandrasekar *et al.*, 1996; Dras, 1999; Brouwers *et al.*, 2014) et décrivaient les typologies d'opérations syntaxiques qui étaient visées pour simplifier la structure des phrases. Actuellement, nous distinguons deux ensembles pour les typologies d'opérations de simplification : l'un basé sur la description linguistique et l'autre basé sur l'édition de chaînes de *tokens*.

2.1 Opérations linguistiques

Le premier ensemble de typologies vise à décrire les opérations linguistiques mises en œuvre lors de la simplification. Cela a été étudié pour différentes langues : l'espagnol (Bott & Saggion, 2014),

l'italien (Brunato *et al.*, 2014, 2022), le français (Koptient *et al.*, 2019; Gala *et al.*, 2020), le portugais du Brésil (Caseli *et al.*, 2009), le basque (Gonzalez-Dios *et al.*, 2018) et l'anglais (Amancio & Specia, 2014). Bien que des opérations soient communes à ces typologies, comme le passage de la voix passive à la voix active, ces typologies présentent des catégories distinctes comme la « *proximization* » (Bott & Saggion, 2014) – faire en sorte que le texte s'adresse au lecteur – ou la « spécification » (Koptient *et al.*, 2019) – conserver un terme difficile et y accoler une explication. Notons que ces deux exemples dépendent du genre de texte : il est peu probable de trouver la *proximization* dans la simplification d'articles encyclopédiques, et il est attendu que la spécification ait lieu dans des textes qui contiennent du lexique spécialisé ou technique. Ces typologies ont été utilisées de manière descriptive, pour renseigner sur les corpus ainsi que sur les pratiques humaines de simplification.

2.2 Éditions de chaînes de tokens

Dans le cadre de ce deuxième ensemble de typologies, les phrases sont vues comme des chaînes de *tokens*, et la simplification consiste à modifier l'agencement de ces *tokens*. Ici, les opérations sont donc décrites comme des modifications à des chaînes de *tokens*, il s'agit par exemple de la suppression, de l'ajout, ou du maintien. Cet angle a été exploré presque exclusivement pour l'anglais (Coster & Kauchak, 2011; Alva-Manchego *et al.*, 2017, 2020a; Vásquez-Rodríguez *et al.*, 2021), avec une exception récente pour l'italien (Brunato *et al.*, 2022). Ces typologies ont été mises en œuvre à diverses fins. Comme pour les typologies linguistiques, elles ont servi à analyser des corpus (Alva-Manchego *et al.*, 2020a). Elles ont aussi servi à étudier la relation des distances entre les chaînes avec les scores attribués par les métriques automatiques (Vásquez-Rodríguez *et al.*, 2021). Certains systèmes de SAT incorporent ce type d'opérations dans leur architecture (Alva-Manchego *et al.*, 2017; Dong *et al.*, 2019; Agrawal *et al.*, 2021). La métrique d'évaluation SARI intègre ce type d'opération dans sa formule : avec les sous-composants KEEP, ADD et DELETE (cf. Section 4 pour plus de détails).

3 Annotation

Deux des trois corpus d'évaluation décrits en section 1 sont librement accessibles : TurkCorpus et ASSET. Nous retenons ASSET, qui a été décrit indépendamment comme meilleur que TurkCorpus (Vásquez-Rodríguez *et al.*, 2021). Cette section présente la typologie d'opérations que nous utilisons pour son annotation (Section 3.1). Nous décrivons le processus d'annotation (Section 3.2), puis nous décrivons la ressource produite (Section 3.3).

3.1 Typologie

Les travaux présentés en section 2.1 proposent des typologies basées sur des analyses manuelles de corpus. Nous construisons la nôtre sur ces travaux. En conséquence, nous n'introduisons aucune nouvelle opération. Nous ne retenons pas les opérations spécifiques au genre de texte, comme celles mentionnées en section 2.1 (*proximization* et spécification), afin d'avoir un ensemble d'opérations génériques. La liste ainsi obtenue est présentée ci-dessous, avec le nom des opérations et leur identifiant dans la ressource. Le nom de certaines opérations est suffisamment descriptif, d'autres opérations sont brièvement clarifiées. Nous présentons les opérations en deux temps : premièrement les opérations qui peuvent directement correspondre à des opérations computationnelles. La correspondance est la

suivante : INSERT (parfois dénotée ADD dans la littérature), DELETE et MOVE, traduits plus bas, sont déjà utilisés tels quels pour les opérations computationnelles. Toutes les autres catégories sont des substitutions.

— **Déplacement** (move)
— **Insérer/Supprimer proposition** (inprop, delprop)
— **Insérer/Supprimer modifieur** (inmod, delmod). Notre définition de modifieur couvre les modifieurs de mots (p. ex. un adjectif qualifiant un nom) et les modifieurs de phrases (p. ex. un complément circonstanciel).
— **Insérer/Supprimer pour la cohérence** (incst, delcst). Toute insertion ou suppression nécessaire suite à une autre opération pour que la phrase reste grammaticale.
— **Insérer/Supprimer autre** (inoth, deloth). Toute insertion ou suppression n'appartenant pas à une des catégories précédentes.
— **Substitution par synonymie** (synonym)
— **Substitution par hyperonymie** (hyperonym)
— **Substitution par hyponymie** (hyponym)
— **Substitution du singulier par le pluriel** (s2p)
— **Substitution du pluriel par le singulier** (p2s)
— **Substitution par pronominalisation** (pron)
— **Substitution par résolution d'antécédent** (fromPron)
— **Modification des traits verbaux** (verbf). Tout changement de mode ou temps d'un verbe.

Deuxièmement, les opérations qui sont le résultat de combinaisons d'opérations computationnelles, ou qui sont trop complexes pour établir une correspondance stable entre les deux types d'opérations.

— **Voix Active vers passive** (a2p)
— **Voix passive vers active** (p2a)
— **Changement de partie du discours** (POSchange)
— **Découpage de phrases** (split)
— **Regroupement de phrases** (merge)
— **Vers forme impersonnelle** (toImp)
— **Vers forme personnelle** (fromImp)
— **Affirmation vers négation** (a2n)
— **Négation vers affirmation** (n2a)

Nous ajoutons également une étiquette **Simplification erronée** (err). Bien que nous n'évaluions pas la simplicité, cela permet de signaler, pendant l'annotation, des erreurs manifestes de grammaticalité ou de préservation du sens, qui rendent la simplification non-désirée en tant que référence dans un protocole d'évaluation.

3.2 Processus

Nous utilisons YAWAT (Germann, 2008) pour l'annotation, un outil déjà utilisé dans ce cadre auparavant (Koptient *et al.*, 2019). [1] La typologie construite fut la base de la rédaction du guide d'annotation. Quatre personnes (travaillant dans la recherche en TAL) ont annoté les mêmes 50 couples de phrases d'ASSET. Cela a servi à (1) évaluer la clarté du guide, (2) former à l'utilisation de l'outil et (3) discuter des points d'amélioration du guide. Cette troisième étape a permis de discuter des cas difficiles et de comment les traiter [2] pour atteindre le consensus. Les discussions n'ont pas

1. Un outil plus récent et commode existe, TS–ANNO (Stodden & Kallmeyer, 2022) mais n'était pas encore disponible au moment de notre annotation.
2. La plupart de ces cas servent d'exemples dans le guide d'annotation.

entraîné de modifications de la typologie. Nous avons réitéré cette étape 2 fois avec 25 nouveaux couples de phrases à chaque fois. Une fois le guide finalisé, nous avons engagé [3] cinq étudiants de master en TAL pour compléter l'équipe d'annotation.

La dernière étape avant l'annotation du corpus intégral – par l'équipe de neuf personnes – fut d'annoter 50 nouveaux couples de phrases d'ASSET. Cela a servi de base au calcul de l'accord inter-annotateurs (Davies & Fleiss, 1982). L'accord est calculé au niveau des tokens et calculé séparément pour les deux côtés du corpus (original et simplifié). L'accord est de 0,61 pour le côté source et de 0,68 pour le côté cible. Nous l'avons également calculé en fusionnant toutes les insertions en une seule catégorie, et toutes les suppressions en une seule catégorie. De la sorte, l'accord est de 0,74 côté source et 0,72 côté cible. Cela indique un compromis entre la granularité de l'annotation et l'accord que l'on peut en obtenir. Le corpus intégralement annoté porte le nom « ASSET$_{ann}$ ».

3.3 Description de la ressource

Cette section décrit le résultat de l'annotation, le corpus ASSET$_{ann}$. Le jeu de test d'ASSET contient 3 590 couples de phrases (359 phrases simplifiées 10 fois chacune). Pendant l'annotation, nous avons trouvé 19 couples de phrases identiques, et 227 simplifications erronées concernant 157 phrases d'origine. ASSET$_{ann}$ contient 3 323 couples de phrases annotées. Un total de 12 827 opérations y sont annotées. *Synonym* est l'opération la plus fréquente, avec 14 % du nombre total d'opérations. Sept opérations (*synonym*, *delcst*, *deloth*, *incst*, *delmod*, *move* et *delprop*) représentent 70 % des opérations à la fois dans le *gold* et dans ASSET$_{ann}$.

4 Analyse de SARI

Nous utilisons ASSET$_{ann}$ pour analyser le comportement de SARI et ses sous-composants en relation avec les opérations de simplification. SARI utilise trois sous-composants, dont la moyenne représente le score final. Ces sous-composants sont *keep*, *add* et *delete*. Pour chaque sous-composant, le score F1 est calculé entre les transformations appliquées de la phrase d'origine pour arriver à la référence ou aux références, et les transformations appliquées de la phrase d'origine pour arriver à la phrase à évaluer. Ces transformations sont observées en n-grammes de *tokens*, où *n* va de 1 à 4 [4] :

$$F1(n, sc) = \frac{2 * prec_{sc}(n) * rappel_{sc}(n)}{prec_{sc}(n) + rappel_{sc}(n)}$$

$$\text{SARI} = \frac{1}{3} \sum_{sc \in \{keep, add, del\}} \frac{1}{k} \sum_{n=1}^{k} F1(n, sc).$$

Nous représentons les couples de phrases par le nombre d'occurrences de chaque opération dans l'annotation. En observant la relation entre la présence d'opérations spécifiques et le score global SARI, nous trouvons une faible corrélation. De plus, même les opérations en correspondance avec les sous-composants de SARI (insertions et suppression) ne sont pas corrélées avec les scores SARI (voir

3. À un taux horaire 25 % supérieur au revenu minimum national.
4. Cette description de SARI correspond à son implémentation dans EASSE (Alva-Manchego *et al.*, 2019), qui est l'outil que nous avons utilisé lors de ce travail.

Table 2 en annexe A), bien qu'elles soient légèrement corrélées avec les scores des sous-composants (voir Table 3 en annexe A).

Pour aller plus loin, nous analysons comment la combinaison des opérations permet de prédire le score SARI, par paire de phrases. Un modèle de régression Lasso (Tibshirani, 1996) avec optimisation des hyperparamètres a ainsi été entraîné et évalué avec R^2, estimé via une validation croisée à 10 plis. Le R^2 de notre modèle est de 0 pour la prédiction du score SARI, ce qui indique que le modèle ne peut pas prédire mieux qu'en utilisant la moyenne. Il ne trouve donc aucun lien entre les opérations de simplification annotées et le score SARI. Nous obtenons le même résultat avec d'autres algorithmes, tels que les arbres de régression (Breiman *et al.*, 1984), les *random forests* (Breiman, 2001) et un perceptron multi-couches (Hinton, 1989) [5].

Cependant, prédire les sous-composants de SARI semble partiellement possible avec Lasso, avec un R^2 moyen de 0,24, 0,03 et 0,23 respectivement pour KEEP, ADD et DEL. La Table 1 présente ainsi les coefficients d'un modèle Lasso entraîné sur le corpus entier pour prédire les sous-composants de SARI. Ces coefficients ont été obtenus avec un R^2 de 0,25, 0,05 et 0,24 pour KEEP, ADD and DEL respectivement. Nous pouvons déjà observer que beaucoup d'opérations ont un coefficient de 0, indiquant qu'elles n'ont pas d'effet sur les sous-composants de SARI.

Il apparaît que bien que SARI montre un certain degré de relation avec les opérations linguistiques, procéder à la moyenne des scores des sous-composants efface cette information. Cela met en lumière deux problèmes de SARI. Premièrement, cette métrique a une variance très faible et n'est pas sensible aux différences entre les sorties des systèmes. Deuxièmement, comme SARI requiert des références, faire la moyenne sur plusieurs références (9 par phrase dans nos expériences) renforce la faible variance. Cette observation est aussi vraie pour les sous-composants, ce qui explique les scores R^2 plutôt faibles.

5 Bilan et perspectives

Nous avons annoté ASSET avec pour objectif d'analyser l'évaluation de la SAT à la lumière des informations linguistiques. Cela nous a permis de porter plusieurs contributions, en plus de la ressource annotée. Concernant l'analyse des pratiques courantes d'évaluation automatique de la SAT, nous avons montré que les sous-composants de SARI peuvent informer sur les opérations linguistiques présentes dans les références et qui apparaissent dans la sortie d'un système, et que cette information est perdue lors du calcul de la moyenne pour produire un score unique. Cela constitue un argument en faveur de rapporter les scores des sous-composants lors de l'évaluation, comme certains travaux commencent à le faire (Zhao *et al.*, 2020; Tanprasert & Kauchak, 2021). Les analyses que nous avons faites encouragent à explorer les liens entre les opérations linguistiques et les opérations computationnelles. Nous pensons qu'ASSET$_{ann}$ et nos expériences représentent une première étape vers des pratiques d'évaluation qui intégreraient ces aspects.

Lors de ce travail, nous avons produit une version annotée du jeu de test d'ASSET. Cette ressource est annotée par 9 personnes, utilisant une typologie que nous proposons en nous basant sur des travaux antérieurs dans plusieurs langues. À partir de l'annotation, nous avons pu nettoyer le jeu de test en excluant 227 couples de phrases avec des erreurs manifestes et produire une version vérifiée manuellement de ces données. Nous avons également mené une analyse poussée de SARI et ses

5. Toutes les expériences ont été menées avec Scikit-learn (Pedregosa *et al.*, 2011)

sous-composants et avons trouvé des liens assez ténus entre ces derniers et les opérations linguistiques. Nous voyons ces résultats comme une direction prometteuse pour l'amélioration de l'évaluation automatique de la SAT, en explorant davantage la relation entre les différents types d'opérations. La ressource décrite ici est librement disponible en ligne. [6]

6 Remerciements

Nous exprimons nos remerciements à Nils Bouckaert, Elena Cao, Angela Kasparian, Melanie Johanns and Luca Matarelli, pour leur aide lors de l'annotation des données. Merci également à Damien de Meyere et Hubert Naets pour leur aide avec YAWAT.

Enfin, nous remercions les relecteurs anonymes pour leurs commentaires et suggestions qui ont contribué à améliorer la qualité de cet article.

Rémi Cardon est financé par le programme *FSR Incoming Postdoc Fellowship* du FSR - Université Catholique de Louvain. Adrien Bibal est financé par la région Wallonne avec un fonds Win2Wal. Rodrigo Wilkens est financé par une convention de recherche avec France Education Internatinal (FEI). David Alfter est financé par le Fonds de la Recherche Scientifique de Belgique (F.R.S-FNRS), référence MIS/PGY F.4518.21.

Références

AGRAWAL S., XU W. & CARPUAT M. (2021). A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3757–3769, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.330.

AL-THANYYAN S. S. & AZMI A. M. (2021). Automated text simplification : A survey. *ACM Computing Surveys (CSUR)*, **54**(2), 1–36.

ALVA-MANCHEGO F., BINGEL J., PAETZOLD G., SCARTON C. & SPECIA L. (2017). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 295–305, Taipei, Taiwan : Asian Federation of Natural Language Processing.

ALVA-MANCHEGO F., MARTIN L., BORDES A., SCARTON C., SAGOT B. & SPECIA L. (2020a). ASSET : A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, p. 4668–4679.

ALVA-MANCHEGO F., MARTIN L., SCARTON C. & SPECIA L. (2019). EASSE : Easier automatic sentence simplification evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing : System Demonstrations*, p. 49–54, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-3009.

ALVA-MANCHEGO F., SCARTON C. & SPECIA L. (2020b). Data-driven sentence simplification : Survey and benchmark. *Computational Linguistics*, **46**(1), 135–187.

6. https://github.com/remicardon/assetann

ALVA-MANCHEGO F., SCARTON C. & SPECIA L. (2021). The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, **47**(4), 861–889.

AMANCIO M. & SPECIA L. (2014). An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, p. 123–130, Gothenburg, Sweden : Association for Computational Linguistics. DOI : 10.3115/v1/W14-1214.

ANEES Y. & ABDUL RAUF S. (2021). Automatic sentence simplification in low resource settings for Urdu. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, p. 60–70, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.nlp4posimpact-1.7.

BOTT S. & SAGGION H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, **48**, 93–120.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

BREIMAN L., FRIEDMAN J. H., OLSHEN R. A. & STONE C. J. (1984). *Classification and Regression Trees*. Belmont, CA : Wadsworth International Group.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2014). Syntactic sentence simplification for french. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL 2014*, p. 47–56.

BRUNATO D., DELL'ORLETTA F. & VENTURI G. (2022). Linguistically-based comparison of different approaches to building corpora for text simplification : A case study on italian. *Frontiers in Psychology*, **13**.

BRUNATO D., DELL'ORLETTA F., VENTURI G. & MONTEMAGNI S. (2014). Defining an annotation scheme with a view to automatic text simplification. In *Proceedings of the Italian Conference on Computational Linguistics and of the International Workshop EVALITA*, p. 87–92.

CARDON R., BIBAL A., WILKENS R., ALFTER D., NORRÉ M., MÜLLER A., PATRICK W. & FRANÇOIS T. (2022). Linguistic corpus annotation for automatic text simplification evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 1842–1866, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

CARDON R. & GRABAR N. (2020). French biomedical text simplification : When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 710–716, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.62.

CASELI H. M., PEREIRA T. F., SPECIA L., PARDO T. A., GASPERIN C. & ALUÍSIO S. M. (2009). Building a brazilian portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, **41**, 59–70.

CHANDRASEKAR R., DORAN C. & SRINIVAS B. (1996). Motivations and methods for text simplification. In *The 16th International Conference on Computational Linguistics*.

COOPER M. & SHARDLOW M. (2020). CombiNMT : An exploration into neural text simplification models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 5588–5594, Marseille, France : European Language Resources Association.

COSTER W. & KAUCHAK D. (2011). Simple English Wikipedia : A new text simplification task. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 665–669.

DAVIES M. & FLEISS J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, **38**, 1047.

DONG Y., LI Z., REZAGHOLIZADEH M. & CHEUNG J. C. K. (2019). EditNTS : An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3393–3402, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1331.

DRAS M. (1999). *Tree adjoining grammar and the reluctant paraphrasing of text*. Macquarie University Sydney.

EVANS R. & ORASAN C. (2019). Sentence simplification for semantic role labelling and information extraction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

GALA N., TODIRASCU A., BERNHARD D., WILKENS R. & MEYER J.-P. (2020). Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés. *SHS Web of Conferences*, **78**, 14006.

GERMANN U. (2008). Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL : HLT Demo Session*, p. 20–23, Columbus, Ohio : Association for Computational Linguistics.

GONZALEZ-DIOS I., ARANZABE M. J. & DÍAZ DE ILARRAZA A. (2018). The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation*, **52**(1), 217–247.

GRABAR N. & SAGGION H. (2022). Evaluation of automatic text simplification : Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 453–463, Avignon, France : ATALA.

HINTON G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, **40**(1), 185–234.

KINCAID J., FISHBURNE R., RODGERS R. & CHISSOM B. (1975). *Derivation of new readability formulas for navy enlisted personnel*. Rapport interne, n°8-75, Research Branch Report.

KODAIRA T., KAJIWARA T. & KOMACHI M. (2016). Controlled and balanced dataset for Japanese lexical simplification. In *Proceedings of the ACL Student Research Workshop*, p. 1–7, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-3001.

KOPTIENT A., CARDON R. & GRABAR N. (2019). Simplification-induced transformations : typology and some characteristics. In *Proceedings of the BioNLP Workshop and Shared Task*, p. 309–318.

NISIOI S., ŠTAJNER S., PONZETTO S. P. & DINU L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 2 : Short papers)*, p. 85–91.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

RENNES E., SANTINI M. & JONSSON A. (2022). The swedish simplification toolkit : – designed with target audiences in mind. In *Proceedings of The 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, p. 31–38, Marseille, France : European Language Resources Association.

SAGGION H. (2017). Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, **10**(1), 1–137.

SIDDHARTHAN A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, **165**(2), 259–298.

SPRING N., RIOS A. & EBLING S. (2021). Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, p. 1339–1349, Held Online : INCOMA Ltd.

ŠTAJNER S. (2021). Automatic text simplification for social good : Progress and challenges. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP*, p. 2637–2652.

STODDEN R. & KALLMEYER L. (2022). TS-ANNO : An annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 145–155, Dublin, Ireland : Association for Computational Linguistics.

SULEM E., ABEND O. & RAPPOPORT A. (2018a). BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 738–744.

SULEM E., ABEND O. & RAPPOPORT A. (2018b). Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 685–696, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1063.

TANPRASERT T. & KAUCHAK D. (2021). Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, p. 1–14, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.gem-1.1.

TIBSHIRANI R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*, **58**(1), 267–288.

TODIRASCU A., WILKENS R., ROLIN E., FRANÇOIS T., BERNHARD D. & GALA N. (2022). HECTOR : A hybrid TExt SimplifiCation TOol for raw texts in French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 4620–4630, Marseille, France : European Language Resources Association.

VÁSQUEZ-RODRÍGUEZ L., SHARDLOW M., PRZYBYŁA P. & ANANIADOU S. (2021). Investigating text simplification evaluation. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP*, p. 876–882, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.77.

VÁSQUEZ-RODRÍGUEZ L., SHARDLOW M., PRZYBYLA P. & ANANIADOU S. (2021). The role of text simplification operations in evaluation. In *Proceedings of the SEPLN Workshop on Current Trends in Text Simplification*, p. 57–69.

WILKENS R., OBERLE B. & TODIRASCU A. (2020). Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, p. 93–100.

XU W., CALLISON-BURCH C. & NAPOLES C. (2015). Problems in current text simplification research : New data can help. *Transactions of the Association for Computational Linguistics*, **3**, 283–297. DOI : 10.1162/tacl_a_00139.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016a). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415.

XU W., NAPOLES C., PAVLICK E., CHEN Q. & CALLISON-BURCH C. (2016b). Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics*, **4**, 401–415. DOI : 10.1162/tacl_a_00107.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating text generation with BERT. In *International Conference on Learning Representations*.

ZHAO Y., CHEN L., CHEN Z. & YU K. (2020). Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**(05), 9668–9675. DOI : 10.1609/aaai.v34i05.6515.

A Tableaux d'analyse de SARI

	keep	*add*	*del*
inoth	0	0.1	0
split	0	0.73	0
deloth	-3.91	-0.19	3.05
delcst	-2.44	0	1.52
move	-1.55	0.14	1.5
delprop	-3.97	-0.74	3.66
incst	0	0.95	0
hyperonym	0	0.14	1.58
synonym	-1.68	0.61	3.63
delmod	-3.39	0	3.02

TABLE 1 – Coefficients des modèles de régression Lasso pour la prédiction des sous-composants de SARI. Les opérations ayant des coefficients différents de zéro pour *add* et *del* impliquent l'ajout et la suppression de *tokens*. Les coefficients négatifs pour la prédiction de *keep* indiquent que l'opération diminue le score du sous-composant *keep* de SARI. Les opérations absentes ont tous leurs coefficients à zéro.

	Spearman correlation	p-value
inoth	0.0161	0.3349
split	0.0915	0
deloth	-0.0137	0.4119
p2s	-0.017	0.3084
delcst	0.0413	0.0134
verbf	0.0519	0.0018
pron	-0.035	0.036
move	0.012	0.4731
inprop	0.0295	0.0775
delprop	-0.0381	0.0223
incst	0.0718	0
s2p	-0.0338	0.0428
fromPron	-0.0038	0.8194
merge	-0.0033	0.8421
p2a	0.0119	0.4752
pos2neg	0.0028	0.8648
neg2pos	-0.018	0.2806
hyponym	-0.0191	0.253
toImp	-0.0283	0.09
fromImp	-0.0349	0.0367
POSchange	0.0177	0.2895
hyperonym	0.0374	0.025
a2p	-0.018	0.2814
synonym	0.163	0
delmod	-0.0025	0.8792
inmod	0.0194	0.2451

TABLE 2 – Corrélation de Spearman entre l'occurrence des opérations dans les couples de phrases et le score SARI. Une p-value marquée à 0 signifie qu'elle est inférieure à 0,0001. Les p-values élevées s'expliquent par un nombre insuffisant d'occurrences des opérations correspondantes dans le corpus.

	keep		add		del	
Transformation	Spearman	p-value	Spearman	p-value	Spearman	p-value
inoth	-0.0752	0	0.0658	0.0001	0.073	0
split	0.0251	0.1328	0.1925	0	-0.0063	0.7052
deloth	-0.2214	0	0.0008	0.9614	0.2015	0
p2s	-0.0478	0.0042	0.0144	0.3899	0.0378	0.0235
delcst	-0.2267	0	0.1482	0	0.2279	0
verbf	-0.0827	0	0.089	0	0.1441	0
pron	-0.0827	0	0.011	0.5084	0.0321	0.0547
move	-0.1764	0	0.0828	0	0.1486	0
inprop	-0.0699	0	0.069	0	0.0947	0
delprop	-0.1694	0	-0.0636	0.0001	0.1809	0
incst	-0.1269	0	0.2198	0	0.1362	0
s2p	-0.108	0	0.0258	0.1217	0.073	0
fromPron	-0.0458	0.006	0.0114	0.4931	0.04	0.0165
merge	-0.0257	0.123	-0.0056	0.736	0.0269	0.107
p2a	-0.0304	0.0684	0.0123	0.4616	0.0421	0.0116
pos2neg	-0.038	0.0229	0.0186	0.2663	0.0346	0.038
neg2pos	-0.038	0.0226	-0.0147	0.3777	0.0378	0.0235
hyponym	-0.0373	0.0256	0.0263	0.1145	0.0068	0.6847
toImp	-0.0826	0	0.0148	0.3767	0.049	0.0033
fromImp	-0.0355	0.0333	-0.0278	0.0957	-0.0055	0.7408
POSchange	-0.1317	0	0.0655	0.0001	0.1538	0
hyperonym	-0.0303	0.0699	0.0836	0	0.0649	0.0001
a2p	-0.0307	0.0658	-0.0281	0.0926	0.0155	0.353
synonym	-0.0607	0.0003	0.2135	0	0.2116	0
delmod	-0.1857	0	0.0195	0.2416	0.2071	0
inmod	-0.0472	0.0047	0.0332	0.0466	0.0725	0

TABLE 3 – Corrélations de Spearman et p-values entre chaque transformation annotée et les sous-composant de SARI. Une p-value marquée à 0 signifie qu'elle est inférieure à 0,0001.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Un mot, deux facettes : traces des opinions dans les représentations contextualisées des mots

Aina Garí Soler Matthieu Labeau Chloé Clavel
LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.garisoler,matthieu.labeau,chloe.clavel}@telecom-paris.fr

RÉSUMÉ

La façon dont nous utilisons les mots est influencée par notre opinion. Nous cherchons à savoir si cela se reflète dans les plongements de mots contextualisés. Par exemple, la représentation d' « animal » est-elle différente pour les gens qui voudraient abolir les zoos et ceux qui ne le voudraient pas ? Nous explorons cette question du point de vue du changement sémantique des mots. Nos expériences avec des représentations dérivées d'ensembles de données annotés avec les points de vue révèlent des différences minimes, mais significatives, entre postures opposées [1].

ABSTRACT

One Word, Two Sides : Traces of Stance in Contextualized Word Representations

The way we use words is influenced by our opinion. We investigate whether this is reflected in contextualized word embeddings. For example, is the representation of "animal" different between people who would abolish zoos and those who would not ? We explore this question from a Lexical Semantic Change standpoint. Our experiments with BERT embeddings derived from datasets with stance annotations reveal small but significant differences in word representations between opposing stances.

MOTS-CLÉS : Représentations contextualisées, changement sémantique, détection de point de vue.

KEYWORDS: Contextualized representations, semantic change, stance detection.

1 Introduction

Nos opinions se reflètent dans notre façon de parler. Les personnes ayant des positions opposées sur un sujet particulier peuvent utiliser des mots différents pour en discuter. Par exemple, seules les personnes contre l'utilisation de masques pendant la pandémie de COVID-19 sont susceptibles de les appeler des « muselières ». Dans cet article, cependant, nous n'étudions pas *quels* mots sont utilisés de part et d'autre : nous comparons plutôt la façon dont les locuteurs qui sont en désaccord sur un sujet utilisent les *mêmes* mots. Plus précisément, nous cherchons à savoir si les modèles contextuels capturent une différence entre la représentation d'un mot (par exemple, « masque ») lorsqu'il est utilisé par des personnes qui sont pour ou contre une certaine cible (par exemple, l'utilisation obligatoire de masques).

Nous abordons cette question du point de vue du changement lexico-sémantique (CLS). Les travaux sur le CLS visent généralement à détecter les changements de sens des mots sur deux périodes de

1. Cet article est une adaptation et traduction de Garí Soler *et al.* (2022).

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 49–57.

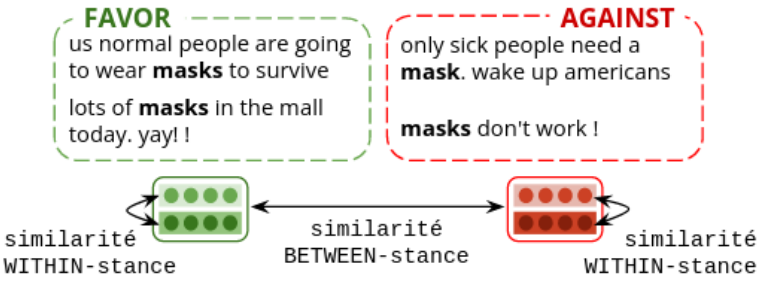


FIGURE 1 – Exemples d'instances de « mask » de l'ensemble de données de position Covid19 (Glandt *et al.*, 2021). Nous comparons la similarité d'utilisation interne aux, et entre, les positions.

temps ou plus (Tahmasebi *et al.*, 2021), mais leurs techniques ont également été employées pour identifier les différences synchroniques dans l'utilisation des mots, par exemple à travers différents âges, sexes, professions (Gonen *et al.*, 2020), domaines (Yin *et al.*, 2018; Schlechtweg *et al.*, 2019) ou cultures (Garimella *et al.*, 2016). Contrairement aux études connexes qui étudient le CLS entre différents points de vue (Azarbonyad *et al.*, 2017; Rodriguez *et al.*, 2021), notre objectif n'est pas d'explorer l'utilisation de mots spécifiques, et nous n'évaluons pas notre méthode en fonction du classement des mots par la stabilité de leur sens. Nous voulons plutôt déterminer si les représentations vectorielles reflètent une plus grande similitude dans l'utilisation des mots au sein d'un point de vue qu'entre différentes positions (voir Figure 1). Nous explorons cette question en nous appuyant sur des ensembles de données annotés avec des informations sur la position. Auparavant, nous testons différents modèles contextuels dans une configuration où les données sont rares, afin de sélectionner un type de représentation robuste [2].

Notre objectif à long terme est de détecter les différences d'utilisation des mots entre les locuteurs d'une conversation, ce qui pourrait indiquer leur niveau d'alignement conceptuel (Stolk *et al.*, 2016) ; c'est-à-dire la mesure dans laquelle les participants au dialogue « veulent dire la même chose en utilisant les mêmes mots » (Schober, 2005). Avec cette étude, nous présentons un premier pas dans cette direction. Les représentations sensibles aux différences d'opinion pourraient être utiles pour identifier les désaccords et les désalignements dans le dialogue.

2 Méthodologie

2.1 Données

Nous utilisons des ensembles de données en anglais contenant des phrases étiquetées comme étant en faveur (FAVOR) ou contre (AGAINST) une cible spécifique. Nous excluons les phrases sans position claire (NONE), le cas échéant. **SemEval2016** (Mohammad *et al.*, 2016b,a) contient des tweets sur six cibles variées. Nous en extrayons 3 253 phrases [3]. **Covid19** (Glandt *et al.*, 2021) est un ensemble de données avec 3 918 tweets centrés sur quatre cibles liées à la pandémie de COVID-19. **P-Stance** (Li *et al.*, 2021) contient 21 574 tweets sur trois politiciens américains. Enfin, IBM-ArgQ-Rank-30kArgs (Gretz *et al.*, 2020), ci-après **ArgQ**, est une collection d'arguments sur 71 cibles. Nous

2. Notre code et nos données sont disponibles sur https://github.com/ainagari/1word2sides
3. Nous omettons la cible « Climate Change is a Real Concern » car elle ne contient que 26 tweets AGAINST.

utilisons 29 972 arguments qui ont une position claire (avec un score de confiance [4] supérieur à 0,6, d'après Bar-Haim *et al.* (2020)). Nous voulons organiser les données de manière à nous permettre de déterminer si des instances du même mot ont une similitude plus élevée au sein d'une position qu'entre les positions. À cette fin, nous prétraitons et organisons les données comme suit.

Prétraitement : L'ensemble de données ArgQ était à l'origine destiné à la détection de la qualité des arguments, et plusieurs arguments mentionnent explicitement leur position. Pour atténuer les biais potentiels que cela pourrait engendrer, nous appliquons une stratégie qui omet automatiquement cette partie d'une phrase. Si le début d'une phrase contient les mêmes mots que la cible (avec l'ajout facultatif de *not* et *n't*) et est suivi de *because* (*of*), *as*, *since*, d'une virgule ou d'un point, on omet la première partie de la phrase jusqu'à ce token (inclus) [5]. Cette procédure modifie 3 223 phrases. Ensuite, les phrases dans tous les ensembles de données sont tokenisées, annotées en parties du discours et lemmatisées avec la librairie `nltk`.

Ensembles de phrases : Pour une cible donnée, nous divisons aléatoirement les phrases de chaque position (f (FAVOR) ou a (AGAINST)) en deux ensembles de taille égale P et Q. Avec ces ensembles, nous exécutons quatre comparaisons, deux intra-position : WITHIN-FAVOR (P_f vs Q_f) et WITHIN-AGAINST (P_a vs Q_a) ; et deux entre-positions : BETWEEN-1 (P_f vs Q_a) et BETWEEN-2 (P_a vs Q_f).

2.2 Représentations vectorielles

Nous voulons générer des représentations vectorielles pour des ensembles d'instances de mots tirés d'une position (par exemple, P_f). Par exemple, on veut obtenir une représentation du mot « woman » à partir de phrases en faveur du « mouvement féministe » (SemEval2016) et la comparer à la représentation de « woman » dans des phrases exprimant une position contre cette cible.

Dans la détection de CLS, les plongements statiques ont tendance à être plus performants que ceux contextualisés (Schlechtweg *et al.*, 2020). Une approche typique consiste à apprendre séparément les plongements statiques pour chaque période, corpus ou point de vue, puis les comparer soit en les alignant (Hamilton *et al.*, 2016) soit avec un approche basée sur les plus proches voisins (Gonen *et al.*, 2020). Dans ces études, même dans celles portant sur la détection de changement à court terme (Stewart *et al.*, 2017; Del Tredici *et al.*, 2019), il est courant de disposer d'un assez grande nombre d'instances d'un mot donné. Cependant, le nombre de phrases disponibles par mot dans un point de vue est limité dans nos données [6]. Nous expérimentons donc avec trois types différents de plongements contextualisés :

Les plongements *À la carte* (ALC) (Khodak *et al.*, 2018) ont été utilisées pour détecter des différences dans l'utilisation des mots selon les points de vue (Rodriguez *et al.*, 2021). Le modèle consiste à appliquer une transformation linéaire à la moyenne des plongements pré-entraînés des mots du contexte du mot cible. Nous utilisons un modèle ALC reposant sur des plongements GloVe 300*d* (Pennington *et al.*, 2014) formés sur 840 $\times 10^9$ tokens de Common Crawl.

4. Ce score reflète la mesure dans laquelle les annotateurs sont d'accord sur la position d'un argument. Il est calculé comme une moyenne pondérée des décisions des annotateurs et il varie de 0 à 1.

5. Par exemple, on retient seulement la partie en italique pour la phrase « Homeschooling should not be banned because *it is a right for parents to educate their children in their comfort of home* . » (pour la cible « Homeschooling should be banned »).

6. Par exemple, Schlechtweg *et al.* (2020) a une moyenne de 788 instances par lemme et période de temps. Dans nos données, le nombre moyen d'instances d'un mot dans un côté d'une comparaison est de 14.

Context2vec (c2v) (Melamud *et al.*, 2016) qui apprend simultanément la représentation d'un mot cible et, à l'aide d'un biLSTM, du contexte qui l'entoure : il est optimisé pour que les représentations du mot et de son contexte, plongées dans le même espace, soient similaires. Nous utilisons un modèle 600*d* entraîné sur le corpus ukWaC (Baroni *et al.*, 2009).

BERT : (Devlin *et al.*, 2019) Nous utilisons des représentations contextualisées générées avec le modèle 768*d* `bert-base-uncased`.

Nous notons V_P le vocabulaire d'un ensemble de phrases P. Nous incluons dans le vocabulaire tous les noms et les verbes apparaissant dans au moins trois phrases différentes de P. Dans les tweets, les mentions et les hashtags sont traités comme des noms. Les mots vides sont exclus. Nous traitons toutes les instances d'un lemme avec une catégorie grammaticale spécifique comme le même mot. Nous extrayons une représentation $\mathbf{w}_P$ pour chaque mot w dans V_P. Pour c2v et BERT, cela se fait en faisant la moyenne des représentations de toutes les instances w de P.

2.3 Tester les représentations

Nous identifions d'abord les représentations les mieux adaptées pour refléter la similarité sémantique lexicale entre de petits ensembles de phrases. En suivant Schlechtweg & Schulte im Walde (2020), nous utilisons SemCor (Miller *et al.*, 1993), un corpus annoté en sens, pour créer un ensemble de données qui simule le changement sémantique. Nous contrôlons en outre le nombre de phrases disponible pour chaque lemme. Le jeu de données est composé de 576 lemmes : 245 noms, 241 verbes, 69 adjectifs et 21 adverbes. Pour chaque lemme, nous avons deux ensembles de 25 instances chacun, P et Q. Pour simuler des situations de données, nous créons des sous-ensembles de taille X de P et Q (P_X , Q_X). Nous expérimentons différentes valeurs de X. Comme dans Schlechtweg & Schulte im Walde (2020), nous déterminons la « vraie » distance sémantique entre deux groupes P_X et Q_X en calculant la divergence de Jensen-Shannon (JSD) entre leurs distributions de sens.

Les prédictions de similarité pour un mot w sont obtenues en calculant simplement la similarité cosinus entre les représentations de ce lemme dans chaque ensemble de phrases, $cos(\mathbf{w}_{P_X}, \mathbf{w}_{Q_X})$. Nous rapportons le coefficient de corrélation τ-b de Kendall entre JSD et les similarités prédites par chaque type de représentation. Les résultats de cette expérience sont présentés dans la section 3.1.

2.4 Calcul de similarité et évaluation

Calcul de similarité : Pour calculer la similarité globale de l'utilisation des mots pour une comparaison entre deux ensembles de phrases P et Q, nous identifions d'abord les mots communs aux deux ensembles, $V_P \cap V_Q$. Cependant, $V_P \cap V_Q$ contient des mots qui ne sont pas nécessairement liés à la cible en question. On calcule donc une similarité basée uniquement sur un sous-ensemble de $V_P \cap V_Q$, appelé V_{PQ}. Le score de similarité est la similarité cosinus moyenne de tous les mots dans V_{PQ} :

$$sim(P, Q) = \frac{\sum_{w \in V_{PQ}} cos(\mathbf{w}_P, \mathbf{w}_Q)}{|V_{PQ}|} \quad (1)$$

Cette mesure de similarité vise à refléter dans quelle mesure les mots sont utilisés de la même manière et dans le même sens dans deux ensembles de phrases. Nous expérimentons avec trois définitions de

V_{PQ}. Dans chacune, nous veillons à utiliser le même nombre de mots pour les quatre comparaisons qui sont faites au sein d'une cible. Dans *all*, on inclut les k premiers mots les plus fréquents dans $V_P \cap V_Q$, où k correspond à la plus petite taille de $V_P \cap V_Q$ disponible pour cette cible. La fréquence est déterminée à partir de l'union des phrases dans P et Q. Nous utilisons également les 10 premiers mots de $V_P \cap V_Q$ avec les scores tf-idf les plus élevés dans cette cible (*tf-idf*). Les scores tf-idf sont calculés sur l'ensemble des jeux de données, en traitant toutes les phrases concernant la même cible comme un seul document. Enfin, on utilise aussi les 10 mots de $V_P \cap V_Q$ avec le plus petit poids de tf-idf (*rev-tf-idf*). Ce sous-ensemble contient des mots qui sont moins pertinents pour la cible, et donc nous nous attendons à ce que les similarités BETWEEN- et WITHIN- aient des valeurs plus proches. Notons que 25% des comparaisons (dans SemEval2016 et ArgQ) ont moins de 20 mots en commun. Dans ces cas, les valeurs *tf-idf* et *rev-tf-idf* sont partiellement calculées avec les mêmes mots.

Évaluation : Nous nous attendons à ce que les comparaisons de type WITHIN présentent une similarité moyenne plus élevée que les comparaisons BETWEEN. Pour mesurer à quel point cela est vrai, nous utilisons la précision par paires : nous vérifions pour combien de paires (WITHIN, BETWEEN) la comparaison de type BETWEEN a une similarité plus faible. Avec 4 comparaisons par cible, nos expériences portent sur un total de 332 paires (WITHIN, BETWEEN). Les résultats sur les données de position sont présentées dans la section 3.2.

3 Résultats

3.1 Sélection d'un type de représentation

Les résultats sur SemCor sont présentés sur la Figure 2. Dans les graphiques *a* et *b*, nous voyons les corrélations obtenues par les différents types de représentations sur différentes quantités de données (X). Naturellement, les performances sont moins bonnes avec des valeurs faibles de X. C'est notamment le cas de ALC, qui à $X = 25$ continue à s'améliorer. Dans le cas de c2v et BERT, cependant, nous n'observons pas de d'amélioration significative après $X = 10$. Dans ce contexte de données rares, les performances des plongements ALC sont bien inférieures à celles de c2v et BERT. Dans l'ensemble, les représentations BERT de la 10ème couche fonctionnent le mieux. On utilise donc des plongements de cette couche pour nos expériences sur les données de points de vue. Nous examinons également les performances des deux meilleurs modèles (c2v et la 10ème couche dans BERT) par PoS (figures *c* et *d*) : nous constatons que les noms et les verbes, qui sont les catégories grammaticales incluses dans nos expériences de position, sont généralement mieux représentés.

3.2 Résultats sur la position

Les précisions par paires obtenues avec la 10ème couche BERT avec différentes définitions de V_{PQ} se trouvent dans la Table 1 (gauche). Nous voyons que, en particulier pour *all* et *tf-idf*, la précision par paires est remarquablement élevée dans tous les ensembles de données. Cela montre que les représentations de mots contextualisées de BERT reflètent des différences dans la façon dont les mots sont utilisés entre deux postures opposées.

Lors de l'utilisation des 10 mots avec le tf-idf le plus faible (*rev-tf-idf*), les performances diminuent,

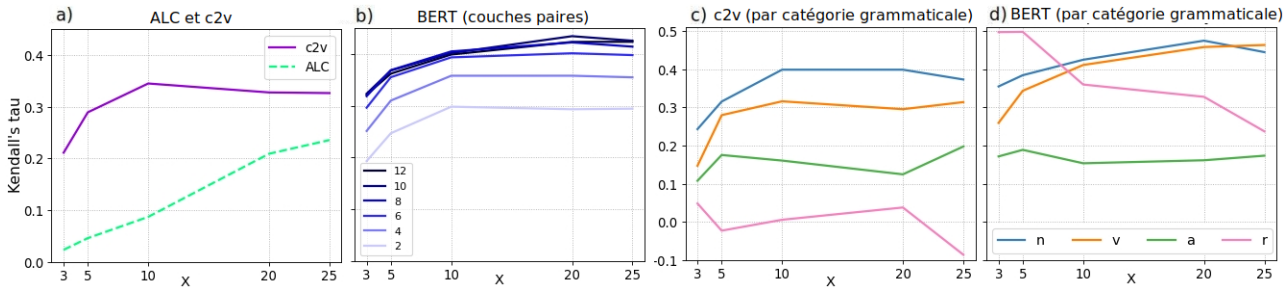


FIGURE 2 – *a* and *b* : τ de Kendall obtenu par différentes représentations vectorielles sur SemCor. Nous n'incluons que des couches paires pour BERT pour une meilleure lisibilité. *c* et *d* : Performances de c2v et BERT (10ème couche) par catégorie grammaticale.

Dataset	***all***	**tf-idf**	**rev-tf-idf**
SemEval2016	0.90	0.85	0.60
Covid19	0.88	0.81	0.50
P-stance	1.00	1.00	0.83
ArgQ	1.00	0.98	0.95
Global	0.99	0.96	0.90

	all	**tf-idf**	**rev-tf-idf**
a) W vs W	0.013	0.010	0.023
b) B vs B	0.013	0.010	0.023
c) W vs B	0.047	0.027	0.041

TABLE 1 – À gauche : Précision par paires par ensemble de données avec différents V_{PQ}. *Global* correspond à tous les jeux de données réunis. À droite : Différences de similarité entre comparaisons.

mais elles restent élevées pour les datasets P-Stance et ArgQ. Nous exécutons des tests du χ^2 pour la qualité d'ajustement sur les prédictions *rev-tf-idf* pour déterminer leur probabilité sous l'hypothèse nulle (H_0 : préc. $= 0.5$). Les valeurs-p sont significatives pour tous les jeux de données ensemble ($p < 0.001$) mais pas pour l'ensemble de toutes les données Twitter ($p = 0.08$, $\alpha = 0.05$). Il semble que les représentations BERT encodent, dans une certaine mesure, les différences dans les mots qui sont moins pertinents pour la cible. Cependant, si pour une raison quelconque, tous les mots ne peuvent pas être utilisés, (s'il y en a trop), alors il est préférable de sélectionner soigneusement un sous-ensemble (par exemple avec tf-idf).

Nous examinons également les mots qui ont les similitudes les plus élevées et les plus faibles dans les comparaisons BETWEEN. Les mots qui sont utilisés le plus différemment entre les positions tendent à être des noms qui sont au centre du sujet (par exemple « religion » dans « athéisme »), tandis que les mots les plus similaires sont souvent non-thématiques (« man » ou « take »).

Nous étudions l'ampleur des différences de similarité entre les comparaisons WITHIN (W) et BETWEEN (B) en examinant les différences de similarité (en valeur absolue) entre les paires de comparaison : a) entre WITHIN-FAVOR et WITHIN-AGAINST (W vs W), b) entre BETWEEN-1 et BETWEEN-2 (B vs B), et c) la différence moyenne trouvée dans les quatre appariements WITHIN vs BETWEEN (W vs B). Nous nous attendons à ce que ce dernier ait une plus grande différence de similarité que a) et b), où les comparaisons sont du même type. Les résultats sont présentés dans la Table 1 (droite). Nous rapportons la moyenne de ces valeurs sur toutes les données. Les différences de similarité sont assez faibles dans l'ensemble, ce qui indique que le contraste (c'est-à-dire la mesure dans laquelle les comparaisons WITHIN affichent une plus grande similarité que les comparaisons BETWEEN) est subtile. Les valeurs sont cependant entre 1,8 et 3,6 fois plus grandes pour les paires de comparaison W vs B. Pour toutes les définitions V_{PQ}, les valeurs de différence des paires de comparaison sont

significativement éloignées de celles en a) et b) ($p < 0{,}001$). [7]

4 Conclusion et travaux futurs

Nous avons montré que les représentations de mots BERT sont sensibles à l'opinion exprimée dans les phrases dont ils sont dérivés. Les différences de similitude trouvées entre les positions concordantes et contradictoires sont petites, mais significatives ; et les mots avec les différences les plus élevées ont tendance à être au cœur du sujet. Notre approche peut servir à identifier les points de divergence par rapport à une cible, et elle peut être utile pour la détection de position et l'analyse des débats. Nos expériences sur SemCor fournissent des informations précieuses sur la quantité suffisante d'instances de mots nécessaire à l'obtention de représentations de qualité, ce qui est pertinent pour étudier le CLS à faibles ressources et, plus généralement, pour dériver des vecteurs de mots à partir de peu de données.

Dans nos futurs travaux, nous prévoyons d'appliquer cette méthodologie au dialogue. Les ensembles P et Q correspondraient aux énoncés des participants à une conversation. La mesure de similarité agirait comme une approximation de l'alignement conceptuel ou de position entre les deux participants, indiquant si les locuteurs partagent des opinions et utilisent des mots de manière similaire.

Remerciements

Nous remercions les relecteurs anonymes pour leurs commentaires utiles. Ce travail a été soutenu par la chaire de recherche Télécom Paris sur la Science des Données et Intelligence Artificielle pour l'Industrie et les Services Numériques (DSAIDIS).

Références

AZARBONYAD H., DEHGHANI M., BEELEN K., ARKUT A., MARX M. & KAMPS J. (2017). Words Are Malleable : Computing Semantic Shifts in Political and Media Discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, p. 1509–1518, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3132847.3132878.

BAR-HAIM R., EDEN L., FRIEDMAN R., KANTOR Y., LAHAV D. & SLONIM N. (2020). From Arguments to Key Points : Towards Automatic Argument Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4029–4039, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.371.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The WaCky wide web : a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, **43**(3), 209–226.

7. Selon les tests de Wilcoxon ou de Student appariés, dépendant de la normalité des données (déterminée par les tests de Shapiro-Wilk).

DEL TREDICI M., FERNÁNDEZ R. & BOLEDA G. (2019). Short-Term Meaning Shift : A Distributional Exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 2069–2075, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1210.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GARÍ SOLER A., LABEAU M. & CLAVEL C. (2022). One word, two sides : Traces of stance in contextualized word representations. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3950–3959, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

GARIMELLA A., MIHALCEA R. & PENNEBAKER J. (2016). Identifying Cross-Cultural Differences in Word Usage. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 674–683, Osaka, Japan : The COLING 2016 Organizing Committee.

GLANDT K., KHANAL S., LI Y., CARAGEA D. & CARAGEA C. (2021). Stance Detection in COVID-19 Tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1596–1611, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.127.

GONEN H., JAWAHAR G., SEDDAH D. & GOLDBERG Y. (2020). Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 538–555, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.51.

GRETZ S., FRIEDMAN R., COHEN E., TOLEDO A., LAHAV D., AHARONOV R. & SLONIM N. (2020). A Large-Scale Dataset for Argument Quality Ranking : Construction and Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 7805–7813. DOI : 10.1609/aaai.v34i05.6285.

HAMILTON W. L., LESKOVEC J. & JURAFSKY D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1489–1501, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1141.

KHODAK M., SAUNSHI N., LIANG Y., MA T., STEWART B. & ARORA S. (2018). A La Carte Embedding : Cheap but Effective Induction of Semantic Feature Vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 12–22, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1002.

LI Y., SOSEA T., SAWANT A., NAIR A. J., INKPEN D. & CARAGEA C. (2021). P-Stance : A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 2355–2365, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.208.

MELAMUD O., GOLDBERGER J. & DAGAN I. (2016). context2vec : Learning Generic Context Embedding with Bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Compu-*

tational Natural Language Learning, p. 51–61, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/K16-1006.

MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A Semantic Concordance. In *Human Language Technology : Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

MOHAMMAD S., KIRITCHENKO S., SOBHANI P., ZHU X. & CHERRY C. (2016a). A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 3945–3952, Portorož, Slovenia : European Language Resources Association (ELRA).

MOHAMMAD S., KIRITCHENKO S., SOBHANI P., ZHU X. & CHERRY C. (2016b). SemEval-2016 Task 6 : Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 31–41, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1003.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1162.

RODRIGUEZ P. L., SPIRLING A. & STEWART B. M. (2021). *Embedding Regression : Models for Context-Specific Description and Inference*. Rapport interne, Working Paper Vanderbilt University.

SCHLECHTWEG D., HÄTTY A., DEL TREDICI M. & SCHULTE IM WALDE S. (2019). A Wind of Change : Detecting and Evaluating Lexical Semantic Change across Times and Domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 732–746, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1072.

SCHLECHTWEG D., MCGILLIVRAY B., HENGCHEN S., DUBOSSARSKY H. & TAHMASEBI N. (2020). SemEval-2020 Task 1 : Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, p. 1–23, Barcelona (online) : International Committee for Computational Linguistics. DOI : 10.18653/v1/2020.semeval-1.1.

SCHLECHTWEG D. & SCHULTE IM WALDE S. (2020). Simulating Lexical Semantic Change from Sense-Annotated Data. In A. RAVIGNANI, C. BARBIERI, M. MARTINS, M. FLAHERTY, Y. JADOUL, E. LATTENKAMP, H. LITTLE, K. MUDD & T. VERHOEF, Éds., *The Evolution of Language : Proceedings of the 13th International Conference (EvoLang13)*. DOI : 10.17617/2.3190925.

SCHOBER M. F. (2005). Conceptual alignment in conversation. *Other minds : How humans bridge the divide between self and others*, p. 239–252.

STEWART I., ARENDT D., BELL E. & VOLKOVA S. (2017). Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, **11**(1), 672–675.

STOLK A., VERHAGEN L. & TONI I. (2016). Conceptual alignment : How brains achieve mutual understanding. *Trends in cognitive sciences*, **20**(3), 180–191.

TAHMASEBI N., BORIN L. & JATOWT A. (2021). Survey of computational approaches to lexical semantic change detection. In N. TAHMASEBI, L. BORIN, A. JATOWT, Y. XU & S. HENGCHEN, Éds., *Computational approaches to semantic change*. Language Science Press. DOI : 10.5281/zenodo.5040302.

YIN Z., SACHIDANANDA V. & PRABHAKAR B. (2018). The global anchor method for quantifying linguistic shifts and domain adaptation. *Advances in neural information processing systems*, **31**.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

PromptORE – Vers l'Extraction de Relations non-supervisée

Pierre-Yves Genest[1, 2] Pierre-Edouard Portier[2] Előd Egyed-Zsigmond[2]
Laurent-Walter Goix[3] *
(1) Alteca, 88 Boulevard des Belges, 69006 Lyon, France
(2) Univ Lyon, INSA Lyon, LIRIS, CNRS UMR5205, 20 Avenue Einstein, 69621 Villeurbanne, France
(3) Nokia, 37 Quai du Président Roosevelt, 92130 Issy-les-Moulineaux, France
pygenest@alteca.fr, pierre-edouard.portier@insa-lyon.fr,
elod.egyed-zsigmond@insa-lyon.fr, laurent-walter.goix@nokia.com

RÉSUMÉ

L'extraction de relations non-supervisée vise à identifier les relations qui lient les entités dans un texte sans utiliser de données annotées pendant l'entraînement. Cette tâche est utile en monde ouvert, où les types de relations et leur nombre sont inconnus. Bien que des modèles récents obtiennent des résultats prometteurs, ils dépendent fortement d'hyper-paramètres dont l'ajustement nécessite des données annotées, signifiant que ces modèles ne sont pas complètement non-supervisés. Pour résoudre ce problème, nous proposons PromptORE, à notre connaissance le premier modèle d'extraction de relations non-supervisé qui ne nécessite pas d'ajustement d'hyper-paramètres. Pour cela, nous adaptons le principe du prompt-tuning pour fonctionner sans supervision. Les résultats montrent que PromptORE surpasse largement les méthodes à l'état de l'art, avec un gain relatif de $20 - 40\%$ en B^3, V-measure et ARI. Le code source est accessible [1].

ABSTRACT

PromptORE – A Novel Approach Towards Fully Unsupervised Relation Extraction

Unsupervised relation extraction aims to identify relations between entities in text, without having access to labeled data during training. This setting is particularly relevant for open-domain relation extraction where relation types and the number of relations are *a priori* unknown. Although recent approaches achieve promising results, they heavily depend on hyperparameters whose tuning requires labeled data, meaning they are not fully unsupervised. To mitigate this issue, we propose PromptORE, to the best of our knowledge, the first unsupervised relation extraction model that does not need hyperparameter tuning. We adapt the novel prompt-tuning paradigm to work in an unsupervised setting. Results shows that PromptORE consistantly outperforms state-of-the-art models with a relative gain of $20 - 40\%$ in B^3, V-measure and ARI. The source code is available[1].

MOTS-CLÉS : extraction de relation non-supervisée ; extraction de relation ouverte ; prompt-tuning.

KEYWORDS: unsupervised relation extraction ; open relation extraction ; prompt-tuning.

∗. Recherche réalisée lorsque l'auteur travaillait chez Alteca.
1. https://github.com/alteca/PromptORE.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 58–64.

1 Introduction

L'Extraction de Relations Non-Supervisée (ER-NS) vise à identifier les relations qui lient les entités dans un texte, sans avoir accès à des données annotées pendant l'entraînement. Cette tâche est particulièrement utile sur des domaines très spécifiques où aucun jeu de données annoté n'est disponible, ou en monde ouvert où les types de relations et leur nombre ne sont pas connus à l'avance. Plusieurs approches récentes s'intéressent à cette thématique et obtiennent des résultats prometteurs (Simon *et al.*, 2019; Tran *et al.*, 2020; Hu *et al.*, 2020), mais elles dépendent toutes de nombreux hyper-paramètres, dont l'ajustement nécessite des données annotées.

Pour répondre à ce problème, nous proposons **PromptORE**, un modèle d'*Extraction de Relations en monde Ouvert basée sur des Prompts* [2], que nous avons présenté lors de la conférence CIKM'22 (Genest *et al.*, 2022). Nous utilisons et modifions la nouvelle technique du *prompt-tuning* pour fonctionner en mode non-supervisé. Nous l'utilisons pour calculer une représentation vectorielle de chaque relation (*relation embedding*). Nous effectuons ensuite un clustering sur ces embeddings pour identifier des groupes de relations similaires, et effectuer des prédictions. À notre connaissance, PromptORE est le premier modèle d'ER-NS qui ne nécessite pas d'ajustement d'hyper-paramètres. Les résultats obtenus sur des datasets de domaines général et spécifique montrent que PromptORE surpasse largement les méthodes à l'état de l'art, avec un gain relatif de $20 - 40\%$ en B^3, V-measure et ARI, tout en étant plus simple. Une analyse qualitative montre aussi la capacité de PromptORE à extraire des clusters sémantiquement cohérents et très proches des vraies relations.

2 État de l'art

L'extraction de relations non-supervisée (ER-NS) vise à identifier la relation binaire r qui lie deux entités $\mathbf{e1}$ et $\mathbf{e2}$ dans le contexte d'une phrase $\mathbf{S}$ [3], sans utiliser un jeu de données annoté. Le plus souvent, il s'agit de regrouper dans un cluster toutes les instances ($\mathbf{S}$, $\mathbf{e1}$, $\mathbf{e2}$) qui expriment la même relation r. Les approches récentes s'appuient sur des modèles de langage, notamment BERT (Devlin *et al.*, 2019), et suivent un processus en deux étapes (Hu *et al.*, 2020; Simon *et al.*, 2019; Marcheggiani & Titov, 2016) : 1. encodage des instances, c'est-à-dire représenter chaque relation par un vecteur appelé *plongement de relation*, et 2. prédiction de la relation en utilisant le plongement. Simon *et al.* (2019); Marcheggiani & Titov (2016) utilisent une approche générative basée sur les auto-encodeurs variationnels. SelfORE (Hu *et al.*, 2020) utilise l'apprentissage auto-supervisé, en générant des pseudo-labels avec un clustering et en les prédisant avec un classificateur pour affiner les paramètres de BERT (*fine-tuning*). Cependant, Tran *et al.* (2020) montrent avec EType+ qu'un modèle très simple, raisonnant sur les types d'entités, obtient de meilleures performances que ces approches complexes.

Bien que tous ces modèles extraient des relations sans utiliser de données annotées, nous ne pensons pas qu'ils soient réellement non-supervisés. En effet, ils possèdent beaucoup d'hyper-paramètres comme le nombre de clusters, le nombre d'époques, d'itérations, le taux d'apprentissage, l'early-stopping, ou la régularisation. Ajuster ces hyper-paramètres est critique pour les performances des modèles, et les auteurs ne décrivent aucune méthode pour les ajuster automatiquement et sans données

2. *Prompt* peut être traduit par réplique ou requête.
3. Les modèles d'EN-RS se concentrent sur l'extraction dans des phrases, bien que l'extraction dans des documents soit plus générale.

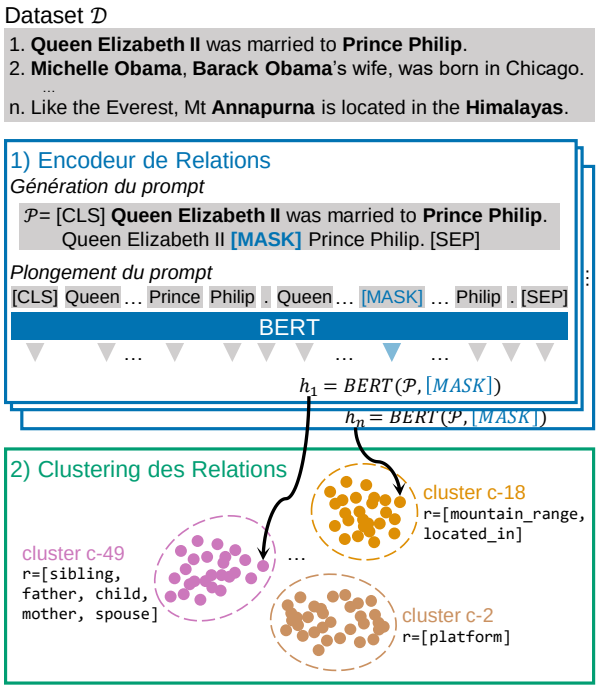


FIGURE 1 – Architecture de PromptORE.

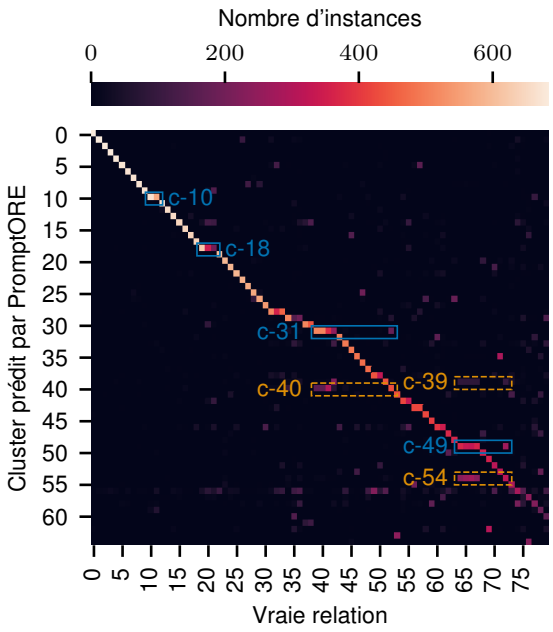


FIGURE 2 – Matrice de confusion entre les relations et les clusters de PromptORE sur FewRel.

annotées. De fait, les approches existantes ne sont pas utilisables dans un contexte non-supervisé réel.

3 PromptORE

C'est pour répondre à ce problème que nous proposons PromptORE. Il a pour objectif d'extraire la relation binaire r entre deux entités $\mathbf{e1}$ et $\mathbf{e2}$ déjà extraites dans une phrase $\mathbf{S}$, sans utiliser de données annotées pendant l'apprentissage et l'ajustement des hyper-paramètres. Nous supposons avoir accès à un dataset $\mathcal{D}$ contenant n instances (voir Figure 1). Chaque instance est décrite par un triplet ($\mathbf{S}$, $\mathbf{e1}$, $\mathbf{e2}$), avec $\mathbf{S}$ le texte de l'instance et $\mathbf{e1}$ et $\mathbf{e2}$ les deux entités. Nous supposons $\mathbf{e1}$ et $\mathbf{e2}$ extraits, et qu'il existe une relation r qui relie $\mathbf{e1}$ et $\mathbf{e2}$ dans le contexte de $\mathbf{S}$ (hypothèse habituelle de l'ER-NS). $\mathcal{R}$ est l'ensemble des k types de relations de $\mathcal{D}$. Nous ne connaissons pas k, ni n'avons d'information sur les relations (label, description, types d'entités, etc.). Comme SelfORE (Hu *et al.*, 2020), PromptORE est composé de deux modules principaux (voir Figure 1) : l'encodeur de relations, et le clustering des relations.

Encodeur de relations Ce module calcule une représentation vectorielle de la relation (*relation embedding*) qui lie $\mathbf{e1}$ à $\mathbf{e2}$ dans le contexte de $\mathbf{S}$. En d'autres termes, l'encodeur de relation abstrait le texte, pour fournir un plongement qui est facile à comparer grâce à une métrique (distance euclidienne). L'objectif est d'avoir un plongement représentatif de la relation sous-jacente : si deux instances ont des plongements proches, alors elles manifestent vraisemblablement la même relation r. Pour notre encodeur, nous utilisons le principe du *prompt-tuning* avec le modèle de langage BERT, que nous adaptons pour fonctionner dans un cadre non-supervisé [4] :

4. Le prompt-tuning a déjà été utilisé pour l'extraction de relation, mais dans un cadre supervisé (e.g., Lv *et al.* (2022)).

Dataset		Modèle	B^3 (F1)	V (F1)	ARI
FewRel (Han *et al.*, 2018)	$k = 80$	EType+ (Tran *et al.*, 2020)	13.7	47.9	8.4
		SelfORE (Hu *et al.*, 2020)	29.2	53.2	24.4
	$\hat{k} = 65$	PromptORE	**49.5**	**71.2**	**42.2**
FewRel PubMed (Gao *et al.*, 2019)	$k = 10$	EType+ (Tran *et al.*, 2020)	18.1	0	0
		SelfORE (Hu *et al.*, 2020)	59.3	63.4	45.4
	$\hat{k} = 10$	PromptORE	**77.4**	**81.1**	**73.8**

TABLE 1 – Résultats de PromptORE et de modèles à l'état de l'art sur deux jeux de données. EType+ et SelfORE ont accès au nombre de relations k, PromptORE l'estime ($\hat{k}$).

1. Pour une instance ($\mathbf{S}$, $\mathbf{e1}$, $\mathbf{e2}$), déterminer un prompt $\mathcal{P}$, qui est une séquence de tokens incluant un token `[MASK]`. Dans un cadre de *prompt-tuning* classique, l'objectif est d'optimiser la formulation de $\mathcal{P}$ pour améliorer les performances. Sans accès à des données annotées, cela nous est impossible, aussi nous proposons d'utiliser la template la plus simple possible :

$$\mathcal{P}(\mathbf{S}, \mathbf{e1}, \mathbf{e2}) = \texttt{"[CLS]}\ \mathbf{S}\ \mathbf{e1}\ \texttt{[MASK]}\ \mathbf{e2}\texttt{.[SEP]"}. \quad (1)$$

   Un exemple de prompt est disponible sur la Figure 1.

2. Prédire le plongement h du token `[MASK]` dans le contexte de $\mathcal{P}$ en utilisant BERT : $h = \text{BERT}(\mathcal{P}, \texttt{[MASK]})$. h est le *relation embedding*, le plongement de la relation.

Les approches précédentes d'ER-NS utilisent la technique de la représentation des paires d'entités pour l'encodeur (Hu *et al.*, 2020; Simon *et al.*, 2019; Marcheggiani & Titov, 2016) : elles calculent un plongement pour $\mathbf{e1}$ et $\mathbf{e2}$, grâce à l'ajout de tokens virtuels, et concatènent les deux plongements pour obtenir le *relation embedding*. Cette approche nécessite de fine-tuner BERT, ce qui est la source de beaucoup d'hyper-paramètres. Notre implémentation basée sur le prompt-tuning, ne nécessite pas de fine-tuning, supprimant de fait tous ces hyper-paramètres.

Clustering des relations Pour identifier les groupes d'instances manifestant la même relation r, nous calculons un clustering avec les k-moyennes (Lloyd, 1957) sur les plongements h de toutes les instances de $\mathcal{D}$ (voir Figure 1). Pour estimer le nombre de relations k, nous proposons d'utiliser l'*elbow rule* (Thorndike, 1953) en conjonction avec le coefficient silhouette (Rousseeuw, 1987). Elle permet de trouver un nombre de clusters $\hat{k}$ qui est un compromis entre la simplicité du clustering et la valeur du coefficient silhouette.

4 Expériences

Jeux de données, Métriques & Comparaisons Nous utilisons deux jeux de données : FewRel (Han *et al.*, 2018), de domaine général (articles Wikipedia) et contenant 56 000 instances avec 80 types de relations ; et FewRel PubMed (Gao *et al.*, 2019), centré sur le domaine biomédical, avec 10 types de relations différents. Étant donné que nous calculons un clustering comme les travaux précédents, nous évaluons PromptORE avec les métriques B^3 (Bagga & Baldwin, 1998), V-measure (V) (Rosenberg & Hirschberg, 2007) et Adjusted Rand Index (ARI) (Hubert & Arabie, 1985). Nous

nous comparons aux deux approches à l'état de l'art SelfORE (Hu *et al.*, 2020) et EType+ (Tran *et al.*, 2020). Ces deux approches sont entraînées avec les valeurs des hyper-paramètres déterminées par leurs auteurs respectifs. En particulier, ils ont besoin de connaître le nombre de relations k, ce qui n'est pas le cas de PromptORE.

Résultats quantitatifs[5] Les résultats sont disponibles sur la Table 1. De manière générale, PromptORE est bien plus performant que SelfORE et EType+ sur les trois métriques : l'écart est de $\approx 19\%$ en B^3, $\approx 18\%$ en V-measure et $\approx 23\%$ en ARI. PromptORE ne voit pas ses performances réduites sur FewRel PubMed (domaine biomédical), ce qui démontre sa flexibilité et son adaptabilité à d'autres domaines. SelfORE aussi n'est pas impacté, par contre EType+ ne prédit qu'une seule relation pour FewRel PubMed (ARI et V-measure à 0), montrant la limite du raisonnement sur les types d'entités pour des domaines très spécifiques. Les performances de PromptORE sont d'autant plus impressionnantes qu'il n'a pas accès à k, contrairement à SelfORE et EType+. PromptORE démontre qu'il est possible d'extraire des relations en mode non-supervisé sans hyper-paramètres à ajuster, tout en ayant de bien meilleures performances que les modèles à l'état de l'art.

Analyse qualitative[5] Sur la Table 1, nous remarquons que l'estimation $\hat{k}$ par l'elbow rule est correcte pour FewRel PubMed, mais qu'elle est sous-estimée pour FewRel : 65 clusters contre 80 relations. Cela signifie que certains clusters doivent être impurs, c'est-à-dire contenir les instances de plusieurs relations à la fois. La matrice de confusion de PromptORE évalué sur FewRel est disponible en Figure 2. Nous voyons que certains clusters sont impurs (e.g., c-10, c-18, c-31, c-49), mais dans l'ensemble nous observons une diagonale claire : PromptORE extrait précisément la grande majorité des relations, malgré son fonctionnement non-supervisé. La plupart des clusters sont relativement complets : peu de clusters partagent les mêmes relations (à l'exception de c-31, c-40 et c-39, c-49, c-54). En observant les clusters contenant plusieurs relations, nous remarquons qu'ils sont sémantiquement cohérents. Pour ne citer que deux exemples :

— c-10. Deux relations : `language_of_film_or_tv_show` et `language_of_work_or_name`. Ce sont des relations centrées sur le domaine du langage d'une oeuvre artistique.

— c-49. Cinq relations : `sibling`, `father`, `child`, `mother` et `spouse`. Relations autour des liens familiaux.

Du point de vue de FewRel, ces clusters sont impurs et incorrects, mais d'un point de vue qualitatif les relations extraites sont sémantiquement proches et cohérentes.

5 Conclusion

Nous avons introduit PromptORE, un modèle d'extraction de relations non-supervisé sans hyper-paramètres à ajuster, qui s'appuie sur la nouvelle méthode du prompt-tuning. Bien que très simple, PromptORE surpasse largement les approches à l'état de l'art. Pour le futur, nous envisageons de *fermer la boucle* de la connaissance, c'est-à-dire bénéficier de la connaissance extraite par PromptORE pour améliorer itérativement les performances de l'extraction.

5. Une évaluation plus détaillée de PromptORE est accessible sur (Genest *et al.*, 2022), contenant plus de datasets, d'approches comparées, et plusieurs formulations pour le prompt $\mathcal{P}$.

Remerciements

Ce travail est soutenu par Alteca et l'Association Nationale de la Recherche et de la Technologie (ANRT) par l'intermédiaire de la thèse CIFRE n°2021/0851. Nous remercions les relecteurs anonymes pour leurs remarques instructives et commentaires pertinents.

Références

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chain. In *Proceedings of the 1st International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, p. 563–566, Granada, Spain : European Language Resources Association.

DEVLIN J., CHANG M. W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, volume 1, p. 4171–4186, Stroudsburg, PA, USA : Association for Computational Linguistics. DOI : 10.18653/V1/N19-1423.

GAO T., HAN X., ZHU H., LIU Z., LI P., SUN M. & ZHOU J. (2019). Fewrel 2.0 : Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, p. 6250–6255, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/d19-1649.

GENEST P.-Y., PORTIER P.-E., EGYED-ZSIGMOND E. & GOIX L.-W. (2022). PromptORE - A Novel Approach Towards Fully Unsupervised Relation Extraction. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, p.1̃1, Atlanta, USA : ACM. DOI : 10.1145/3511808.3557422.

HAN X., ZHU H., YU P., WANG Z., YAO Y., LIU Z. & SUN M. (2018). Fewrel : A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 4803–4809, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/d18-1514.

HU X., WEN L., XU Y., ZHANG C. & YU P. S. (2020). SelfORE : Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, p. 3673–3682, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.299.

HUBERT L. & ARABIE P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218. Publisher : Springer, DOI : 10.1007/BF01908075.

LLOYD S. P. (1957). Least squares quantization in PCM. *Technical Report RR-5497*.

LV B., JIN L., ZHANG Y., WANG H., LI X. & GUO Z. (2022). Commonsense Knowledge-Aware Prompt Tuning for Few-Shot NOTA Relation Classification. *Applied Sciences*, **12**(4), 2185–2185. Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/app12042185.

MARCHEGGIANI D. & TITOV I. (2016). Discrete-State Variational Autoencoders for Joint Discovery and Factorization of Relations. *Transactions of the Association for Computational Linguistics*, **4**, 231–244. Publisher : MIT Press - Journals, DOI : 10.1162/tacl_a_00095.

ROSENBERG A. & HIRSCHBERG J. (2007). V-Measure : A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 410–420, Prague, Czech Republic : Association for Computational Linguistics.

ROUSSEEUW P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**(C), 53–65. Publisher : North-Holland, DOI : 10.1016/0377-0427(87)90125-7.

SIMON E., GUIGUE V. & PIWOWARSKI B. (2019). Unsupervised information extraction : Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 1378–1387, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/p19-1133.

THORNDIKE R. L. (1953). Who belongs in the family ? *Psychometrika*, **18**(4), 267–276. Publisher : Springer, DOI : 10.1007/BF02289263.

TRAN T. T., LE P. & ANANIADOU S. (2020). Revisiting Unsupervised Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7498–7505, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.669.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Injection de connaissances temporelles dans la reconnaissance d'entités nommées historiques

Carlos-Emiliano González-Gallardo[1] Emanuela Boros[2*] Edward Giamphy[1,3]
Ahmed Hamdi[1] José G. Moreno[4] Antoine Doucet[1]
(1) La Rochelle Université, L3i, 17000 La Rochelle, France
(2) EPFL, Digital Humanities Laboratory, Lausanne, Suisse
(3) Preligens, 75009 Paris, France
(4) Université de Toulouse, IRIT UMR 5505 CNRS, 31000 Toulouse, France
carlos.gonzalez_gallardo@univ-lr.fr,{prénom.nom}@univ-lr.fr,
emanuela.boros@epfl.ch, edward.giamphy@preligens.com
jose.moreno@irit.fr

RÉSUMÉ

Dans cet article nous abordons la reconnaissance d'entités nommées (NER) dans des documents historiques multilingues. Cette tâche présente des multiples défis, tels que les erreurs générées suite à la numérisation et de la reconnaissance optique des caractères de ces documents. De plus, ces collections sont distribuées sur une période de temps assez longue et suivent plusieurs conventions orthographiques qui évoluent au fil du temps. Pour répondre à ce défi nous récupérons des contextes supplémentaires, sémantiquement pertinents en exploitant des graphes de connaissances temporelles à partir des informations temporelles fournies par les collections historiques. Ces contextes sont ensuite inclus en tant que représentations mises en commun dans un modèle NER basé sur des Transformeurs. Nous menons des expérimentations avec deux collections historiques multilingues récentes en anglais, français et allemand, composées de journaux historiques (XIX^{e} - XX^{e} siècles) et de commentaires classiques (XX^{e} siècle). Les résultats démontrent l'efficacité de l'injection de connaissances temporelles dans ces ensembles de données.

ABSTRACT

Injecting Temporal-aware Knowledge in Historical Named Entity Recognition.

In this paper we address the detection of named entities in multilingual historical collections. This task presents multiple challenges as a result of digitization and optical caracter recogonition processes. In addition, these collections are distributed over a fairly long period of time and are affected by changes and evolution of natural language. To address this challenge we retrieve semantically-relevant additional contexts from temporal knowledge graphs by extracting the time information provided on historical data collections and include them as mean-pooled representations in a Transformer-based NER model. We experiment with two recent multilingual historical collections in English, French, and German, consisting of historical newspapers (19C-20C) and classical commentaries (19C). The results show the effectiveness of injecting temporal-aware knowledge into the different datasets.

MOTS-CLÉS : Reconnaissance d'entités nommées, Extraction d'informations temporelles, Humanités numériques.

KEYWORDS: Named entity recognition, Temporal information extraction, Digital humanities.

∗. Ce travail a été réalisé à l'Université de La Rochelle, à La Rochelle, France.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 65–73.

1 Introduction

Ces dernières décennies ont vu la mise à disposition d'un nombre croissant de corpus textuels pour les sciences humaines et sociales. Des exemples représentatifs proviennent de *Gallica*, la bibliothèque numérique de la Bibliothèque nationale de France [1], et de *Trove*, l'agrégateur de bases de données et service de documents en texte intégral, d'images numériques et de stockage de données provenant de la Bibliothèque nationale d'Australie [2]. L'accès à ces données massives offre de nouvelles perspectives à un nombre croissant de disciplines, allant de l'histoire sociopolitique et culturelle à l'histoire économique, ainsi que de la linguistique à la philologie.

Des milliards d'images de documents historiques, y compris des documents manuscrits numérisés, des registres médiévaux et des journaux anciens numérisés, sont désormais stockés et leur contenu est transcrit, soit manuellement grâce à des interfaces dédiées, soit automatiquement en utilisant la reconnaissance optique de caractères (OCR) ou la reconnaissance de texte manuscrit. Le processus de numérisation en masse, initié dans les années 1980 par des projets internes à petite échelle, a conduit à la montée en puissance de la numérisation, qui a atteint une certaine maturité au début des années 2000 avec des campagnes de numérisation à grande échelle dans toute l'industrie (Ehrmann *et al.*, 2020a,c).

Alors que ce processus de numérisation de masse se poursuit de plus en plus d'approches du domaine du traitement du langage naturel (TAL) sont dédiées aux documents historiques, offrant de nouveaux moyens d'accéder à des archives enrichies sémantiquement en texte intégral (Oberbichler *et al.*, 2022), tels que la reconnaissance d'entités nommées (NER) (Boroş *et al.*, 2020a; Ehrmann *et al.*, 2023; Hamdi *et al.*, 2021), l'annotation sémantique (Linhares Pontes *et al.*, 2022) et la détection d'événements (Boros *et al.*, 2022; Nguyen *et al.*, 2020).

La NER est une tâche d'extraction d'information dédiée à l'identification d'entités d'intérêt dans les textes, généralement de type personne, organisation et lieu. Ces entités agissent comme des ancrages référentiels qui sous-tendent la sémantique des textes et guident leur interprétation. Par exemple, en Europe, à l'époque médiévale, la plupart des personnes étaient identifiées par un simple mononyme ou un seul nom propre. Les noms de famille ou patronymes ont commencé à être utilisés à partir du XIII^{e} siècle, mais beaucoup plus tard dans certaines régions ou classes sociales (XVII^{e} siècle pour les Gallois). De nombreuses personnes partageaient le même nom et la même orthographe dans les langues vernaculaires et latines, mais aussi au sein d'une même langue (e.g., Guillelmus, Guillaume, Willelmus, Guillaume, Wilhelm). Les lieux ont pu disparaître ou changer complètement. Pour ceux qui ont survécu de la préhistoire jusqu'au XXI^{e} siècle (e.g., l'Écosse, le Pays de Galles, l'Espagne), ils sont très ambigus et possèdent de très différentes orthographes, ce qui rend leur identification très difficile (Boroş *et al.*, 2020b).

Dans cet article, nous nous concentrons sur l'exploration de la temporalité dans la NER à partir de collections historiques. Nous proposons une nouvelle technique pour injecter des connaissances temporelles supplémentaires en s'appuyant sur Wikipédia et Wikidata pour fournir des informations contextuelles sémantiquement proches.

1. https://gallica.bnf.fr/
2. https://trove.nla.gov.au/

2 Contextes basés sur des connaissances temporelles

Plusieurs travaux ont montré que les erreurs d'OCR peuvent avoir un impact sur des tâches en TAL (van Strien *et al.*, 2020), et plus particulièrement sur la NER (Hamdi *et al.*, 2022). Pour remédier à cela, des efforts ont été déployés pour élaborer des corpus et/ou des systèmes de NER adaptés (Ehrmann *et al.*, 2023). Dans ce travail, nous proposons d'introduire des contextes externes grammaticalement corrects dans les systèmes de NER. Ces contextes supplémentaires contribuent à améliorer les performances des systèmes de NER en dépit les erreurs d'OCR (Wang *et al.*, 2022). De plus, l'inclusion de tels contextes, en tenant compte de la temporalité, pourrait encore améliorer la détection des entités, qui sont particulièrement sensibles au contexte temporel. Ainsi, nous proposons plusieurs configurations pour inclure ces contextes supplémentaires, basés sur Wikidata5m [3] (Wang *et al.*, 2021), un graphe de connaissances (KG) comportant cinq millions d'entités Wikidata [4]. Wikidata5m contient des entités du domaine général (e.g., des célébrités, des événements, des concepts, des objets) qui sont alignées sur une description correspondant au premier paragraphe de sa page Wikipédia."

2.1 Intégration de l'information temporelle

Nous agrégeons la temporalité dans Wikidata5m en utilisant le graphe de connaissances temporelles (TKG) créé par (Leblay & Chekol, 2018) et mis au point par (García-Durán *et al.*, 2018) [5]. Ce TKG contient plus de 11 000 entités, 150 000 faits, et un champ temporel couvrant les années 508 à 2017. Pour une entité donnée, il fournit un ensemble de faits décrivant les interactions de l'entité dans le temps. Il est donc nécessaire de combiner ces faits en un seul élément en utilisant un opérateur d'agrégation sur leurs éléments temporels.

Nous effectuons une transformation sur les informations temporelles de chaque fait d'une entité afin de les combiner en un seul élément d'information temporelle. Soit e une entité décrite par les faits : $F_e i = 1^n = (e, r1, e_1, t_1), \ldots (e, r_i, e_i, t_i), \ldots (e, r_n, e_n, t_n)$, où le fait (e, r_i, e_i, t_i) est composé de deux entités e et e_i reliées par la relation r_i et l'horodatage t_i. Un horodatage est un point discret dans le temps qui correspond à une période (une année dans ce travail). L'opérateur d'agrégation est la fonction $AGG \rightarrow t_e$ qui prend en entrée l'information temporelle de F_e et génère l'information temporelle associée à e. Plusieurs opérateurs sont possibles (moyenne, médiane, minimum et maximum). Le minimum d'un ensemble de faits est défini par le fait le plus ancien tandis que le maximum correspond au fait le plus récent. Si une entité est associée à quatre faits s'étendant sur les années 1891, 1997, 2006 et 2011, l'opérateur d'agrégation minimum consiste à conserver le plus ancien, ce qui fait de l'année 1891 l'information temporelle de l'entité.

Étant donné que nos ensembles de données correspondent à des documents entre le XIX^{e} et le XX^{e} siècles, l'opérateur d'agrégation minimum est plus susceptible de créer un contexte temporel approprié pour les entités. Il met en évidence les entités correspondant à une période en accentuant les faits plus anciens. À la fin de l'opération d'agrégation, 8 176 entités de Wikidata5m ont été associées à une année comprise entre 508 et 2001, ce qui permet de filtrer la plupart des faits survenus au cours du XXI^{e} siècle.

3. https://deepgraphlearning.github.io/project/wikidata5m
4. https://www.wikidata.org/
5. https://github.com/mniepert/mmkb/tree/master/TemporalKGs/wikidata

2.2 Recherche de contexte

Notre système de base de connaissances repose sur une instance locale d'ElasticSearch [6] et utilise une correspondance de similarité sémantique multilingue, ce qui présente un avantage pour les requêtes multilingues. Cette correspondance est réalisée avec des index de champ vectoriel dense. Ainsi, à partir d'un vecteur de requête, une API de recherche des k plus proches voisins (k-NN) récupère les k vecteurs les plus proches et renvoie les documents correspondants en tant que résultats de recherche.

Pour chaque entité Wikidata5m, nous créons une entrée ElasticSearch comprenant un identifiant, un champ de description et un champ contenant le vecteur dense de la description, obtenus à l'aide du modèle multilingue pré-entraîné Sentence-BERT (Reimers & Gurevych, 2019, 2020). Nous construisons un index sur l'identifiant de l'entité et un index vectoriel dense sur les vecteurs de description. Nous proposons deux configurations différentes pour la récupération du contexte :

— `non-temporelle` : cette configuration n'utilise aucune information temporelle. Lors de la recherche de contexte pour une phrase d'entrée, nous commençons par obtenir la représentation vectorielle dense correspondante avec le même modèle Sentence-BERT utilisé pendant la phase d'indexation. Ensuite, nous interrogeons la base de connaissances afin de récupérer les entités les plus proches sur le plan sémantique en utilisant une recherche de similarité cosinus via l'algorithme des *k-NN* sur l'index du vecteur dense de la description. Le contexte C est finalement composé de k descriptions d'entités.

— `temporelle-`δ : cette configuration intègre les informations temporelles. Après la récupération des entités sémantiquement similaires avec `non-temporelle`, nous appliquons une opération de filtrage pour garder ou exclure les entités du contexte. En utilisant l'année t_{input} liée aux métadonnées de la phrase d'entrée lors de la recherche de contexte, nous conservons une entité si son année associée t_e se situe dans l'intervalle $t_{input} - \delta \leq t_e \leq t_{input} + \delta$, où δ est le seuil de l'intervalle d'années. Sinon, l'entité est rejetée. La valeur de t_e correspond à l'année la plus ancienne parmi tous les faits de l'entité e dans le TKG, conformément à l'opération d'agrégation *AGG*. Si t_e est absent, l'entité e est également conservée. Cette opération est répétée jusqu'à ce que $|C| = k$.

2.3 Architecture

Notre modèle de base se compose d'une approche d'apprentissage hiérarchique et multitâche, avec un encodeur ajusté basé sur BERT. Ce modèle comprend un encodeur avec deux couches de Transformeur (Vaswani *et al.*, 2017) dotées de modules adaptateurs (Houlsby *et al.*, 2019; Pfeiffer *et al.*, 2020) au-dessus du modèle pré-entraîné BERT. Les adaptateurs sont ajoutés à chaque couche de Transformeur après la projection suivant l'attention multitêtes et ils s'adaptent non seulement à la tâche, mais aussi à l'entrée bruitée, ce qui a prouvé augmenter les performances de la reconnaissance d'entités dans de telles conditions spéciales (Boroş *et al.*, 2020a). Enfin, la couche de prédiction multitâche est constituée de couches distinctes de champs conditionnels aléatoires (CRF).

Pour inclure les contextes supplémentaires, nous introduisons les *jokers contextuels*. Chaque contexte supplémentaire passe par l'encodeur pré-entraîné [7] générant un $JokerTokRep$ qui est ensuite réduit

6. https://www.elastic.co/guide/en/elasticsearch/reference/8.1/release-highlights.html

7. Dans ce cas, nous n'utilisons pas les couches de Transformeur supplémentaires avec les adaptateurs, car ceux-ci ont été spécifiquement proposés pour le texte bruité/non standard et n'apportent aucune amélioration des performances sur le texte

à la moyenne le long de l'axe de la séquence. Nous appelons cette représentation le *joker contextuel*. Nous les considérons comme des jokers insérés discrètement dans la représentation de la phrase actuelle pour améliorer la reconnaissance des entités. Cependant, nous considérons également que ces jokers peuvent affecter les résultats d'une manière qui n'est pas immédiatement apparente et peuvent nuire aux performances d'un système de NER.

Configuration expérimentale Notre configuration expérimentale comprend un modèle de base et quatre configurations avec différents niveaux de contextes basés sur les connaissances :

— *sans-contexte* : dans cette configuration de base, aucun contexte n'est ajouté aux représentations des phrases d'entrée.
— *non-temporelle* : les *jokers contextuels* sont générés et intégrés avec la première configuration de recherche de contexte sans information temporelle.
— *temporelle-(50|25|10)* : les *jokers contextuels* sont générés et intégrés à l'aide de la deuxième configuration de recherche de contexte avec un seuil d'intervalle d'année $\delta \in \{50, 25, 10\}$.

L'évaluation est réalisée en termes de [P]récision, de [R]appel et de mesure F1 au niveau micro (Ehrmann *et al.*, 2020a) dans un cadre strict (correspondance exacte des limites) [8].

Jeux de données Nous avons sélectionné deux collections de documents historiques comprenant des journaux historiques et des commentaires classiques.

— `hipe-2020` (Ehrmann *et al.*, 2020b) : couvre les XIX^{e} et XX^{e} siècles et rassemble des articles de journaux suisses, luxembourgeois et états-uniens en français, en allemand et en anglais provenant de diverses sources telles que la Bibliothèque nationale suisse (BN), la Bibliothèque nationale du Luxembourg (BnL), la Médiathèque et des Archives d'Etat du Valais et les Archives économiques suisses (AES) [9] dans le cadre du projet *impresso*.
— `ajmc` (Romanello *et al.*, 2021) : se compose de commentaires classiques rédigés en français, en allemand et en anglais provenant du projet *Ajax Multi-Commentary*. Ces commentaires, datant du XIX^{e} siècle, fournissent une analyse détaillée de la tragédie grecque *Ajax* de Sophocle datant du début de la période médiévale [10].

3 Résultats

Le tableau 1 présente nos résultats pour les trois langues et les deux ensembles de données. Les meilleurs résultats sont en gras. Nous pouvons observer que les modèles avec des *jokers contextuels* présentent une amélioration par rapport au modèle de base sans contextes supplémentaires. De plus, l'inclusion d'informations temporelles conduit à de meilleurs résultats que les contextes non temporels. Les scores sur `ajmc` s'avèrent plus élevés que ceux obtenus sur `hipe-2020`, quel que

standard, comme l'ont observé Boroş *et al.* (2020a).

8. Nous avons utilisé l'évaluateur HIPE disponible sur `https://github.com/hipe-eval/HIPE-scorer`.

9. BN : `https://www.nb.admin.ch`; BnL : `https://bnl.public.lu`; AES : `https://wirtschaftsarchiv.ub.unibas.ch`

10. Bien que la date exacte de sa première représentation soit inconnue, la plupart des spécialistes la datent du début de la carrière de Sophocle (peut-être la plus ancienne pièce de Sophocle encore existante), quelque part entre 450 et 430 avant J.-C., peut-être vers 444 avant J.-C.

Français						Allemand						Anglais					
hipe-2020			ajmc			hipe-2020			ajmc			hipe-2020			ajmc		
P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
sans-contexte																	
0,755	0,757	0,756	0,829	0,806	0,817	0,754	0,730	0,742	0,910	0,877	0,893	0,604	0,563	0,583	0,789	0,859	0,823
non-temporelle																	
0,762	**0,767**	**0,765**	0,829	0,783	0,806	0,759	**0,767**	**0,763**	**0,930**	0,898	0,913	0,565	0,601	0,583	0,828	0,871	0,849
temporelle-50																	
0,765	0,765	**0,765**	0,839	0,822	0,830	0,748	0,756	0,752	0,921	**0,911**	**0,916**	**0,643**	0,617	**0,630**	0,855	0,882	0,868
temporelle-25																	
0,759	0,756	0,757	**0,848**	**0,839**	**0,844**	0,757	0,743	0,750	0,925	0,903	0,914	0,621	0,630	0,625	0,833	0,876	0,854
temporelle-10																	
0,762	0,764	0,763	**0,848**	**0,839**	**0,844**	**0,760**	0,765	0,762	0,917	0,898	0,907	0,605	**0,646**	0,625	**0,866**	**0,888**	**0,877**

TABLE 1 – Résultats sur le français, l'allemand et l'anglais, pour les deux jeux de données.

	Français		Allemand		Anglais	
	train	test	train	test	train	test
temporelle-50 / 25 / 10						
hipe-2020	120 / 154 / 217	42 / 47 / 61	325 / 393 / 482	12 / 14 / 14	192 / 222 / 246	77 / 85 / 96
ajmc	10 /12 / 12	0 / 0 / 0	71 / 71 / 73	20 / 20 / 20	2 / 2 / 2	0 / 0 / 0

TABLE 2 – Nombre de contextes remplacés par période.

soit la langue et les contextes utilisés. Nous expliquons ce comportement par la faible diversité de certains types d'entités dans ajmc. Par exemple, les dix entités les plus fréquentes du type "personne" représentent respectivement 55%, 51,5% et 62,5% de toutes les entités "personne" dans les ensembles d'entraînement, de développement et de test. Il existe également une intersection de 80% entre les dix entités les plus fréquentes des ensembles d'entraînement et de test, ce qui signifie que huit des dix entités les plus fréquentes apparaissent à la fois dans les ensembles d'entraînement et de test. Le jeu de données hipe-2020 en anglais présente les scores les plus bas par rapport au français et à l'allemand, indépendamment des contextes. Nous attribuons cette baisse de performance à l'absence d'un corpus d'entraînement en anglais.

Impact des intervalles de temps Le jeu de données ajmc en allemand contient des commentaires provenant de deux années (1853 et 1894), le ajmc en anglais provient également de deux années (1881 et 1896), tandis que le ajmc en français ne concerne qu'une seule année (1886). En raison de la taille de la collection, hipe-2020 couvre un plus grand nombre d'années. En termes de couverture, les articles en français ont été collectés de 1798 à 2018, les articles en allemand de 1798 à 1948, et les articles en anglais de 1790 à 1960. Ainsi, nous avons examiné la différence entre les contextes récupérés par les configurations temporelles et non temporelles.

Le tableau 2 résume ces différences pour les ensembles d'entraînement et de test et indique le nombre de contextes qui ont été filtrés et remplacés par non-temporelle pour chaque intervalle de temps. En général, plus l'intervalle d'années est court, plus le nombre de contextes remplacés est élevé. Nous remarquons que le nombre de contextes remplacés est plus faible pour ajmc que pour hipe-2020. Cela s'explique par la taille limitée de la plage temporelle et le manque de diversité des entités pendant cette période. En comparant avec les résultats du tableau 1, nous pouvons déduire qu'en général, l'utilisation d'intervalles de temps plus courts, tels que $\delta = 10$, est bénéfique. En effet, la configuration temporelle-10 présente les scores F1 les plus élevés.

Impact des erreurs de numérisation Les commentaires sur la littérature grecque classique de `ajmc` présentent les difficultés typiques de l'océrisation historique. Avec des mises en page complexes, souvent composées de plusieurs colonnes et lignes de texte, la qualité de la numérisation des commentaires peut avoir un impact significatif sur la NER et d'autres tâches en aval, telles que la liaison d'entités. Statistiquement, environ 10% des entités sont affectées par des erreurs d'OCR dans les corpus `ajmc` en anglais et en allemand, tandis que ce chiffre s'élève à 27,5% dans le corpus en français. Les modèles intégrant un contexte supplémentaire, en particulier les approches temporelles, contribuent à la reconnaissance correcte des entités nommées, qu'elles soient contaminées ou non par des erreurs d'OCR. Cette amélioration est particulièrement significative pour la reconnaissance des entités nommées contaminées, par rapport à celles qui ne le sont pas (même si ces dernières sont plus fréquentes). Par exemple, dans le corpus en allemand, la configuration `temporelle-50` apporte une amélioration d'environ 14 points de pourcentage par rapport au modèle de base pour les entités contaminées, tandis que cette amélioration est de seulement 2 points de pourcentage pour les entités non contaminées. En outre, les trois quarts des entités présentant un taux d'erreur sur les caractères de 67% sont correctement reconnus, tandis que le modèle de base n'en reconnaît qu'un quart. Enfin, les entités avec des taux d'erreur supérieurs à 70% ne sont pas du tout reconnues par tous les modèles.

4 Conclusions et perspectives

Dans cet article, nous avons exploré l'apport de l'injection d'informations temporelles dans la tâche de reconnaissance d'entités nommées à partir de collections historiques. Nos résultats ont démontré que l'injection de *jokers contextuels* sur de courtes périodes offre de meilleurs résultats pour les collections présentant une diversité d'entités limitée et des intervalles de temps restreints. De manière symétrique, l'utilisation de *jokers contextuels* sur une période plus longue est plus bénéfique pour les intervalles d'années plus larges. Nous avons également montré que notre approche est performante dans la détection des entités affectées par des erreurs de numérisation, même lorsque le taux d'erreur des caractères atteint 67%. Enfin, nous avons observé que la qualité des contextes récupérés dépend de l'adéquation entre la collection historique et la base de connaissances. Ainsi, dans de futures recherches, il serait intéressant d'inclure des informations sur la temporalité en prédisant les intervalles d'années à partir d'un large ensemble de pages Wikipédia, afin de les utiliser comme contextes complémentaires.

Limitations Idéalement, le système requiert des métadonnées indiquant l'année de rédaction des ensembles de données, ou du moins un intervalle temporel. Dans le cas contraire, il sera nécessaire de recourir à d'autres systèmes pour prédire l'année de publication (Rastas *et al.*, 2022). Cependant, les erreurs générées par ces systèmes se propageront et pourront influencer les résultats de la reconnaissance d'entités nommées.

Remerciements

Ce travail a été soutenu par les projets ANNA (2019-1R40226), TERMITRAD (AAPR2020-2019-8510010), Pypa (AAPR2021-2021-12263410) et Actuadata (AAPR2022-2021-17014610) financés par la Région Nouvelle-Aquitaine, France.

Références

BOROŞ E., HAMDI A., PONTES E. L., CABRERA-DIEGO L.-A., MORENO J. G., SIDERE N. & DOUCET A. (2020a). Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th conference on computational natural language learning*, p. 431–441.

BOROS E., NGUYEN N. K., LEJEUNE G. & DOUCET A. (2022). Assessing the impact of ocr noise on multilingual event detection over digitised documents. *International Journal on Digital Libraries*, p. 1–26.

BOROŞ E., ROMERO V., MAARAND M., ZENKLOVÁ K., KŘEČKOVÁ J., VIDAL E., STUTZMANN D. & KERMORVANT C. (2020b). A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters. In *2020 17th International conference on frontiers in handwriting recognition (ICFHR)*, p. 79–84 : IEEE.

EHRMANN M., HAMDI A., LINHARES PONTES E., ROMANELLO M. & DOUVET A. (2023). A Survey of Named Entity Recognition and Classification in Historical Documents. *ACM Computing Surveys*.

EHRMANN M., ROMANELLO M., BIRCHER S. & CLEMATIDE S. (2020a). Introducing the CLEF 2020 HIPE shared task : Named entity recognition and linking on historical newspapers. In J. M. JOSE, E. YILMAZ, J. MAGALHÃES, P. CASTELLS, N. FERRO, M. J. SILVA & F. MARTINS, Éds., *Advances in information retrieval*, p. 524–532, Cham : Springer International Publishing.

EHRMANN M., ROMANELLO M., CLEMATIDE S., STRÖBEL P. B. & BARMAN R. (2020b). Language resources for historical newspapers : the impresso collection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 958–968.

EHRMANN M., ROMANELLO M., FLÜCKIGER A. & CLEMATIDE S. (2020c). Overview of clef hipe 2020 : Named entity recognition and linking on historical newspapers. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, p. 288–310 : Springer.

GARCÍA-DURÁN A., DUMANČIĆ S. & NIEPERT M. (2018). Learning sequence encoders for temporal knowledge graph completion. *arXiv preprint arXiv :1809.03202*.

HAMDI A., LINHARES PONTES E., BOROS E., NGUYEN T. T. H., HACKL G., MORENO J. G. & DOUCET A. (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2328–2334.

HAMDI A., PONTES E. L., SIDERE N., COUSTATY M. & DOUCET A. (2022). In-depth analysis of the impact of ocr errors on named entity recognition and linking. *Natural Language Engineering*, p. 1–24.

HOULSBY N., GIURGIU A., JASTRZEBSKI S., MORRONE B., DE LAROUSSILHE Q., GESMUNDO A., ATTARIYAN M. & GELLY S. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, p. 2790–2799 : PMLR.

LEBLAY J. & CHEKOL M. W. (2018). Deriving validity time in knowledge graph. In *Companion Proceedings of the The Web Conference 2018*, p. 1771–1776.

LINHARES PONTES E., CABRERA-DIEGO L. A., MORENO J. G., BOROS E., HAMDI A., DOUCET A., SIDERE N. & COUSTATY M. (2022). Melhissa : a multilingual entity linking architecture for historical press articles. *International Journal on Digital Libraries*, **23**(2), 133–160.

NGUYEN N. K., BOROS E., LEJEUNE G. & DOUCET A. (2020). Impact analysis of document digitization on event extraction. In *4th workshop on natural language for artificial intelligence (NL4AI 2020) co-located with the 19th international conference of the Italian Association for artificial intelligence (AI* IA 2020)*, volume 2735, p. 17–28.

OBERBICHLER S., BOROŞ E., DOUCET A., MARJANEN J., PFANZELTER E., RAUTIAINEN J., TOIVONEN H. & TOLONEN M. (2022). Integrated interdisciplinary workflows for research on historical newspapers : Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology*, **73**(2), 225–239.

PFEIFFER J., VULIĆ I., GUREVYCH I. & RUDER S. (2020). MAD-X : An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 7654–7673, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.617.

RASTAS I., RYAN Y. C., TIIHONEN I., QARAEI M., REPO L., BABBAR R., MÄKELÄ E., TOLONEN M. & GINTER F. (2022). Explainable publication year prediction of eighteenth century texts with the bert model. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, p. 68–77.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

REIMERS N. & GUREVYCH I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4512–4525.

ROMANELLO M., NAJEM-MEYER S. & ROBERTSON B. (2021). Optical character recognition of 19th century classical commentaries : the current state of affairs. In *The 6th International Workshop on Historical Document Imaging and Processing*, p. 1–6.

VAN STRIEN D., BEELEN K., ARDANUY M. C., HOSSEINI K., MCGILLIVRAY B. & COLAVIZZA G. (2020). Assessing the impact of ocr quality on downstream nlp tasks.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

WANG X., GAO T., ZHU Z., ZHANG Z., LIU Z., LI J. & TANG J. (2021). Kepler : A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, **9**, 176–194.

WANG X., SHEN Y., CAI J., WANG T., WANG X., XIE P., HUANG F., LU W., ZHUANG Y., TU K. *et al.* (2022). Damo-nlp at semeval-2022 task 11 : A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv :2203.00545*.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Oui mais.. ChatGPT peut-il identifier des entités dans des documents historiques ?

Carlos-Emiliano González-Gallardo[1] Emanuela Boros[2*] Nancy Girdhar[1]
Ahmed Hamdi[1] Jose G. Moreno[3] Antoine Doucet[1]
(1) La Rochelle Université, L3i, 17000 La Rochelle, France
(2) EPFL, Digital Humanities Laboratory, Lausanne, Suisse
(3) Université de Toulouse, IRIT UMR 5505 CNRS, 31000 Toulouse, France
carlos.gonzalez_gallardo@univ-lr.fr,{prénom.nom}@univ-lr.fr,
emanuela.boros@epfl.ch, jose.moreno@irit.fr

RÉSUMÉ

Les modèles de langage de grande taille (LLM) sont exploités depuis plusieurs années maintenant, obtenant des performances de l'état de l'art dans la reconnaissance d'entités à partir de documents modernes. Depuis quelques mois, l'agent conversationnel ChatGPT a suscité beaucoup d'intérêt auprès de la communauté scientifique et du grand public en raison de sa capacité à générer des réponses plausibles. Dans cet article, nous explorons cette compétence à travers la tâche de reconnaissance et de classification d'entités nommées (NERC) dans des sources primaires (des journaux historiques et des commentaires classiques) d'une manière *zero-shot* et en la comparant avec les systèmes de pointe basés sur des modèles de langage. Nos résultats indiquent plusieurs lacunes dans l'identification des entités dans le texte historique, allant de l'uniformité des directives d'annotation des entités, de la complexité des entités et du *code-switching*, à la spécificité de l'invite. De plus, comme prévu, la relative absence sur Internet des archives historiques et donc dans le corpus d'entraînement de ChatGPT a également un impact sur sa performance.

ABSTRACT

Yes but.. Can ChatGPT Identify Entities in Historical Documents ?

Large language models (LLM) have been leveraged for several years now, obtaining state-of-the-art performance in recognizing entities from modern documents. For the last few months, the conversational agent ChatGPT has "prompted" a lot of interest in the scientific community and public due to its capacity of generating plausible-sounding answers. In this paper, we explore this ability by probing it in the named entity recognition and classification (NERC) task in primary sources (i.e., historical newspapers and classical commentaries) in a zero-shot manner and by comparing it with state-of-the-art LM-based systems. Our findings indicate several shortcomings in identifying entities in the historical text that range from the consistency of entity annotation guidelines, entity complexity, and code-switching, to the specificity of prompting. Moreover, as expected, the inaccessibility of historical archives to the public (and thus on the Internet) also impacts its performance.

MOTS-CLÉS : Reconnaissance et classification d'entités nommées, Modèles de langage de grande taille, Transformeur génératif pré-entraîné, Documents historiques.

KEYWORDS: Named entity recognition and classification, Large language models, Generative pretrained transformer, Historical documents.

∗. Ce travail a été réalisé à l'Université de La Rochelle, à La Rochelle, France.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 74–82.

1 Introduction

Depuis qu'OpenAI a lancé ChatGPT lors du trente-sixième colloque sur les systèmes neuronaux de traitement de l'information (NeurIPS) en novembre 2022, sa capacité à fournir des réponses d'apparence humaines et plausibles a rendu le modèle extrêmement populaire au-delà de la communauté des chercheurs, avec plus d'un million d'utilisateurs en moins d'une semaine. ChatGPT est un agent conversationnel basé sur GPT-3.5 (Transformeur génératif pré-entraîné), un grand modèle de langage avec plus de 175 milliards de paramètres (Ouyang *et al.*, 2022). Étant donné sa grande popularité et son accessibilité, la question de savoir comment ce modèle hautement médiatisé se comporte dans différentes tâches de traitement du langage naturel (TAL) s'est déjà posée dans plusieurs domaines (Biswas, 2023; Pavlik, 2023).

Les modèles de langage de grande taille (LLM) sont exploités depuis plusieurs années maintenant, obtenant des performances de pointe dans la majorité des tâches de TAL, en étant généralement affinés sur des tâches en aval telles que la reconnaissance et la classification d'entités nommées (NERC) et moins dans des paramètres *zero-shot* (Li *et al.*, 2020). Ainsi, pour la reconnaissance et la classification d'entités nommées, mais aussi de manière générale, les efforts sont consacrés à la manière de transférer efficacement les connaissances pour l'adaptation au domaine en développant des systèmes robustes inter-domaines et en explorant l'apprentissage *zero-shot* ou *few-shot* pour traiter la cohérence et l'inadéquation des domaines et des annotations dans des contextes inter-domaines (Ehrmann *et al.*, 2020c, 2022). Simultanément, dans les documents historiques comme la presse ancienne, la tâche de NERC est confrontée à de nouveaux défis, outre l'hétérogénéité des domaines, tels que le bruit des entrées, la dynamique de la langue et le manque de ressources (Ehrmann *et al.*, 2021; Schweter & Baiter, 2019; González-Gallardo *et al.*, 2023; Boroş *et al.*, 2020; Najem-Meyer & Romanello, 2022; Schweter *et al.*, 2022; Boros *et al.*, 2022).

Dans ce court travail préliminaire, nous menons une étude exploratoire pour étudier le potentiel de ChatGPT, qui a été entraîné sur une quantité massive de données Internet (e.g., Common Crawl, WebText2, Wikipedia) (Brown *et al.*, 2020) et des ensembles de données d'invites pour l'apprentissage par renforcement à partir des préférences humaines (RLHF) (Ouyang *et al.*, 2022). Nous menons cette étude sur la tâche de NERC dans une configuration *zero-shot* et en comparant les performances de ChatGPT avec celles des systèmes de pointe.

2 Méthodologie

Nous avons suivi une approche de type *zero-shot* pour récupérer les entités nommées extraites par ChatGPT via son interface web officielle [1] entre le 11 janvier et le 7 février 2023. Une mise à jour a été publiée le 30 janvier pour améliorer la factualité et les capacités mathématiques du modèle [2], cependant, nous n'avons pas perçu de différence en ce qui concerne la capacité du modèle à détecter les entités.

1. https://chat.openai.com
2. https://help.openai.com/en/articles/6825453-chatgpt-release-notes

Jeux de données Nous avons sélectionné trois collections de documents historiques englobant une période d'environ 200 ans. Ces ensembles comprennent des commentaires classiques et des journaux historiques provenant de bibliothèques impliquées dans divers projets de recherche internationaux, notamment *NewsEye* [3] et *impresso* [4].

	Tokens	Entités	PERS	LOC	ORG	HumanProd	TIME	SCOPE
NewsEye	30 458	1 298	463	597	217	21	–	–
hipe-2020	48 854	1 600	502	854	130	61	53	–
ajmc	5 390	360	139	9	–	80	3	129

TABLE 1 – Description statistique des jeux de donnés (PERS = personne, LOC = emplacement, ORG = organisation, HumanProd = ouvrage/production, TIME = date/intervalle, SCOPE = partie spécifique de l'ouvrage).

Le jeu de données NewsEye (Hamdi *et al.*, 2021a) a été collecté auprès des bibliothèques nationales de France (BnF), d'Autriche (ONB) et de Finlande (NLF) [5]. Il se compose de quatre corpus (français, allemand, finnois et suédois). Le corpus français est constitué de textes provenant des archives numérisées de neuf journaux, notamment *L'Oeuvre*, *La Fronde*, *La Presse*, *Le Matin*, *Marie-Claire*, *Ce soir*, *Marianne*, *Paris Soir* et *Regards*, couvrant la période de 1854 à 1946.

Le jeu de données hipe-2020 (Ehrmann *et al.*, 2020b) est composé d'articles de journaux suisses, luxembourgeois et états-uniens en français, allemand et anglais couvrant les XIX^{e} et XX^{e} siècles. Il a été collecté principalement auprès de la Bibliothèque nationale suisse (BN), de la Bibliothèque nationale du Luxembourg (BnL), de la Médiathèque et des Archives d'Etat du Valais et des Archives économiques suisses (AES) [6] dans le cadre du projet *impresso*.

Le jeu de données ajmc (Romanello *et al.*, 2021) se compose de commentaires classiques provenant du projet *Ajax Multi-Commentary*. Ce projet rassemble des commentaires numérisés du XIX^{e} siècle rédigés en français, en allemand et en anglais. Ces commentaires offrent une analyse approfondie ainsi qu'une explication détaillée de la tragédie grecque *Ajax* de Sophocle.

Les collections mentionnées ont été annotées à deux niveaux de granularité (grossier et fin) pour la tâche de NERC. Les entités identifiées comprennent des catégories universelles telles que les personnes, les lieux et les organisations ; ainsi que des entités spécifiques au domaine telles que des références bibliographiques à la littérature primaire et secondaire. Les données ont été réparties en ensembles d'entraînement, de développement et de test.

Au cours de ce travail préliminaire, seules les partitions de test en français avec une granularité grossière ont été prises en compte pour simplifier la complexité des invites. Le tableau 1 fournit des informations sur le nombre et le type d'entités identifiées dans les ensembles de données spécifiés.

Nous avons défini les trois invites présentées dans le tableau 2 en fonction des différents types d'entités entre les ensembles de données et pour respecter la casse des étiquettes correspondantes. Même si le format IOB/BIO [7] est explicitement demandé pour chaque mot, la tokénisation par ChatGPT était

3. https://www.newseye.eu/
4. https://impresso-project.ch/
5. BnF : https://bnf.fr ; ONB : https://onb.ac.at ; NLF : https://kansalliskirjasto.fi
6. BN : https://www.nb.admin.ch ; BnL : https://bnl.public.lu ; AES : https://wirtschaftsarchiv.ub.unibas.ch
7. https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging)

incohérente avec les fichiers de jeux de données tokénisés IOB. Ainsi, à des fins d'évaluation, une vérification a été nécessaire pour assurer la cohérence de l'évaluation.

Des études antérieures ont démontré que l'inclusion d'instructions détaillées et d'exemples dans les invites ont un impact sur la qualité des résultats obtenus (Wang & Jin, 2023). Cependant, l'introduction d'une telle complexité dans les invites aurait été en contradiction avec notre étude de NERC dans une configuration *zero-shot*. Par conséquent, nous avons pris la décision de maintenir les invites aussi simples que possible afin de minimiser toute influence potentielle sur les résultats.

NewsEye	hipe-2020	ajmc
Quels sont les emplacements (LOC), les personnes (PER), organisations (ORG) et productions humaines (HumanProd) présents dans le texte historique suivant ? ***{PHRASE}*** Répondez, pour chaque mot, en utilisant IOB ou BIO séparé par tabulation. Si un mot n'a pas d'entité, ajoutez O.	Quels sont les emplacements (loc), les personnes (pers), organisations (org), produits (prod) et périodes (temps) présents dans le texte historique suivant ? ***{PHRASE}*** Répondez, pour chaque mot, en utilisant IOB ou BIO séparé par tabulation. Si un mot n'a pas d'entité, ajoutez O.	Quels sont les emplacements (loc), les personnes (pers), les périodes de temps (date), œuvres humaines (HumanProd), objets physiques (objet) et partie spécifique des travaux (étendue - *scope*) présents dans le texte historique suivant ? ***{PHRASE}*** Répondez, pour chaque mot, en utilisant IOB ou BIO séparé par tabulation. Si un mot n'a pas d'entité, ajoutez O.

TABLE 2 – Invites du jeu de données utilisées pour collecter les prédictions.

3 Résultats

Le tableau 3 présente les performances de ChatGPT par rapport au NERC avec une granularité grossière (types d'entités de haut niveau) en termes de précision (P), de rappel (R) et de F-mesure (F1) à micro-niveau *strict* et *fuzzy*, évaluées avec CLEF-HIPE-2020-scorer [8]. Nous présentons également les performances de deux systèmes NERC de l'état de l'art basés sur modèles de langage qui ont été entraînés avec l'ensemble d'entraînement des jeux de données figurant à la section 2.

	NewsEye			hipe-2020			ajmc		
	P	R	F1	P	R	F1	P	R	F1
strict									
Stacked NERC	**75,0**	70,6	**72,7**	–	–	–	–	–	–
Temporal NERC	–	–	–	**76,5**	**76,5**	**76,5**	**84,8**	**83,9**	**84,4**
ChatGPT	70,9	**72,3**	71,6	32,5	50,0	39,4	21,8	26,1	23,8
fuzzy									
Stacked NERC	**85,4**	**80,5**	**82,9**	–	–	–	–	–	–
Temporal NERC	–	–	-	**86,7**	**86, 7**	**86,7**	**90,2**	**89,2**	**89,7**
ChatGPT	77,8	79,4	78,6	49,0	75,4	59,4	25,5	30,6	27,8

TABLE 3 – Résultats comparatifs utilisant les trois jeux de données (micro).

Stacked NERC est basé sur le modèle pré-entraîné BERT proposé par (Devlin *et al.*, 2019) avec un empilement de deux blocs de Transformeur au-dessus, finalisé par une couche de prédiction à champ aléatoire conditionnel (CRF). *Stacked NERC* a démontré une performance accrue dans les documents historiques, tout en ne dégradant pas la performance sur les données modernes (Boros *et al.*, 2020; Boroş *et al.*, 2020). La même architecture a été utilisée comme base de référence dans la description du jeu de données NewsEye (Hamdi *et al.*, 2021a).

8. https://github.com/hipe-eval/HIPE-scorer

Temporal NERC s'appuie sur *Stacked NERC*, et il comprend une amélioration au niveau des données en exploitant des graphes de connaissances temporelles pour générer des informations temporelles contextuelles supplémentaires et une amélioration au niveau du modèle qui incorpore ces informations avec des contextes regroupés par la moyenne (González-Gallardo *et al.*, 2023). *Temporal NERC* a prouvé l'importance de la temporalité pour les journaux historiques et les commentaires classiques, en fonction des intervalles de temps et du taux d'erreur de numérisation.

D'après le tableau 3, il est clair que la capacité de ChatGPT à identifier des entités nommées dépend fortement de l'ensemble de données et du type d'entités. Des performances nettement inférieures en termes de F1 sont observées pour `ajmc`, avec une diminution de plus de 71% pour la métrique *strict* et de plus de 69% pour la métrique *fuzzy*. Pour `hipe-2020`, les performances ont diminué de manière moins radicale, avec plus de 48% et 31,48% respectivement. En ce qui concerne `NewsEye`, les scores sont légèrement similaires, avec une baisse d'environ 1,5% et un peu plus de 5% respectivement. Bien que les résultats soient globalement équilibrés, nous observons également un rappel plus élevé dans le cas de `hipe-2020`, ce qui pourrait indiquer que la complexité de l'annotation des entités permet à ChatGPT de les détecter, mais pas de les classer correctement [9]. La section suivante présente une analyse des faiblesses de ChatGPT dans le processus de reconnaissance des entités dans des textes historiques, en tenant compte de la définition des entités nommées, de leur complexité et des erreurs liées au processus de numérisation.

4 Analyse des erreurs

Définition des entités nommées L'annotation d'entités nommées suit des directives d'annotation bien définies (*guidelines*) pour décrire la nature et les limites des types d'entités, cependant il est nécessaire de faire confiance à l'intuition et à la conscience linguistiques de l'annotat·eur·rice (Hamdi *et al.*, 2021b; Ehrmann *et al.*, 2020a; Romanello & Najem-Meyer, 2022). Alors que les définitions des types d'entités universelles sont similaires entre les directives d'annotation, les types d'entités spécifiques à un domaine sont très variables. Celles de `hipe-2020` (Ehrmann *et al.*, 2020a) et `NewsEye` (Hamdi *et al.*, 2021b) définissent une entité de type « production humaine » comme étant tout ce qui est diffusé dans la presse, à la radio ou à la télévision, comme les journaux, les magazines, les émissions ou les catalogues de vente (e.g., *Die Zeit*, *Le Figaro*, *Le sept à huit*, *La ferme célébrités*) et excluent les produits médiatiques tels que les films et les téléfilms, ainsi que les doctrines politiques, philosophiques et religieuses/sectaires, comme *Der Sozialismus*, *Theheravada Buddhismus*, *Le socialisme*, *Le bouddhisme theravâda*. De même, l'entité de type « ouvrage » est décrite par les directives d'annotation pour `ajmc` (Romanello & Najem-Meyer, 2022) comme une entité désignant une création humaine, qu'elle soit intellectuelle ou artistique, qui peut être désignée par son titre. Pour le FRBR [10], une œuvre *« est une création intellectuelle ou artistique distincte »*, notamment les œuvres littéraires, les œuvres religieuses, les éditions de sources papyrologues et épigraphiques (e.g., *IG2*, *P.Oxy 1.119*), et les revues.

Étant donné que nous explorons la tâche de NERC dans une configuration *zero-shot* et bien que nous convenions que l'uniformité des annotations est une préoccupation majeure en raison de l'ambiguïté de la langue, nous ne pouvons que supposer que la variété des définitions crée nécessairement une difficulté pour ChatGPT, qui n'est pas entraîné sur des jeux de données annotés par des annotateurs.

9. Toutes les prédictions sont disponibles sur `https://github.com/cic4k/NERC_ChatGPT`.
10. `https://www.oclc.org/research/activities/frbr.html`

Complexité des entités Les directives d'annotation de `NewsEye` définissent à une entité nommée comme un objet du monde réel désignant un individu unique avec un nom propre. Historiquement, le nom d'une personne a joué un rôle influent en reflétant les attributs clés de son travail ou de sa vie, la majorité des entités ont également été annotées avec l'intitulé du métier de la personne. Prenons pour illustration l'exemple suivant : *« Becthoven. - Par le Quatuor de la fondation Beethoven : MM.A, Géioso1er violon ; A. Tracol ; 2e violon ;P. Monteux. aito, F. Schnéklud, violoncelle ; César Geloso, pianiste »*. « César Geloso, le pianiste » est considéré comme une entité « personne », cependant, ChatGPT n'a pas été capable de détecter au-delà de la mention du prénom et du nom. Si des métiers doivent être détectés à l'intérieur des entités, nous supposons que des informations supplémentaires doivent être ajoutées dans l'invite.

L'écriture bicamérale semble également problématique. Alors qu'il est courant que les noms d'organisations et les titres d'articles de journaux soient tout en majuscules (e.g., *LE SPORT*, *NOUVELLES BREVES*), ChatGPT les a tous identifiés comme des organisations. Les démonymes, les lieux et les personnes (e.g., *Carcassonnais, Russe, Mexicains, « Italien, 17 3/4 », « Japon 1899, 73 », « Portugais 3 %, 2 1/4 », « Russe 1906 »*) ont également posé des problèmes. Il n'est cependant pas clair si la confusion provient du fait que ces mots commencent par des majuscules ou si elle est due à un autre élément de contexte, car cette limitation se retrouve couramment dans des systèmes de NERC de l'état de l'art.

Les directives d'annotation de `NewsEye` et `impresso` considèrent des adresses telles que *« 130, rue de la Courselle »* et *« 56, rue de la Montagne-Sainte-Geneviève, 5e »* comme des emplacements, mais ChatGPT ne semble pas capturer ce niveau fin de granularité. Étant donné que le mot « emplacement » définit l'espace d'une manière sémantiquement plus générique qu'un lieu où une position spécifique, une adresse fait référence aux particularités d'un lieu qui, s'il n'est pas spécifié dans l'invite, ne peut pas être identifié correctement par ChatGPT.

Erreurs d'océrisation D'un point de vue quantitatif, ChatGPT n'a identifié que 7% des entités nommées comportant des erreurs de reconnaissance optique de caractères (OCR) dans le jeu de données `ajmc`, tandis que *Temporal NERC* a correctement identifié environ 40% de ces entités bruitées. *Temporal NERC* a reconnu des entités nommées dont jusqu'à 70% des caractères avaient subi une erreur d'océrisation (i.e., suppression, insertion et substitution).

Cependant, ces erreurs ne devraient pas dépasser 20% pour que les entités puissent être reconnuespar ChatGPT. Notamment, ChatGPT s'est montré incapable de reconnaître les entités nommées comportant des erreurs de segmentation. Par exemple, dans *« 13659 - 4360.Hxépra. ... . elloug. Ulysse paraît faire allusion à l'amertume des peroles que vient de prononcer À gamemnon ; Agamemnon répond comme si Ulysse avait eu en vue l'amertume de ses propres remontrances, 4866. 'EvOaë' Tlouet, j'en arriverai là, c'est-à-dire, je mourrai, .Dindorf : Kai'* αὐτὸς ἴξομαι πρὸς τὸ ϑάπτειν αὐτόν. », le nom d'une personne *Agamemnon* a été correctement identifié. En revanche, *« À gamemnon »* n'a pas été reconnu. Il n'est pas surprenant de constater que lorsque l'entité est fortement impactée par le processus d'océrisation, comme *Becthoven* au lieu de *Beethoven* dans *« XIe Quatuor (op. 95) Becthoven »*, ChatGPT ne parvient pas à détecter ces entités. Enfin, en raison d'une quantité abondante de bruit, comme dans *« m _'¦'— "..'i -i' i' —'i m Nota. Les Avis à insérer dans cette Feuille, qui ne seront pas remis au Bureau le mardi matin, à neuf-heures au plus tard, seront renvoyés irrévocablement _. [...] »*, ChatGPT « abandonne » avec cette déclaration : *« Le texte que vous avez fourni n'est pas dans un langage cohérent, et il est difficile de comprendre ce qu'il essaie de communiquer. Le texte contient principalement de la ponctuation et des caractères spéciaux sans entités reconnaissables. »*.

Alternance codique L'alternance codique (*code-switching*) est un phénomène qui consiste à alterner entre deux ou plusieurs langues dans une même phrase, une même expression ou un même mot. Le biais monolingue dans le multilinguisme est un type de biais qui peut se produire dans les modèles de langage (Talat *et al.*, 2022; Choudhury & Deshpande, 2021). `ajmc` présente de l'alternance codique entre le français et le grec ancien. GPT-3.5 a été entraîné dans plus de 100 langues, l'anglais étant surreprésenté avec 93% du nombre de mots, 1,82% pour le français et 0,032% pour le grec moderne, tandis que le grec ancien n'est pas représenté (Brown *et al.*, 2020). ChatGPT a été entraîné à l'aide d'autres ensembles de données pour un réglage fin, un retour humain et des messages-guides (Ouyang *et al.*, 2022), mais on peut s'attendre à une distribution similaire des langues. Ainsi, pour un exemple tel que *« À la marge d'un exemplaire de Sophocle, on lit la traduction suivante de ces deux vers, due à notre ÆRacine : " O mon fils, sois un jour plas beareux que ton pèren ; Da reste avec honneur tu peux lui ressembler". »*, ChatGPT répond *« Je suis désolé, mais je ne suis pas en mesure de comprendre le texte que vous avez fourni. Le texte semble être un mélange de grec ancien et de français, avec quelques références à la littérature ancienne et des annotations, ce qui rend difficile l'extraction d'informations significatives. De plus, le texte n'est pas un texte historique, mais plutôt un texte littéraire, ce qui rend également difficile l'extraction d'entités historiques. »*.

5 Conclusions et perspectives

ChatGPT rencontre plusieurs difficultés dans la reconnaissance des entités dans les documents historiques, qui relèvent de la cohérence des directives d'annotation des entités, de la complexité des entités, du multilinguisme, de l'alternance codique et de la spécificité de l'invite. De plus, alors qu'une quantité sans précédent de documents historiques est disponible en format numérique, peu de choses sont disponibles gratuitement avec de nombreuses archives historiques qui restent inaccessibles au public. Par exemple, les sources primaires telles que les journaux et les articles de magazines (comme dans le cas de la majorité des ensembles de données de cette étude) sont disponibles à la fois en ligne et en bibliothèque, mais néanmoins, elles sont généralement filigranées ou derrière un mur payant.

Par conséquent, ChatGPT est « ignorant » (pour l'instant) de ces connaissances, ce qui contribue au degré de confusion du modèle en ce qui concerne les documents historiques. Même si ces ressources deviennent accessibles et que les systèmes basés sur les LLM améliorent leurs capacités de « compréhension » des documents historiques, leur mise en œuvre dans les bibliothèques numériques devra être prise avec précaution pour éviter les réponses biaisées et hors domaine.

Déclaration éthique

Bien qu'il puisse générer un texte à la consonance plausible, le contenu généré par ChatGPT n'est pas nécessairement vrai. Néanmoins, nous considérons qu'il ne concerne pas la tâche de reconnaissance d'entités nommées, car nous n'ajoutons aucune autre considération éthique que celles posées par ChatGPT. Nous sommes conscients de la position intentionnelle (Dennett, 2009) de termes comme « abandonne », « compréhension » et « ignorant » lorsqu'ils sont appliqués à un agent conversationnel, cependant, dans ce papier, nous les adoptons pour souligner la capacité de ChatGPT à interagir avec un utilisateur.

Remerciements

Ce travail a été soutenu par les projets ANNA (2019-1R40226), TERMITRAD (AAPR2020-2019-8510010), Pypa (AAPR2021-2021-12263410) et Actuadata (AAPR2022-2021-17014610) financés par la Région Nouvelle-Aquitaine, France.

Références

BISWAS S. (2023). Chatgpt and the future of medical writing.

BOROS E., GONZÁLEZ-GALLARDO C.-E., GIAMPHY E., HAMDI A., MORENO J. G. & DOUCET A. (2022). Knowledge-based contexts for historical named entity recognition & linking. *Conference and Labs of the Evaluation Forum (CLEF 2020)*.

BOROŞ E., HAMDI A., PONTES E. L., CABRERA-DIEGO L.-A., MORENO J. G., SIDERE N. & DOUCET A. (2020). Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th conference on computational natural language learning*, p. 431–441.

BOROS E., PONTES E. L., CABRERA-DIEGO L. A., HAMDI A., MORENO J. G., SIDÈRE N. & DOUCET A. (2020). Robust named entity recognition and linking on historical multilingual documents. In *Conference and Labs of the Evaluation Forum (CLEF 2020)*, volume 2696, p. 1–17 : CEUR-WS Working Notes.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

CHOUDHURY M. & DESHPANDE A. (2021). How linguistically fair are multilingual pre-trained language models ? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, p. 12710–12718.

DENNETT D. (2009). Intentional Systems Theory. In *The Oxford Handbook of Philosophy of Mind*. Oxford University Press. DOI : 10.1093/oxfordhb/9780199262618.003.0020.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

EHRMANN, WATTER, ROMANELLO, CLEMATIDE & FLÜCKIGER (2020a). *Impresso Named Entity Annotation Guidelines*. DOI : 10.5281/zenodo.3604227.

EHRMANN M., HAMDI A., PONTES E. L., ROMANELLO M. & DOUCET A. (2021). Named entity recognition and classification on historical documents : A survey. *arXiv preprint arXiv :2109.11406*.

EHRMANN M., ROMANELLO M., CLEMATIDE S., STRÖBEL P. & BARMAN R. (2020b). Language resources for historical newspapers : the impresso collection.

EHRMANN M., ROMANELLO M., FLÜCKIGER A. & CLEMATIDE S. (2020c). Extended overview of clef hipe 2020 : named entity processing on historical newspapers. In *CEUR Workshop Proceedings*, volume 2696 : CEUR-WS.

EHRMANN M., ROMANELLO M., NAJEM-MEYER S., DOUCET A. & CLEMATIDE S. (2022). Extended overview of hipe-2022 : Named entity recognition and linking in multilingual historical documents. In *Proceedings of the 13th International Conference of the CLEF Association (Lecture Notes in Computer Science)*, volume 13390 : Springer.

GONZÁLEZ-GALLARDO C.-E., BOROS E., GIAMPHY E., HAMDI A., MORENO J. G. & DOUCET A. (2023). Injecting temporal-aware knowledge in historical named entity recognition. In *Advances in Information Retrieval : 45th European Conference on IR Research, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, p. 65–79 : Springer.

HAMDI A., LINHARES PONTES E., BOROS E., NGUYEN T. T. H., HACKL G., MORENO J. G. & DOUCET A. (2021a). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2328–2334.

HAMDI A., PONTES E. L. & DOUCET A. (2021b). Annotation Guidelines for Named Entity Recognition, Entity Linking and Stance Detection. DOI : 10.5281/zenodo.4574199.

LI J., SUN A., HAN J. & LI C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, **34**(1), 50–70.

NAJEM-MEYER S. & ROMANELLO M. (2022). Page layout analysis of text-heavy historical documents : a comparison of textual and visual approaches. In *Proceedings of the Computational Humanities Research Conference 2022 Antwerp, Belgium, December 12-14, 2022.*, p. 36–54.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A. *et al.* (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv :2203.02155*.

PAVLIK J. V. (2023). Collaborating with chatgpt : Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, p. 10776958221149577.

ROMANELLO M. & NAJEM-MEYER S. (2022). *Guidelines for the Annotation of Named Entities in the Domain of Classics*. DOI : 10.5281/zenodo.6368101.

ROMANELLO M., NAJEM-MEYER S. & ROBERTSON B. (2021). Optical character recognition of 19th century classical commentaries : The current state of affairs. In *The 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, p. 1–6, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3476887.3476911.

SCHWETER S. & BAITER J. (2019). Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, p. 96–103, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-4312.

SCHWETER S., MÄRZ L., SCHMID K. & ÇANO E. (2022). hmbert : Historical multilingual language models for named entity recognition. *Conference and Labs of the Evaluation Forum (CLEF 2020)*.

TALAT Z., NÉVÉOL A., BIDERMAN S., CLINCIU M., DEY M., LONGPRE S., LUCCIONI S., MASOUD M., MITCHELL M., RADEV D. *et al.* (2022). You reap what you sow : On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, p. 26–41.

WANG S. & JIN P. (2023). A brief summary of prompting in using gpt models.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

De l'interprétabilité des dimensions à l'interprétabilité du vecteur : parcimonie et stabilité

Simon Guillot[1] Thibault Prouteau[1] Nicolas Dugué[1]
(1) LIUM, Le Mans, France
prenom.nom@univ-lemans.fr

RÉSUMÉ

Les modèles d'apprentissage de plongements parcimonieux (SPINE, SINr) ont pour objectif de produire un espace dont les dimensions peuvent être interprétées. Ces modèles visent des cas d'application critiques du traitement de la langue naturelle (*e.g.* usages médicaux ou judiciaires) et une utilisation des représentations dans le cadre des humanités numériques. Nous proposons de considérer non plus seulement l'interprétabilité des dimensions de l'espace de description, mais celle des vecteurs de mots en eux-mêmes. Pour cela, nous introduisons un cadre d'évaluation incluant le critère de stabilité, et redéfinissant celui de la parcimonie en accord avec les théories psycholinguistiques. Tout d'abord, les évaluations en stabilité indiquent une faible variabilité sur les modèles considérés. Ensuite, pour redéfinir le critère de parcimonie, nous proposons une méthode d'éparsification des vecteurs de plongements en gardant les composantes les plus fortement activées de chaque vecteur. Il apparaît que pour les deux modèles SPINE et SINr, de bonnes performances en similarité sont permises par des vecteurs avec un très faible nombre de dimensions activées. Ces résultats permettent d'envisager l'interprétabilité de représentations éparses sans remettre en cause les performances.

ABSTRACT

From dimension to vector interpretability : sparseness and stability

Sparse word embeddings models (SPINE, SINr) are designed to embed words in interpretable dimensions. These models are useful for critical downstream tasks in natural language processing (*e.g.* medical or legal NLP), and digital humanities applications. We propose to shift attention from the interpretability of dimensions to the interpretability of word vectors. We thus introduce stability to the interpretability framework, and redefine sparseness. First, stability evaluations show some variability for both models. Then, our sparsifiaction approach redefines the sparseness criterion by keeping only a limited number of components among the strongest in each vector. Both SPINE and SINr show interesting performances on the similarity task with very few activated dimensions. These results are encouraging and pave the way towards intrinsically interpretable word embeddings.

MOTS-CLÉS : Sémantique distributionnelle, traits sémantiques, interprétabilité, plongements.

KEYWORDS: Distributional semantics, semantic features, interpretability, word embeddings.

1 Introduction

L'interprétabilité (Rudin, 2019) telle qu'elle est définie dans la littérature pour les modèles de plongements lexicaux correspond à assurer la possibilité de trouver une cohérence sémantique (ou syntaxique) aux dimensions de l'espace de représentation (Murphy *et al.*, 2012; Senel *et al.*, 2018;

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 83–91.

Subramanian *et al.*, 2018; Prouteau *et al.*, 2022). Cette interprétabilité facilite également la mise en perspective des représentations du lexique avec des conceptions linguistiques de celui-ci. En effet, des dimensions de description sémantiquement cohérentes sont analysables comme traits sémantiques ou sèmes, unités utilisées dans une variété de modèles théoriques de représentation du sens allant de modèles d'inspiration cognitive (Jackendoff, 1983) à des modèles structuralistes ou néosaussuriens (Pottier, 1963; Rastier, 2009). Ces unités de sens varient quant à leur nombre, leur caractère universel ou relatif (à une langue, ou à un corpus), leur statut de primitives sémantiques et leur portée (le référent ou le concept) selon les cadres théoriques (Rastier, 2009).

Les modèles de plongements denses de l'état de l'art (Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Devlin *et al.*, 2018) ont permis des évolutions importantes dans le traitement automatique. Ils consistent à plonger le lexique dans des espaces de représentation denses aux dimensions opaques. Il est possible d'obtenir une compréhension a posteriori de ces modèles par le sondage (Rogers *et al.*, 2021) ou l'analyse des matrices de plongements (Shin *et al.*, 2018). Ces méthodes correspondent à l'*explicabilité* en apprentissage automatique. L'article fondateur de Murphy *et al.* (2012) ouvre la réflexion sur les représentations distributionnelles psycholinguistiquement plausibles. Les auteurs postulent un ensemble de contraintes sur l'espace de représentation : la parcimonie, la positivité, et la performance. La parcimonie est justifiée par la difficulté de couvrir l'ensemble du vocabulaire par un faible nombre de traits partagés. Ainsi, de nombreuses dimensions sont nécessaires, mais seules certaines sont activées pour la description d'un mot. La positivité tient au caractère non économique du stockage d'informations négatives, critère déjà utilisé pour des travaux psychologiquement motivés (Palmer, 1977) sur la factorisation de matrices d'éléments perceptifs (Lee & Seung, 1999). Les travaux sur les modèles de plongement épars ou parcimonieux connaissent des développements avec `SPOWV` (Faruqui *et al.*, 2015), `SPINE` (Subramanian *et al.*, 2018) et `SINr` (Prouteau *et al.*, 2021). Les deux premiers transforment un espace de représentation dense en espace épars, tandis que le troisième construit un espace épars depuis la matrice de cooccurrences.

La positivité et la parcimonie constituent des caractéristiques nécessaires à l'interprétabilité des représentations distributionnelles du lexique. Néanmoins, cette interprétabilité est cantonnée au niveau de la dimension d'un espace de représentation. En effet, les tests d'intrusion qui servent à la caractériser dans Murphy *et al.* (2012); Senel *et al.* (2018); Subramanian *et al.* (2018); Prouteau *et al.* (2022) examinent les dimensions une à une pour vérifier leur cohérence sémantique interne, ainsi que l'activation de la dimension pour un faible nombre d'items lexicaux.

Dans ce travail, nous proposons de concevoir également l'**interprétabilité au niveau du vecteur** de mot et non plus seulement au niveau des dimensions. Cette interprétabilité devient possible si un mot est décrit par une faible quantité de dimensions dont il est possible de faire sens. En effet, d'une part les tâches d'association utilisées pour établir des listes de traits sémantiques comme celles de (Garrard *et al.*, 2001; McRae *et al.*, 2005) indiquent l'ordre de grandeur d'une dizaine de traits par item à décrire en sommant les réponses de leurs annotateurs. D'autre part, (Miller, 1956; Peterson & Peterson, 1959) posent une limite pour la manipulation d'items lexicaux en mémoire de travail cohérente avec cet ordre de grandeur. Nous prenons ainsi le parti de définir cette quantité de descripteurs comme un horizon souhaitable pour les représentations interprétables, dans la mesure où de tels vecteurs restreints en dimensions sont plus manipulables par d'éventuels locuteurs pour en faire sens. Par ailleurs, nous proposons de définir la stabilité comme un critère supplémentaire à l'interprétabilité des vecteurs. Que ce soit dans le cadre des humanités numériques (Hellrich & Hahn, 2016a,b), ou dans des utilisations critiques du traitement automatique de la langue pour des applications juridiques ou médicales (Digan *et al.*, 2020), la reproductibilité des résultats est un critère essentiel. Or, l'entraînement non déterministe des modèles de plongements neuronaux cause

une certaine instabilité dans les représentations et les voisinages, même pour des représentations entraînées avec les mêmes hyperparamètres sur les mêmes données (Pierrejean, 2020). Rudin (2019) encourage à prioriser les approches interprétables sur les approches explicables qui souffrent de cette instabilité, même si la littérature voit émerger de récents efforts pour la création de méthodes d'explicabilité *a posteriori* déterministes (Zafar & Khan, 2021).

Ainsi, dans cet article, nous évaluons la capacité des modèles parcimonieux de l'état de l'art à se conformer aux contraintes d'interprétabilité des vecteurs que nous définissons : la stabilité de ces vecteurs, et la capacité à produire des représentations performantes avec un très faible nombre de dimensions activées, eu égard aux travaux en psycholinguistiques. Nous décrirons d'abord en Section 2 les deux modèles considérés dans nos expérimentations : `SPINE` et `SINr`. Puis nous détaillerons en Section 3 notre cadre expérimental : les corpus utilisés (OANC et BNC) pour apprendre les plongements et les jeux de données pour les évaluer sur des tâches de similarité. Nous introduirons également notre approche d'éparsification des vecteurs pour produire progressivement des modèles qui tendent vers une dizaine de dimensions activées par vecteur. Enfin, nous présenterons en Section 4 les résultats en stabilité et en similarité sur des représentations éparsifiées qui permettent de tendre vers la définition réactualisée de l'interprétabilité que nous proposons.

2 Les modèles interprétables

SPINE. `SPINE` est introduit par Subramanian *et al.* (2018) et permet, à partir d'une représentation dense obtenue avec `Word2Vec` (Mikolov *et al.*, 2013) ou `GloVe` (Pennington *et al.*, 2014), de produire un modèle parcimonieux dans un espace de plus grande taille (*e.g.* 1000 dimensions). Pour cela, `SPINE` est un auto-encodeur dont la couche cachée est en plus grande dimension que l'entrée à reconstruire. De plus, l'auto-encodeur est dit *k-sparse*, l'objectif est donc de n'activer que k neurones dans la couche cachée pour réussir à reconstruire l'entrée. Pour apprendre un tel modèle, trois fonctions de coût sont employées. La fonction de coût de reconstruction (*Reconstruction Loss*) pénalise la mauvaise reconstruction de l'entrée à partir de la représentation parcimonieuse fournie par la couche cachée. Pour imposer l'activation d'un faible nombre de neurones dans la couche cachée et donc la parcimonie de la matrice de plongements, les auteurs introduisent la fonction de coût sur la parcimonie moyenne (*Average Sparsity Loss*) et la fonction de coût sur la parcimonie partielle (*Partial Sparsity Loss*). Ces deux fonctions pénalisent un trop grand nombre d'activations tout en forçant les valeurs d'activation vers 0 ou 1. Le modèle `SPINE` présente plusieurs hyperparamètres : le niveau minimal de parcimonie, le nombre d'époques d'apprentissage et la dimension de la représentation.

SINr. Introduite dans Prouteau *et al.* (2021), `SINr` est une approche basée graphes pour la représentation distributionnelle du lexique. En appliquant une fenêtre glissante sur les phrases du corpus, la méthode extrait un graphe de co-occurrences pondéré (les sommets sont des mots et les poids des arêtes le nombre de co-occurrences observées). Prouteau *et al.* (2021) appliquent ensuite un algorithme de détection de communautés (l'algorithme de *Louvain* (Blondel *et al.*, 2008)) pour extraire des communautés de mots densément connectés donc fréquemment co-occurrents. À partir de cette partition du graphe en communautés, `SINr` calcule la distribution des liens de chaque sommet à travers les communautés détectées. Cela permet ainsi d'avoir une distribution de chaque mot sur les groupes de mots découverts de façon non supervisée par l'algorithme. Les plongements de mots produits sont naturellement parcimonieux, un sommet n'est pas connecté à toutes les communautés.

Le modèle SINr n'a qu'un seul hyperparamètre qui agit sur le nombre de communautés détectées.

3 Cadre expérimental

Modèles. Outre les modèles parcimonieux présentés Section 2, nous utilisons Word2Vec comme contrôle. Nous utilisons *Skipgram with Negative sampling* avec les paramètres décrits dans Levy & Goldberg (2014), la dimension des plongements de mots est de 300 avec une fenêtre de contexte de taille 5. Puisque le nombre de dimensions pour SPINE peut être fixé facilement contrairement à la méthode SINr (il dépend du nombre de communautés détectées), on fixe le nombre dimensions de SPINE en fonction de celles obtenues avec SINr. Les performances de la méthode SINr semblent optimales en réglant l'hyperparamètre agissant sur la détection de communautés à 50, aboutissant à 8454 dimensions pour BNC et 4460 pour OANC, ces nombres sont donc également choisis pour SPINE. Les plongements de mots SPINE sont appris à partir du modèle Word2Vec présenté ci-avant. Le niveau de parcimonie obtenu avec SPINE est peu sensible à l'hyperparamètre censé permettre de le régler. Ainsi, un grand nombre de modèles a été lancé et les résultats sont présentés sur le modèle qui obtient les meilleures performances sur la tâche d'évaluation en similarité avec une parcimonie (1000 époques aboutissant à 95% de parcimonie) permettant au modèle de subir le protocole que nous décrivons ci-après.

Protocole expérimental. Dans ces travaux, nous introduisons un protocole expérimental permettant d'évaluer l'interprétabilité au niveau des vecteurs. Pour cela, nous commençons par considérer le compromis performance-parcimonie. Nous supposons que des vecteurs plus épars sont à la fois plus interprétables et plus plausibles psycholinguistiquement. Pour travailler avec des modèles à la parcimonie contrôlée, nous introduisons notre approche d'éparsification des vecteurs : des représentations de chaque modèle de plongements sont construites en conservant pour chaque vecteur les k composantes ayant les valeurs les plus élevées, en variant k de 250 à 10. Les autres composantes sont fixées à 0. Si ce processus est naturel pour les modèles *a priori* parcimonieux et positifs, cela l'est moins pour Word2Vec, notre modèle de contrôle. Dans ce cas, nous avons fait le choix d'utiliser les k valeurs les plus élevées de la valeur absolue des composantes des vecteurs.

Pour contrôler la qualité de l'espace une fois ce protocole d'éparsification effectué, nous utilisons l'évaluation en similarité, soit la corrélation entre la similarité entre paires de mots dans les espaces appris et la similarité donnée par des humains. Les jeux de données sélectionnés correspondent dans la mesure du possible à une variété de relations : MEN, WS353, SCWS. Pour évaluer la stabilité des vecteurs produits par les deux modèles SPINE et SINr, le second critère d'interprétabilité que nous considérons, nous reproduisons ces expérimentations dix fois et présentons des résultats consolidés.

Puisque les jeux de données en similarité disponibles portent très majoritairement sur l'anglais, le British National Corpus (BNC) et l'Open American National Corpus (OANC) sont choisis comme corpus d'apprentissage. Le BNC compte environ 100 millions de tokens, tandis que la section de données textuelles de OANC en compte environ 11 millions. Les deux corpus sont composites en domaines et en genres textuels. Un prétraitement commun est appliqué aux deux corpus avec la bibliothèque spaCy : tokenisation et regroupement des entités nommées, suppression des mots de longueur inférieure ou égale à deux caractères comme approximation de suppression des mots vides, suppression de la ponctuation et des caractères numériques, et conservation dans le vocabulaire à représenter des items lexicaux apparaissant au moins vingt fois dans les corpus. Après ce prétraitement,

les données considérées sont : 20 814 types pour 4 millions de tokens pour OANC, 58 687 types pour 40 millions de tokens pour BNC.

4 Résultats et discussions

Résultats en stabilité. Les premiers résultats que nous considérons Table 1 ont le double emploi d'établir un point de référence sur les résultats en similarité des modèles avant l'application de notre protocole d'éparsification, et de proposer un indice sur la stabilité des représentations. Dix représentations sont produites avec chaque modèle (sur les mêmes données avec les mêmes hyperparamètres) et sont évaluées en similarité sur les trois jeux de données choisis.

BNC	MEN		WS353		SCWS	
	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ
Word2Vec	$0,72$	2e−3	$0,65$	5e−3	$0,57$	2e−3
SPINE	$0,65$	6e−3	$0,57$	1e−2	$0,60$	4e−3
SINr	$0,66$	6e−4	$0,63$	2e−3	$0,54$	1e−3
OANC	MEN		WS353		SCWS	
	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ	$\overline{Pearson}$	σ
Word2Vec	$0,43$	2e−3	$0,50$	5e−3	$0,46$	3e−3
SPINE	$0,36$	9e−3	$0,43$	1e−2	$0,39$	1e−2
SINr	$0,39$	8e−4	$0,44$	2e−3	$0,39$	2e−3

TABLE 1 – Stabilité des résultats en similarité sur le corpus BNC en haut, sur OANC en bas. Le coefficient de corrélation de Pearson moyen et l'écart-type sont donnés pour 10 exécutions.

Les trois modèles proposent des performances du même ordre de grandeur, quoique légèrement meilleures pour Word2Vec. SPINE semble légèrement plus instable que Word2Vec dont il est tributaire, et que SINr. La variabilité de SINr et de Word2Vec sont comparables. Quoique légère, l'instabilité constatée sur ces différents modèles nuit à leur interprétabilité.

Performance en similarité en fonction de la parcimonie. Les résultats récapitulés dans la Figure 1 permettent d'observer les performances en similarité en fonction du nombre de composantes conservées par vecteur suivant notre protocole d'éparsification. Il est d'abord notable que les trois modèles proposent des performances comparables à celles obtenues en Table 1 en similarité malgré l'éparsification, et ce jusqu'à ne conserver que cinquante dimensions activées par vecteur. Il apparaît même que pour SINr, le processus d'éparsification améliore les performances sur cette tâche, possiblement en retirant du bruit présent dans les représentations originales. La relation entre parcimonie et performance n'est donc pas nécessairement un compromis. De même, la conservation de résultats convenables malgré l'éparsification forcée sur notre modèle de contrôle, bien qu'il soit originellement dense, est un résultat intéressant sur la répartition de l'information sémantique dans celui-ci.

Pour tendre vers l'horizon de parcimonie fixée Section 1, soit dix dimensions, l'expérience est égale-

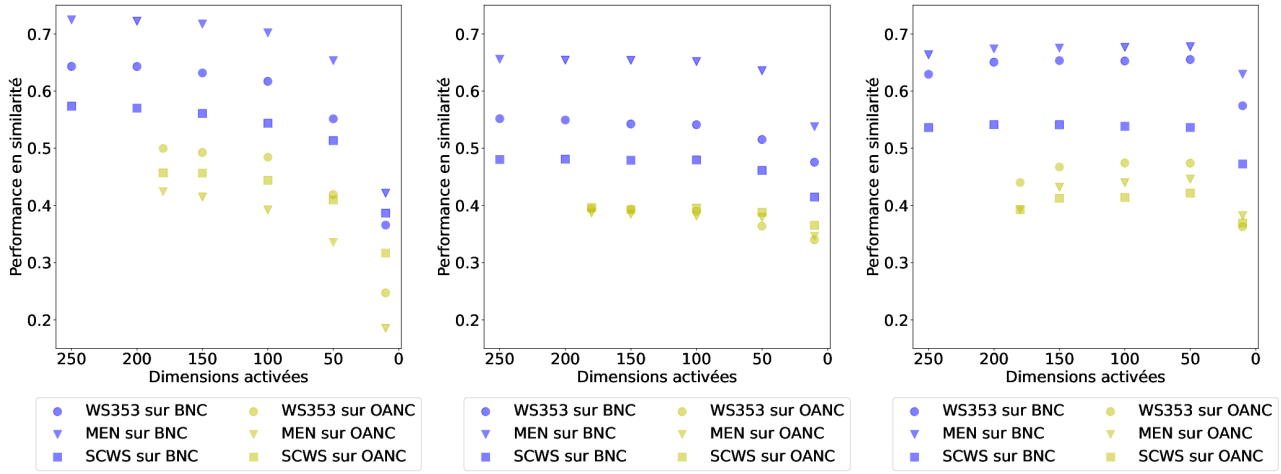


FIGURE 1 – La similarité(en ordonnée, corrélation de Pearson) en fonction du nombre de dimensions conservées par vecteur (en abscisse) pour `Word2Vec` à gauche, `SPINE` au milieu et `SINr` à droite. En jaune, les performances sur le corpus OANC, en bleu celles sur le corpus BNC.

ment menée à ce palier. Bien que les performances baissent significativement pour trois modèles, et particulièrement pour `Word2Vec`, une partie importante de l'information sémantique semble conservée dans ces dix dimensions dans la mesure où elles permettent de résoudre, au moins partiellement, la tâche de similarité. Quoique cette restriction de dix dimensions activées sur les vecteurs ne semble pas intéressante pour des utilisations en aval des vecteurs (considérant la chute de performance), elle permet de construire des vecteurs interprétables. Par ailleurs, cette très faible quantité de composantes rend plus plausible la compatibilité de ces représentations aux modèles théoriques utilisant des traits sémantiques, ouvrant ainsi d'éventuelles opportunités empiriques.

5 Conclusion

Dans cet article, nous proposons de définir l'interprétabilité au niveau des vecteurs, et plus seulement au niveau des dimensions de l'espace de plongement. Nous proposons ainsi un protocole d'évaluation basé sur deux critères : la stabilité des représentations produites, et la contrainte de parcimonie que nous redéfinissons en accord avec la plausibilité psycholinguistique. Il nous paraît en effet souhaitable, non seulement de pouvoir trouver une cohérence sémantique interne à une dimension de description dans une représentation du lexique, mais de pouvoir décrire chaque mot avec un faible nombre de ces dimensions. Nous faisons l'hypothèse que les vecteurs correspondant à ces contraintes sont interprétables par un locuteur, puisqu'il devient possible de manipuler ce faible nombre de dimensions en mémoire de travail.

Il apparaît que les modèles de plongement lexicaux interprétables conservent des résultats intéressants sur la tâche de similarité même en forçant des parcimonies élevées, la méthode `SINr` bénéficiant même du processus d'éparsification appliqué. Ce résultat permet de reconsidérer le compromis entre interprétabilité et performance pour les représentations lexicales. La construction de représentations lexicales aux vecteurs interprétables ouvre par ailleurs la perspective de mettre en relation les modèles théoriques décrivant le lexique depuis des traits sémantiques avec des plongements lexicaux.

Remerciements

Ces travaux ont été financés dans le cadre du projet ANR-21-CE23-0010 DIGING.

Références

BLONDEL V. D., GUILLAUME J. L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, **2008**(10), P10008.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DIGAN W., NÉVÉOL A., NEURAZ A., WACK M., BAUDOIN D., BURGUN A. & RANCE B. (2020). Can reproducibility be improved in clinical natural language processing ? A study of 7 clinical NLP suites. *Journal of the American Medical Informatics Association*, **28**(3), 504–515. DOI : 10.1093/jamia/ocaa261.

FARUQUI M., TSVETKOV Y., YOGATAMA D., DYER C. & SMITH N. (2015). Sparse overcomplete word vector representations. *arXiv preprint arXiv :1506.02004*.

GARRARD P., LAMBON RALPH M., HODGES J. & PATTERSON K. (2001). Prototypicality, distinctiveness, and intercorrelation : Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, **18**(2), 125–174.

HELLRICH J. & HAHN U. (2016a). An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In *SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, p. 111–117.

HELLRICH J. & HAHN U. (2016b). Bad Company Neighborhoods in neural embedding spaces considered harmful. In *Conference on Computational Linguistics*, p. 2785–2796.

JACKENDOFF R. (1983). *Semantic and Cognition*. MIT Press.

LEE D. D. & SEUNG H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature*, **401**, 788–791.

LEVY O. & GOLDBERG Y. (2014). Neural word embedding as implicit matrix factorization. In Z. GHAHRAMANI, M. WELLING, C. CORTES, N. LAWRENCE & K. WEINBERGER, Éds., *Advances in Neural Information Processing Systems*, volume 27 : Curran Associates, Inc.

MCRAE K., CREE G. S., SEIDENBERG M. S. & MCNORGAN C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**(4), 547.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

MILLER G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review*, **63**(2), 81–97.

MURPHY B., TALUKDAR P. & MITCHELL T. (2012). Learning effective and interpretable semantic models using non-negative sparse embedding. In *Conference on Computational Linguistics*, p. 1933–1950.

PALMER S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, **9**(4), 441–474. DOI : https ://doi.org/10.1016/0010-0285(77)90016-0.

PENNINGTON J., SOCHER R. & MANNING C. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing*, p. 1532–1543.

PETERSON L. & PETERSON M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**(3), 193. DOI : 10.1037/h0049234.

PIERREJEAN B. (2020). *Qualitative Evaluation of Word Embeddings : Investigating the Instability in Neural-Based Models*. Thèse de doctorat, Université Toulouse 2 - Jean Jaurès.

POTTIER B. (1963). *Recherches sur l'analyse sémantique en linguistique et en traduction mécanique*. Publications linguistiques de la Faculté des lettres et sciences humaines de Nancy.

PROUTEAU T., CONNES V., DUGUÉ N., PEREZ A., LAMIREL J.-C., CAMELIN N. & MEIGNIER S. (2021). SINr : Fast Computing of Sparse Interpretable Node Representations is not a Sin ! In *Intelligent Data Analysis*, volume 12695, p. 325–337.

PROUTEAU T., DUGUÉ N., CAMELIN N. & MEIGNIER S. (2022). Are embedding spaces interpretable ? results of an intrusion detection evaluation on a large french corpus. In *Language Ressources and Evaluation Conference*.

RASTIER F. (2009). Principes et conditions de la sémantique componentielle. In *Sémantique interprétative*, Formes sémiotiques, p. 17–37.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2021). A Primer in BERTology : What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866.

RUDIN C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, **1**(5), 206–215.

SENEL L. K., UTLU I., YUCESOY V., KO.C A. & CUKUR T. (2018). Semantic structure and interpretability of word embeddings. *Transactions on Audio, Speech, and Language Processing*, **26**(10), 1769–1779.

SHIN J., MADOTTO A. & FUNG P. (2018). Interpreting word embeddings with eigenvector analysis. *Advances in Neural Information Processing Systems*, **32**.

SUBRAMANIAN A., PRUTHI D., JHAMTANI H., BERG-KIRKPATRICK T. & HOVY E. (2018). Spine : Sparse interpretable neural embeddings. In *AAAI conference on artificial intelligence*, volume 32.

ZAFAR M. R. & KHAN N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, **3**(3), 525–541.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Effet de l'anthropomorphisme des machines sur le français adressé aux robots: Étude du débit de parole et de la fluence

Natalia Kalashnikova[1] Mathilde Hutin[1, 2] Ioana Vasilescu[1]
Laurence Devillers[1, 3]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Institut Langage & Communication, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgique
(3) Sorbonne Université, Paris, France
{natalia.kalashnikova,hutin,ioana,devil}@lisn.upsaclay.fr

RÉSUMÉ

"Robot-directed speech" désigne la parole adressée à un appareil robotique, des petites enceintes domestiques aux robots humanoïdes grandeur-nature. Les études passées ont analysé les propriétés phonétiques et linguistiques de ce type de parole ou encore l'effet de l'anthropomorphisme des appareils sur la sociabilité des interactions, mais l'effet de l'anthropomorphisme sur les réalisations linguistiques n'a encore jamais été exploré. Notre étude propose de combler ce manque avec l'analyse d'un paramètre phonétique (débit de parole) et d'un paramètre linguistique (fréquence des pauses remplies) sur la parole adressée à l'enceinte *vs* au robot humanoïde *vs* à l'humain. Les données de 71 francophones natifs indiquent que les énoncés adressés aux humains sont plus longs, plus rapides et plus dysfluents que ceux adressés à l'enceinte et au robot. La parole adressée à l'enceinte et au robot est significativement différente de la parole adressée à l'humain, mais pas l'une de l'autre, indiquant l'existence d'un type particulier de la parole adressée aux machines.

ABSTRACT

The Effect of Human-Likeliness in French Robot-Directed Speech : A study of Speech Rate and Fluency

Robot-directed speech refers to speech to a robotic device, ranging from small home speakers to full-size humanoid robots. Studies have investigated the phonetic and linguistic properties of this type of speech or the effect of anthropomorphism of the devices on the social aspect of interaction. However, none have investigated the effect of the device's human-likeliness on linguistic realizations. This preliminary study proposes to fill this gap by investigating one phonetic parameter (speech rate) and one linguistic parameter (use of filled pauses) in speech directed at a home speaker *vs* a humanoid robot *vs* a human. The data from 71 native speakers of French indicate that human-directed speech shows longer utterances at a faster speech rate and more filled pauses than speech directed at a home speaker and a robot. Speaker- and robot-directed speech is significantly different from human-directed speech, but not from each other, indicating a unique device-directed type of speech.

MOTS-CLÉS : parole adressée aux robots, interaction humain-machine, débit de parole, pauses remplies, anthropomorphisme.

KEYWORDS: robot-directed speech, human-computer interaction, speech rate, filled pause, human-likeliness.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 92–100.

1 Introduction

Robot-directed speech (RDS), littéralement "parole adressée au robot", désigne les productions langagières d'un humain à un outil robotique, qu'il s'agisse de petites enceintes intelligentes ou de robots humanoïdes grandeur nature. Ce type de parole peut être comparé au *computer-directed speech* ("parole adressée à l'ordinateur") en ce qu'il appartient lui aussi à la catégorie du *device-directed speech* ("parole adressée à la machine"), dans la catégorie des registres langagiers spéciaux, tels que la parole adressée aux enfants, aux locuteurs non-natifs, ou même aux animaux. L'exploration de tels styles de parole permet non seulement une meilleure compréhension des interactions humain-machine, mais aussi de l'ajustement dans le dialogue et de l'accommodation entre les interlocuteurs, ce qui à son tour peut permettre de développer des outils concrets tels que la détection de destinataire (*addressee-detection*).

Certaines études ont spécifiquement analysé les propriétés acoustiques de la parole adressée à l'ordinateur ou au robot à celle adressée aux adultes. Ces travaux de recherche ont montré que, lorsqu'ils s'adressent aux ordinateurs, les humains ont tendance à produire plus d'énoncés (Amalberti *et al.*, 1993) et à hyperarticuler les voyelles (en termes de formants et de durée) sans avoir une fréquence fondamentale (F0) plus élevée (Burnham *et al.*, 2010). En revanche, dans la parole adressée aux robots, les voyelles sont aussi hyperarticulées (Kriz *et al.*, 2010) mais le débit de parole montre peu de différence (Raveh *et al.*, 2019) et la F0 et l'intensité sont plus élevées (Kriz *et al.*, 2010; Raveh *et al.*, 2019). En ce qui concerne les particularités linguistiques de la communication avec les ordinateurs, les humains ont tendance à contrôler et simplifier leur usage du langage (moins de dysfluences, quantité limitée d'informations par proposition, etc.) : en particulier, les locuteurs utilisent moins de pauses remplies (comme *euh* ou *um*) (Amalberti *et al.*, 1993). Toutes ces études indiquent également que les différences entre la parole adressée à un appareil et la parole adressée à un humain diminuent au cours de la conversation (Amalberti *et al.*, 1993) et que les adultes montrent plus de variation intra-locuteur dans la parole adressée à un appareil que dans la parole adressée à un enfant (Kriz *et al.*, 2009; Fischer *et al.*, 2011).

Ces différences entre la parole adressée à l'ordinateur et la parole adressée au robot laissent supposer que les appareils ne sont pas tous considérés de la même façon par les humains. De plus, Gong (2008) a montré que, sur une échelle de quatre niveaux d'anthropomorphisme, plus l'appareil ressemble à un humain, plus il reçoit de réponses sociales de la part des utilisateurs lors de l'interaction. Ces résultats ont été confirmés d'un point de vue neurologique, puisque plus l'appareil est anthropomorphique, plus l'activité corticale est forte dans les région du cerveau associées au raisonnement à propos de l'intention d'autrui (la théorie de l'esprit) (Krach *et al.*, 2008).

Dans le présent article, nous présentons les résultats préliminaires d'une étude plus large en opposant les participants dialoguant avec une enceinte Google Home *vs* un robot Pepper *vs* un autre humain. Plus précisément, nous analysons les facteurs suivants :

— les **facteurs phonétiques** : si les participants produisent (i) des énoncés (en l'occurrence des tours de parole) plus longs ; (ii) avec un débit de parole moins rapide pour la condition enceinte *vs* humain ;
— les **facteurs linguistiques** : si les participants produisent (iii) moins de pauses remplies pour la condition enceinte *vs* humain ;
— la **variation** : (iv) si ces mesures diffèrent plus intra-locuteurs dans la parole adressée à l'enceinte que dans la parole adressée à l'humain et (v) évoluent au cours de la conversation ;
— l'**effet de l'anthropomorphisme** : (vi) si la parole adressée au robot Pepper ressemble plus à

la parole adressée à l'humain ou à la parole adressée à l'enceinte (ou se positionne quelque part entre les deux).

2 Méthode

2.1 Protocole expérimental

Procédure. D'abord, les membres de notre équipe expliquent le déroulé de l'expérience aux participants, qui signent la notice de consentement et remplissent un formulaire de questions de base. Dans ce questionnaire, ils notent leur degré de volonté à adopter une série de 8 habitudes écologiques. Ensuite, ils suivent un ou une bénévole dans l'une des 3 salles (qui correspondent chacune à un agent conversationnel), où deux membres de notre équipe contrôlent la configuration de l'expérience. À la fin de l'enregistrement, les expérimentateurs remercient les participants et les invitent à retourner auprès des organisateurs où ils peuvent poser leurs questions sur l'expérience en prenant une collation. Le déroulement de l'expérience a été validé par un comité d'éthique et a respecté les mesures sanitaires contre le covid-19.

Enregistrements. La procédure d'enregistrement a été inspirée de l'étude de Mehenni *et al.* (2020). Lors de l'enregistrement, l'agent conversationnel (enceinte, robot Pepper ou humain) pose des questions sur les habitudes écologiques des participants. Les protocoles de l'enceinte et du robot Pepper sont réalisés dans un dispositif en Magicien d'Oz, c'est-à-dire que les réponses pré-enregistrées sont envoyées à la machine par un des expérimentateurs sans que le participant s'en rende compte. La voix de synthèse fournie par les paramètres par défaut du robot Pepper est une voix enfantine, qui a été enregistrée séparément afin de pouvoir l'utiliser aussi pour l'enceinte. La condition de la parole adressée à l'humain est réalisée par un des membres de notre équipe, qui lit à haute voix le même script qui a été enregistré pour le robot et l'enceinte. L'échange oral consiste en 4 étapes :

- S0 : L'agent instaure la conversation avec le participant en posant quelques questions anodines.
- S1 : L'agent présente les situations hypothétiques dans lesquelles les participants doivent choisir entre l'option par défaut et l'option respectueuse de l'environnement qui demande plus d'investissement (d'argent ou de temps).
- S2 : L'agent fournit une information présentant les conséquences sur l'environnement de chaque habitude et pose cette fois les questions de base du formulaire écrit.
- S3 : L'agent présente d'autres situations hypothétiques et les participants doivent à nouveau choisir entre l'option par défaut et l'option respectueuse de l'environnement.

Les données audio ont été enregistrées avec le microphone unidirectionnel (AKG45) sur Audacity à 44.1 kHz.

Participants. En avril et juin 2022, notre équipe a recruté les visiteurs, le personnel et les étudiants de Collège des Bernardins à Paris, France. Les 71 locuteurs natifs du français (46 femmes, 25 hommes) se répartissent de façon équilibrée dans les groupes d'âge (de 18 à 65+ ans). Parmi tous les participants, 21 (16 femmes, 5 hommes) ont participé à la condition avec l'agent humain, 28 (18 femmes, 10 hommes) à la condition avec l'agent robot, et 22 (12 femmes, 10 hommes) à la condition avec l'agent enceinte. Au total, 16 heures de parole ont été enregistrées.

2.2 Méthodologie

Les fichiers audio ont été segmentés manuellement et transcrits en français par deux annotateurs. Les pauses remplies ont été annotées dans la transcription. Les fichiers finaux de transcription contiennent l'horodatage des tours de parole. Tous les calculs sont faits avec Python 3 (Van Rossum & Drake, 2009).

Nous utilisons ces fichiers d'annotation pour calculer la durée totale des énoncés en secondes pour chaque tour de parole de chaque participant. Pour le calcul du débit de parole, nous utilisons l'espace entre les mots de la transcription pour segmenter le discours en unités de type (pseudo-)mots. Nous calculons ensuite la somme des unités pour chaque étape de l'expérience et divisons cette somme par la durée de l'étape. Ainsi, nous obtenons le débit de parole exprimé en nombre d'unités par seconde.

Pour les pauses remplies, nous calculons le nombre total des pauses remplies et divisons ce nombre par la durée de tours de parole pour obtenir le nombre de pauses remplies par seconde.

Pour l'analyse de la variation intra-individuelle, nous calculons l'écart-type de chaque paramètre pour chaque participant et ensuite la moyenne des écarts-types pour chaque condition.

La significativité de nos résultats est testée avec un *t-test* pour les deux échantillons indépendants appliqués avec SciPy (Virtanen *et al.*, 2020). Nous avons ensuite appliqué la correction de Bonferroni (seuil alpha = 0.5) avec Statsmodels (Seabold & Perktold, 2010).

3 Résultats

3.1 Longueur de l'énoncé et débit de parole

De façon générale, les participants ont tendance à produire des énoncés plus longs en parlant avec un humain (moyenne = 372 sec.) qu'en parlant avec l'enceinte (moyenne = 191 sec., Δ=181 sec., *t*=2,19, p=0,09) et surtout au robot humanoïde (moyenne = 147 sec., Δ=225 sec., *t*=2,87, p=0,01). La durée des énoncés de la parole adressée au robot et à l'enceinte n'est pas significativement différente (Δ=44 sec., *t*=-1,08, p=0,85). En regardant chaque étape séparément, comme dans la Figure 1, il apparaît

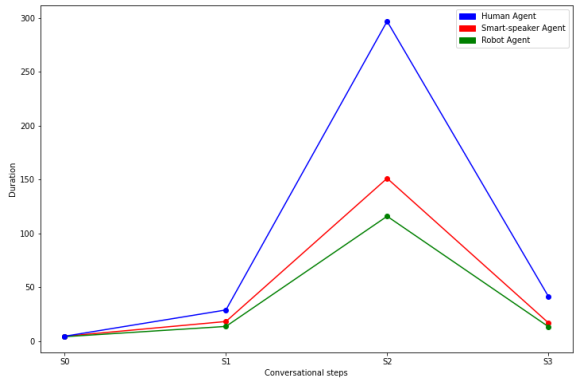


FIGURE 1 – Durée moyenne des énoncés de la parole adressée à l'humain (bleu), à l'enceinte (rouge) et au robot humanoïde (vert) à chaque étape de la conversation.

que les seules différences significatives sont observées entre la parole adressée à l'humain et celle adressée au robot dans les étapes "S2" (Δ=181,15 sec., p=0,01), et "S3" (Δ=28,14 sec., p=0,02).

Contrairement à la littérature comparant la parole adressée à l'appareil et l'enfant (Kriz *et al.*, 2009; Fischer *et al.*, 2011), les participants montrent plus de variation intra-locuteur dans la condition de la parole adressée à l'humain, avec un écart-type à travers des participants de 137,54 secondes, que dans la parole adressée à l'enceinte (écart-type moyen = 69,57, Δ=67,97 sec., t=2,47, p=0,06). Ce paramètre est significativement différent entre la parole adressée à l'humain et au robot (écart-type moyen = 53,18 sec., Δ=84,36 sec., t=3,24, p=0,004). Néanmoins, les différences entre la parole adressée à l'enceinte et au robot ne sont pas significativement différents l'un de l'autre (Δ=16,39, t=-1,3, p=0,2).

En ce qui concerne le débit de parole (Figure 2), la parole adressée à l'humain est plus lente au début mais s'accélère autour de l'étape S1, au moment où la conversation commence réellement. De manière générale, les participants s'adressent à l'humain avec un débit de 2,96 unités/sec, au robot avec un débit de 2,63 unités/sec, et à l'enceinte avec un débit de 2,5 unités/sec. Comme attendu, la différence entre la parole adressée à l'humain et la parole adressée à l'enceinte est significative (Δ=0,46 unités/sec, t=3,43, p=0,004). Contrairement à Raveh *et al.* (2019), la différence entre la parole adressée à l'humain et la parole adressée au robot est aussi significative (Δ=0,32 unités/sec, t=2,8, p=0,009), mais pas entre la parole adressée au robot et la parole adressée à l'enceinte (Δ=0,14 unités/sec, t=1,32, p=0,58). L'analyse de chaque étape montre que les deux différences significatives

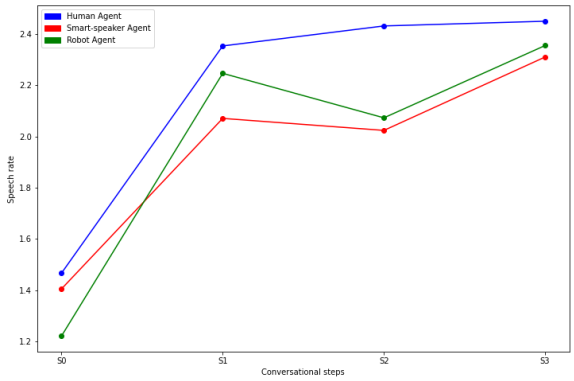


FIGURE 2 – Débit de parole moyen des énoncés de la parole adressée à l'humain (bleu), à l'enceinte (rouge) et au robot humanoïde (vert) à chaque étape de la conversation.

du débit de parole sont observées à l'étape "S2" dans les conditions entre l'humain et l'enceinte (Δ=0,4 unités/sec., p=0,002), et entre l'humain et le robot sur la même étape (Δ=0,36 token/sec., p=0,003). Aucune différence significative n'a été observée entre les étapes de la parole adressée au robot et la parole adressée à l'enceinte.

Contrairement à ce qui était attendu, on observe moins de variation intra-locuteur dans la parole adressée à l'enceinte, avec un écart-type moyen de 0,6 unités/sec. contre un écart-type moyen de 0,68 unités/sec. dans la parole adressée au robot (Δ=-0,08 unités/sec., t=0,92, p>1.) et un écart-type moyen de 0,62 unités/sec dans la parole adressée à l'humain (Δ=-0,02 unités/sec., t=0,19, p>1). La différence entre les paroles adressées au robot et à l'humain n'est pas significative (Δ=0,06 unités/sec., t=-0.7, p>1.).

Ces résultats indiquent que l'anthropomorphisme des appareils impacte de façon modérée la longueur

des énoncés et le débit de parole. Les participants ont tendance à parler plus longtemps à l'humain qu'au robot, mais pas à l'enceinte, tandis qu'ils parlent plus rapidement à l'humain qu'à l'enceinte, et surtout plus rapidement qu'au robot. Cependant, lorsqu'on compare les résultats pour le robot et l'enceinte, aucune différence de longueur des énoncés ni de débit de la parole n'est observée.

3.2 Pauses remplies

Les pauses remplies sont des instances de dysfluences, des accidents dans la production de la parole qui sont extrêmement fréquents dans les interaction humain-humain (Shriberg, 1994). De façon générale, les participants produisent significativement moins de pauses remplies quand ils parlent à l'enceinte que quand ils parlent à l'humain (Δ=-0,4, t=2,92, p=0,01) et au robot (Δ=-0,29, t=2,62, p=0,03). Cependant, la différence de fréquence des pauses remplies entre la parole adressée à l'humain et la parole adressée au robot est non-significative (Δ=0,11, t=0,8, p>1.).

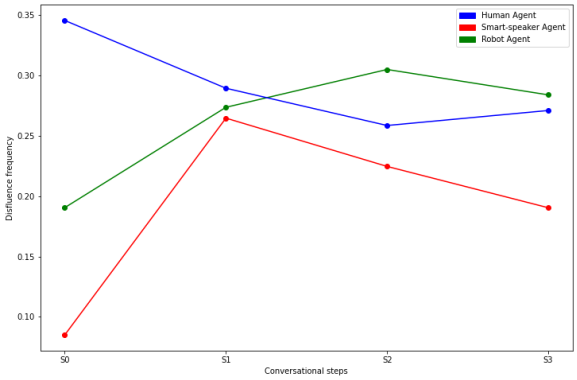


FIGURE 3 – Score moyen des disfluences des énoncés de la parole adressée à l'humain (bleu), à l'enceinte (rouge) et au robot humanoïde (vert) à chaque étape de la conversation.

Cette situation évolue avec le temps (Figure 3). Au début de la conversation, surtout à l'étape "S0" (échange de banalités), les participants produisent plus des pauses remplies quand ils parlent à l'humain (0,35 pauses remplies par seconde), que quand ils parlent au robot (0,19 pauses/sec., Δ=0,16 pauses/sec., p=0,4), ou surtout à l'enceinte (0,09 pauses/sec., Δ=0,26 pauses/sec., p=0,04). A la fin de l'interaction, par exemple à l'étape "S3", les participants produisent presque autant de pauses remplies quand ils parlent à l'humain (0,27 pauses/sec.), que quand ils parlent au robot (0,28 pauses/sec., Δ=0,01 pauses/sec., p>1.), mais plus quand ils parlent à l'enceinte (0,19 pauses/sec., Δ=0,08, p=0,4).

La variation intra-locuteur est similaire dans les trois conditions, avec un écart-type de 0,17, 0,17, (Δ=0, t=0,004, p=1) et 0,12 (Δ=0,05, t=1,1, p=0,28) pauses remplies par seconde dans la parole adressée respectivement à l'humain, au robot et à l'enceinte.

Ces résultats en accord avec les recherches passées (Amalberti *et al.*, 1993) indiquent que les participants produisent plus de dysfluences dans la parole adressée à l'humain que dans la parole adressée à un appareil, mais cet écart diminue avec le temps. Cependant, la parole adressée au robot n'est pas significativement différente de la parole adressée ni à l'humain ni à l'enceinte.

4 Conclusion et discussion

Dans cet article, nous avons analysé 16 heures de parole de 71 locuteurs natifs du français pour étudier le débit de parole et l'utilisation des pauses remplies dans la parole adressée à l'humain, à l'enceinte et au robot. Partiellement conforme aux recherches passées sur la parole adressée à l'ordinateur (Amalberti *et al.*, 1993), nous montrons que les participants parlent plus vite et produisent plus de pauses remplies en parlant avec l'humain qu'avec l'enceinte. En ce qui concerne le robot humanoïde, en se basant sur les études préconisant l'effet de l'anthropomorphisme sur les interactions parlées (Gong, 2008; Krach *et al.*, 2008), nous avons fait l'hypothèse que la parole adressée au robot partagerait plus de caractéristiques avec la parole adressée à l'humain qu'avec la parole adressée à l'enceinte. Cependant, les participants ont produit des énoncés plus longs de la façon plus rapide avec plus de pauses remplies dans la parole adressée à l'humain que dans la parole adressée au robot. De plus, la parole adressée au robot et celle adressée à l'enceinte ne sont pas différentes en longueur de l'énoncé, ni en débit de parole et en fréquence de dysfluences. Nous nous attendions également à plus de variation intra-locuteur dans la parole adressée à l'appareil qu'à l'humain : nous trouvons la tendance inverse pour la longueur des énoncés, et aucune différence pour la fréquence des pauses remplies et le débit de la parole. De façon générale, la parole adressée à l'humain est opposée à la parole adressée à l'enceinte et la parole adressée au robot.

La similarité entre les paroles adressées au robot et à l'enceinte peut être due au fait que les deux appareils communiquent avec une voix d'enfant, tandis que l'agent conversationnel humain était adulte. Il est possible que les participants se soient davantage alignés avec la parole adressée à l'enfant qu'avec la parole adressée à l'appareil. Cependant, les conditions du robot et de l'enceinte sont comparables mais ne montrent aucune différence significative, indiquant que le comportement des participants peut être considéré comme un type unique de "parole adressée à l'appareil". Enfin, le robot et l'enceinte ont utilisé la même voix alors que l'agent humain a été incarné à tour de rôle par 3 femmes et 1 homme. Il est possible que les données de la condition avec l'agent humain aient été impactées par la différence entre la parole adressée à la femme et celle adressée à l'homme.

Les futures analyses devront aussi prendre en compte l'âge des participants et leur utilisation quotidienne des appareils robotiques pour établir l'effet de la familiarité et de l'habitude. Dans les futures étapes de cette recherche, nous envisageons d'étudier les propriétés linguistiques du discours, tels que les marqueurs discursifs, la complexité de la syntaxe, etc. Nous comptons également analyser la similarité des caractéristiques phonétiques et linguistiques entre la parole des agents et la parole des sujets afin d'étudier les mécanismes de l'alignement linguistique pendant la conversation.

Remerciements

Les auteurs remercient les relecteurs pour leurs commentaires et leurs suggestions. Cet article est écrit dans le cadre de la thèse financée par Chaire AI HUMAAINE (ANR-19-CHIA-0019) dirigé par Laurence Devillers.

Références

AMALBERTI R., CARBONELL N. & FALZON P. (1993). User representations of computer systems

in human-computer speech interaction. *International Journal of Man-Machine Studies*, **38**(4), 547–566. DOI : https ://doi.org/10.1006/imms.1993.1026.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BURNHAM D., JOEFFRY S. & RICE L. (2010). Computer-and human-directed speech before and after correction. In *Proceedings Of The 13Th Australasian International Conference On Speech Science And Technology*, volume 6, p. 13–17.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FISCHER K., FOTH K., ROHLFING K. J. & WREDE B. (2011). Mindful tutors : Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, **12**, 134–161. DOI : https ://doi.org/10.1075/is.12.1.06fis.

GONG L. (2008). How social is social responses to computers ? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, **24**(4), 1494–1509. Including the Special Issue : Integration of Human Factors in Networked Computing, DOI : https ://doi.org/10.1016/j.chb.2007.05.007.

KRACH S., HEGEL F., WREDE B., SAGERER G., BINKOFSKI F. & KIRCHER T. (2008). Can Machines Think ? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLoS ONE*, **3**(7), 1494–1509. Including the Special Issue : Integration of Human Factors in Networked Computing, DOI : https ://doi.org/10.1371/journal.pone.0002597.

KRIZ S., ANDERSON G., BUGAJSKA M. & TRAFTON J. G. (2009). Robot-directed speech as a means of exploring conceptualizations of robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, p. 271–272, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1514095.1514171.

KRIZ S., ANDERSON G. & TRAFTON J. G. (2010). Robot-directed speech : Using language to assess first-time users' conceptualizations of a robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, p. 267–274. DOI : 10.1109/HRI.2010.5453187.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

MEHENNI H. A., KOBYLYANSKAYA S., VASILESCU I. & DEVILLERS L. (2020). Nudges with conversational agents and social robots : a first experiment with children at a primary school. In *11th International Workshop on Spoken Dialog System Technology*, Madrid, Spain. HAL : hal-03083526.

RAVEH E., STEINER I., SIEGERT I., GESSINGER I. & MÖBIUS B. (2019). Comparing phonetic changes in computer-directed and human-directed speech. In P. BIRKHOLZ & S. STONE, Éds., *Studientexte zur Sprachkommunikation : Elektronische Sprachsignalverarbeitung 2019*, p. 42–49 : TUDpress, Dresden.

SEABOLD S. & PERKTOLD J. (2010). statsmodels : Econometric and statistical modeling with python. In *9th Python in Science Conference*.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHRIBERG E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. Thèse de doctorat, University of California, Berkeley.

VAN ROSSUM G. & DRAKE F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA : CreateSpace.

VIRTANEN P., GOMMERS R., OLIPHANT T. E., HABERLAND M., REDDY T., COURNAPEAU D., BUROVSKI E., PETERSON P., WECKESSER W., BRIGHT J., VAN DER WALT S. J., BRETT M., WILSON J., MILLMAN K. J., MAYOROV N., NELSON A. R. J., JONES E., KERN R., LARSON E., CAREY C. J., POLAT İ., FENG Y., MOORE E. W., VANDERPLAS J., LAXALDE D., PERKTOLD J., CIMRMAN R., HENRIKSEN I., QUINTERO E. A., HARRIS C. R., ARCHIBALD A. M., RIBEIRO A. H., PEDREGOSA F., VAN MULBREGT P. & SCIPY 1.0 CONTRIBUTORS (2020). SciPy 1.0 : Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**, 261–272. DOI : 10.1038/s41592-019-0686-2.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Détection de la nasalité du locuteur à partir de réseaux de neurones convolutifs et validation par des données aérodynamiques

Lila Kim[1] Cédric Gendrot[1] Amélie Elmerich[1] Angélique Amelot[1]
Shinji Maeda[1]
(1) Laboratoire de Phonétique et Phonologie, 4 rue des irlandais, 75005 Paris, France
lila.kim@sorbonne-nouvelle.fr, cedric.gendrot@sorbonne-nouvelle.fr,
amelie.elmerich@gmail.com, angelique.amelot@gmail.com, smaeda75@gmail.com

RÉSUMÉ

Ce travail se positionne dans le domaine de la recherche d'informations sur le locuteur, reconnue comme une des tâches inhérentes au traitement automatique de la parole. A partir d'un nouveau masque pneumotachographe acoustiquement transparent, nous avons enregistré simultanément des données aérodynamiques (débit d'air oral et nasal) et acoustiques pour 6 locuteurs masculins français, impliquant des consonnes et voyelles orales et nasales sur des logatomes. Un CNN entraîné sur d'autres corpus acoustiques en français a été testé sur les données recueillies à partir du masque pour la distinction de nasalité phonémique, avec une classification correcte de 88% en moyenne. Nous avons comparé ces résultats CNN avec les débit d'air nasal et oral captés par le masque afin de quantifier la nasalité présente par locuteur. Les résultats montrent une corrélation significative entre les erreurs produites par le CNN et des distinctions moins nettes de débit d'air entre nasales et orales.

ABSTRACT

Detection of speaker nasality from convolutional neural networks and validation with aerodynamic data

This work is positioned in the field of speaker information retrieval, which is recognized as one of the inherent tasks of automatic speech processing. Using a new acoustically transparent pneumotachograph mask, we simultaneously recorded aerodynamic (oral and nasal airflow) and acoustic data for 6 French male speakers, involving oral and nasal consonants and vowels on logatomes. A CNN trained on other French acoustic corpora was tested on the data collected from the mask for nasality distinction, with a correct classification of 88% on average. We compared these CNN results with the nasal and oral airflow captured by the mask to quantify the nasality present per speaker. The results show a significant correlation between the errors produced by the CNN and less clear distinctions in airflow between nasal and oral.

MOTS-CLÉS : CNN, RI, nasalité, caractérisation des locuteurs, aérodynamique.

KEYWORDS: CNN, nasality, speaker characterisation, nasality, aerodynamic.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 101–108.

1 INTRODUCTION

La nasalité est un trait distinctif dans environ un tiers des langues du monde (Basbøll, 1985). Les connaissances de base impliquent que le palais mou doit être suffisamment abaissé pour que l'orifice vélopharyngé soit ouvert et permette à l'air de passer par le nez. L'abaissement du voile du palais et le passage de l'air par le nez ont une incidence sur la composante acoustique du signal de parole, ce qui gêne généralement l'analyse acoustique pour les phonéticiens(Styler, 2017).

Il existe des variations temporelles et spatiales dans la réalisation de la caractéristique [nasale]. Elle varie en fonction du sexe et de l'anatomie du locuteur (Clarke, 1975; Amelot, 2004), de la stratégie du locuteur (Croft *et al.*, 1981; Skolnick *et al.*, 1973; Vaissière, 1988), de la langue (Clumeck, 1976), du style d'élocution (Basset *et al.*, 2001), du débit de parole (Bell-Berti & Krakow, 1991), du type de son de parole, du contexte phonétique et prosodique (Krakow, 1993), etc.

Plus précisément, l'ouverture de l'orifice vélopharyngé diffère d'un locuteur à l'autre et la morphologie des fosses nasales est très variable d'un individu à l'autre (Clarke, 1975; Amelot, 2004). Les voyelles et consonnes nasales sont importantes pour l'identification du locuteur car elles contiennent plus d'informations acoustiques relatives aux locuteurs que les autres sons (Ajili *et al.*, 2016; Kahn *et al.*, 2011).

Les réseaux neuronaux profonds ont récemment connu un développement important dans le domaine de la parole. Des études ont été menées dans le domaine clinique avec des réseaux neuronaux artificiels pour diagnostiquer des pathologies du langage, notamment l'hyper- ou l'hypo-nasalisation (Wang *et al.*, 2019; Mohammed *et al.*, 2020; Abderrazek *et al.*, 2022b). En effet, il a été démontré que les réseaux de neurones artificiels ont la capacité de se spécialiser sur des caractéristiques phonétiques telles que le lieu d'articulation ou le mode articulatoire (Abderrazek *et al.*, 2022b,a; Pellegrini & Mouysset, 2016).

L'objectif de la présente étude est d'évaluer la détection de la nasalité à partir de données acoustiques avec un CNN en la comparant avec les données aérodynamiques collectées lors de l'enregistrement acoustique. Nous prenons l'étiquette phonémique de la voyelle "nasale" ou "orale" comme référence et vérifions si la classification CNN est correcte à partir des données acoustiques. Dans un deuxième temps, le débit d'air nasal fourni validera la classification CNN ou nous aidera à comprendre les erreurs de classification. Enfin, nous étudierons si le niveau de nasalité de chaque locuteur peut être approché avec le CNN.

Ce travail se situe dans le domaine de la recherche d'information en TAL. En effet, la nasalité est une caractéristique inhérente du locuteur dûe à la morphologie peu malléable des cavités nasales ainsi qu'aux habitudes de production idiolectales. Dans le cadre de la vérification du locuteur, les informations propres à la voix du locuteur sont cruciales pour l'explicabilité du résultat.

2 MATÉRIAUX ET MÉTHODES

2.1 Corpus et acquisition de données

Pour cette étude, 6 locuteurs natifs masculins français (âge moyen : 36 ans) ont été enregistrés dans une pièce insonorisée. Les échantillons de parole sont constitués de séquences VCV, où

C=[p,b,t,d,v,s,z,m,n] et V=[i,a,y,u,e,ɑ̃,ɛ̃,ɔ̃]. Les séquences ont été insérées dans la phrase cadre, par exemple : " Non, tu n'as pas dit apa quatre fois, tu as dit aba et ada quatre fois ". Finalement, nous avons un total de 270 séquences avec C= 270 et V= 540. Les données aérodynamiques et acoustiques ont été enregistrées simultanément à l'aide d'un masque pneumotachographe.

Les avantages de ce masque sont les suivants : i) le débit d'air oral et nasal peut être enregistré séparément, ii) il est possible d'adapter la taille et la position de la plaque pour séparer le débit d'air nasal (NAF) du débit d'air oral (OAF) pour chaque locuteur, iii) il n'y a pas de distorsion acoustique.

Il peut y avoir de légères différences dans les mesures de débit en fonction du masque, de la position du capteur et de la taille et de la position de la plaque séparant le débit d'air nasal et le débit d'air oral. Par conséquent, un étalonnage doit être effectué séparément pour chaque masque (masque individuel pour chaque locuteur) : un étalonnage pour le compartiment buccal et un autre pour le compartiment nasal. L'étalonnage des 2 modules de capteurs de pression permet de convertir les valeurs de débit d'air dans l'unité physique (litres/s, voir Figure 1). Le masque offre une faible résistance, nécessaire pour mesurer le débit d'air sans affecter la propagation du son.

Les données acoustiques ont été capturées à l'aide d'un microphone (AKG C520 L). Tous les capteurs aérodynamiques et acoustiques sont reliés à une carte d'acquisition (DT9003). Les données acoustiques et aérodynamiques ont été enregistrées à une fréquence d'échantillonnage de 20kHz. Les données ont été segmentées manuellement dans Praat. Un script Python a été utilisé pour extraire automatiquement la moyenne de l'OAF et du NAF pour chaque voyelle (l/sec).

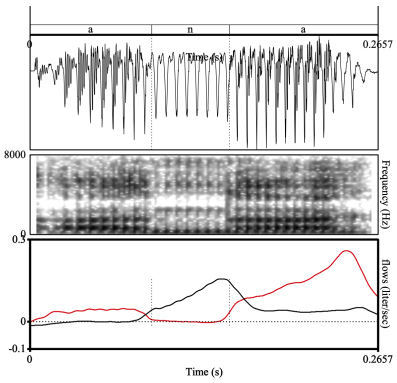


FIGURE 1 – Exemple d'enregistrements acoustiques et de débit d'air de [ana]. De haut en bas, (1) signal audio capturé avec un microphone, (2) spectrogramme, (3) débit d'air nasal (NAF en noir) et débit d'air oral (OAF en rouge)

2.2 Réseau de neurones convolutif

Pour la tâche de classification automatique acoustique nasal-non nasal, nous avons choisi de travailler avec des réseaux neuronaux convolutifs (CNN). Pour des raisons d'espace, nous détaillons uniquement les résultats sur 6 voyelles /a, e, o, ɑ̃, ɛ̃, ɔ̃/, donc 3 qualités de voyelles /a, e, o/ et leurs homologues nasales. Nous sommes conscients qu'il n'y a pas de correspondance articulatoire exacte entre nos 3 voyelles orales et nasales (Zerling, 1984) et nous avons décidé d'inclure les 6 voyelles dans le système de classification (au lieu de comparer par paires) afin de contourner cette asymétrie.

Le choix d'un CNN par rapport à d'autres réseaux neuronaux est lié à notre intention d'utiliser des spectrogrammes de voyelles, l'objectif final étant de localiser, à l'aide d'un algorithme de type

gradCam, l'endroit où se trouve l'information sur la nasalité.

L'ensemble de données d'entraînement est composé des productions de ces voyelles extraites de 3 corpus français avec différents types de discours : NCCFr (Torreira *et al.*, 2010), ESTER (Gravier *et al.*, 2004) et PTSVOX (Chanclu *et al.*, 2020). Dans tous ces corpus, des segmentations automatiques ont été fournies au niveau des phonèmes. Les voyelles ont été extraites aléatoirement à leurs frontières sous la forme d'un spectrogramme, sans aucune sélection du contexte prosodique, lexical ou phonémique. Pour les 2 premiers corpus, le nombre de voyelles de chaque type a été vérifié. 10 887 productions de chaque type ont été extraites de NCCFr et 9 186 d'ESTER. Pour PTSVOX, nous avons pris toutes les voyelles possibles en respectant la fréquence naturelle des phonèmes, ce qui a donné de meilleurs résultats.

	Entraînement & validation			Test
Source	NCCFr	ESTER	PTSVOX	Données enregistrées avec le masque
nasal	32,661	27,558	65,669	198
non nasal	32,661	27,558	135,119	198

TABLE 1 – Nombre de voyelles dans les jeux d'entraînement et de test selon les corpus

Dans la phase d'évaluation du modèle, nous avons sélectionné au hasard des productions de voyelles dans les données acoustiques présentées à la section 2.1. et extrait leurs spectrogrammes. Cet ensemble de test contient 66 productions de chaque type de voyelle (6 locuteurs * 11 occurrences), soit 198 voyelles pour chaque catégorie. Toutes ces images de spectrogrammes avec une bande de fréquence de 0 à 8000 Hz ont été réduites en 48x48 pixels et présentées comme entrée à notre réseau.

Pour la partie extraction des caractéristiques, le modèle est composé de deux blocs de couches de convolution et de pooling. Les couches de convolution ont été réalisées avec un noyau de taille 5x5 et ont donc produit respectivement 32 et 64 filtres. Après chaque couche de convolution, une couche de batch normalisation a été insérée avant d'appliquer une couche d'activation afin de permettre au modèle de se généraliser (Ioffe & Szegedy, 2015) sur différents types de corpus et de données. Les couches de max-pooling ont ensuite été utilisées pour réduire la taille des images avec une taille de pool de 2x2. Avec les caractéristiques extraites, 3 couches denses ont effectué la tâche de classification avec 1024 neurones. La fonction d'activation ReLU a été appliquée après chaque couche de batch normalisation et chaque couche dense. Enfin, une fonction d'activation softmax a été utilisée dans la dernière couche dense pour la classification nasale-orale. Au cours de l'apprentissage du modèle, nous avons tenté de minimiser les erreurs du modèle en appliquant Adam comme technique d'optimisation et categorical crossentropy comme fonction de perte pour mesurer la performance du modèle.

3 Résultats

3.1 Résultats du CNN

Pour une voyelle donnée, le classifieur renvoie une valeur entre 0 et 1, que nous appelons la probabilité de nasalité. Lorsqu'une voyelle est identifiée comme nasale par le modèle, la probabilité de nasalité attendue est supérieure à 0,5 et, inversement, une voyelle classée comme non nasale a une valeur proche de 0 (ou au moins inférieure à 0,5). Notre classifieur a pu identifier avec précision 95% des voyelles non nasales et 82% des voyelles nasales en atteignant une exactitude globale de 88% et

un score F1 de 88% (k = 0,77). Nous avons également testé un autre modèle incluant -en plus des voyelles- les consonnes /m, n, l, b, d, v/, et obtenu des résultats comparables : 89% d'exactitude globale en testant des voyelles et consonnes nasales et orales confondues.

La variabilité inter-locuteurs peut être observée sur la figure 2 (uniquement pour les voyelles pour une question de place). Certains locuteurs génèrent beaucoup d'erreurs de classification alors que pour d'autres, le modèle fait considérablement moins d'erreurs. Par exemple, le modèle fait le plus d'erreurs de classification pour les locuteurs MT01 et MT04. Sur le total des erreurs de classification, 37% des voyelles mal classées proviennent du locuteur MT04 (soit 17 erreurs sur 46). Le locuteur MT01 recueille 13 occurrences incorrectes (soit 28% du total des erreurs de classification) tandis que le locuteur MT03 n'a qu'une seule erreur. En outre, les erreurs sur les voyelles non nasales ne se produisent que pour les locuteurs MT01 et MT05. Pour les autres locuteurs, le modèle fonctionne correctement sur les voyelles non nasales, les erreurs ne se produisant que pour les nasales.

Nous observons que les erreurs de classification apparaissent principalement entre /a/ et /ɑ̃/. Plusieurs contextes phonétiques peuvent être considérés comme des facteurs de confusion. D'une part, lorsqu'il y a une pause dans le contexte gauche, les voyelles /a/ ont tendance à être classées comme nasales par notre modèle (5 erreurs sur 10 de /a/, soit 50%). En revanche, la présence d'une consonne labiale ou coronale devant les voyelles /ɑ̃/ peut influencer la décision de sa classe (respectivement 5 et 6 sur 14 erreurs de /ɑ̃/). Nous remarquons la même influence lorsqu'une pause est située après ces voyelles (5 des 14 erreurs de classification de /ɑ̃/). Ces 3 contextes pour les voyelles /ɑ̃/ apparaissent également dans les erreurs pour les autres voyelles nasales. Sur 36 classifications incorrectes de voyelles nasales, 12 erreurs sont causées par le contexte des consonnes labiales et coronales précédant les voyelles nasales, et 9 erreurs se produisent avec une pause en contexte gauche.

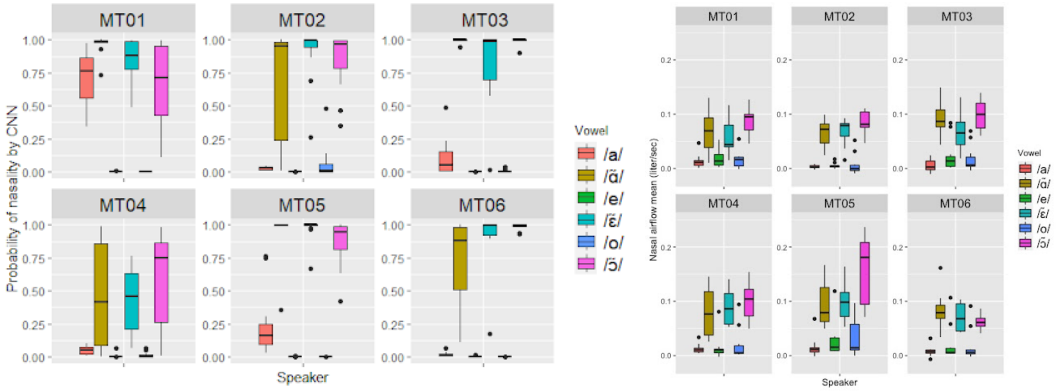


FIGURE 2 – Probabilités de nasalité obtenues par le modèle CNN pour chaque locuteur et chaque type de voyelle (à gauche), Moyenne du débit d'air nasal pour chaque locuteur et chaque type de voyelle (à droite)

3.2 Résultats aérodynamiques

Dans la figure 2 à droite, nous observons une plus grande quantité de débit d'air nasal pour les 3 voyelles nasales. Les résultats de l'ANOVA étaient statistiquement significatifs avec p<0,001 pour tous les locuteurs pour la distinction de la NAF entre les voyelles nasales et orales.

Nous constatons que chaque locuteur a un niveau minimum de débit d'air nasal qui diffère entre les voyelles. De plus, pour chaque paire de voyelles (/a/ vs. /ɑ̃/, /e/ vs. /ɛ̃/, et /o/ vs. /ɔ̃/), le débit d'air

nasal maximum pour les voyelles orales peut être plus important que le débit d'air nasal minimum pour les voyelles nasales. Ceci explique évidemment certaines erreurs de classification faites par le CNN, en particulier pour les locuteurs MT01, MT04 et MT05. Pour le locuteur MT02, le débit d'air nasal moyen pour /ɑ̃/ est de 0,065 l/s, et toutes les occurrences de /ɑ̃/, sauf une, ont été mal classées en dessous de ce seuil. En ce qui concerne le locuteur MT04, toutes les occurrences de /ɛ̃/, sauf deux, ont été mal classées en dessous d'un seuil de 0,059 l/s. De nombreux exemples de mauvaises classifications ont été trouvés selon ces critères. Globalement, un coefficient de corrélation de Pearson a révélé que la prédiction de la nasalité et de la non-nasalité est corrélée avec la mesure moyenne de la NAF (avec r = 0,66).

4 Discussion et conclusion

Le principal résultat de ce travail est la classification automatique correcte de la nasalité pour les voyelles jusqu'à 88% à partir d'un nouveau corpus acoustique, et 89% lorsqu'on inclue des consonnes dans l'entraînement et dans le test. Nous avons montré qu'il existe un lien significatif entre les probabilités CNN et les données aérodynamiques. Les erreurs de classification du CNN pour les locuteurs MT01 et MT04 ou pour la distinction entre /a/ et /ɑ̃/ sont corrélées à des différences plus faibles dans le débit d'air nasal. La même tendance des erreurs de classification de CNN entraîné avec des voyelles a également été observée pour le modèle incluant des voyelles et consonnes : les erreurs sont globalement les plus fréquentes chez les locuteurs MT01, MT04 et MT05.

Notre objectif était également d'établir une corrélation entre les probabilités CNN et le niveau moyen de débit d'air nasal par locuteur afin d'évaluer la nasalité globale par locuteur, et ce point doit encore être approfondi. Une première analyse a montré que les valeurs de probabilité données par le CNN n'étaient pas liées à la quantité de débit d'air nasal, mais que les erreurs de classification parviennent à donner de bonnes indications à ce sujet.

Reste à déterminer pourquoi les erreurs de classification portent plus fréquemment sur les voyelles orales pour le locuteur MT01 et sur les voyelles nasales pour le locuteur MT04. Pour le premier, les valeurs de débit d'air nasal sur les voyelles orales sont moyennes comparativement aux autres locuteurs alors que le débit d'air nasal est bas sur les voyelles nasales. Pour le deuxième, les valeurs de débit d'air nasal sur les voyelles nasales sont moyennes comparativement aux autres locuteurs alors que le débit d'air nasal est haut sur les voyelles orales. Ce résultat surprenant pourrait s'expliquer par d'autres paramètres mis en place par les locuteurs telles que la durée phonétique ou l'intensité, et qui montreraient une stratégie différente de la norme dans la distinction orale vs. nasale. Une étude perceptive de plusieurs items pour ces locuteurs permettrait de répondre à cette question.

Dans un futur proche, en incluant tous les phonèmes de la parole, nous travaillerons sur les fonctions d'activation afin de mieux relier les valeurs de probabilité avec le niveau de débit d'air nasal. L'analyse des zones spectrales utilisées par un CNN pourrait être un élément important de notre travail car la modélisation de la relation entre l'acoustique et l'articulation est encore problématique pour les nasales. Par exemple, des études articulatoires ont montré que le vélum du /a/ est plus bas que celui des autres voyelles, ce qui devrait avoir un impact sur le taux de classification (Durand, 1953). Dans l'ensemble, les implications de ces résultats devraient aider les phonéticiens dans leur analyse des voyelles nasales, et il est prévu de partager ce modèle et ce masque aérodynamique avec la communauté. Pour conclure, le projet de comparaison entre le test de perception et le CNN en anglais ainsi que l'implémentation du système Wav2vec sont en cours de développement.

Références

ABDERRAZEK S., FREDOUILLE C., GHIO A., LALAIN M., MEUNIER C. & WOISARD V. (2022a). Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders step 2 : contribution of the emergence of phonetic traits. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7387–7391 : IEEE.

ABDERRAZEK S., FREDOUILLE C., GHIO A., LALAIN M., MEUNIER C. & WOISARD V. (2022b). Validation of the neuro-concept detector framework for the characterization of speech disorders : A comparative study including dysarthria and dysphonia. In *Interspeech 2022*.

AJILI M., BONASTRE J.-F., ROSSETTO S. & KAHN J. (2016). Inter-speaker variability in forensic voice comparison : a preliminary evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2114–2118 : IEEE.

AMELOT A. (2004). *Etude aérodynamique, fibroscopique, acoustique et perceptive des voyelles nasales du français*. Thèse de doctorat, Université de la Sorbonne nouvelle-Paris III.

BASBØLL H. (1985). Ian maddieson (1984). patterns of sounds. with a chapter contributed by sandra ferrari disner.(cambridge studies in speech science and communication) cambridge : Cambridge university press. pp. ix+ 422. *Phonology*, **2**(1), 343–353.

BASSET P., AMELOT A., VAISSIÈRE J. & ROUBEAU B. (2001). Nasal airflow in french spontaneous speech. *Journal of the international phonetic association*, **31**(1), 87–99.

BELL-BERTI F. & KRAKOW R. A. (1991). Anticipatory velar lowering : A coproduction account. *The Journal of the Acoustical Society of America*, **90**(1), 112–123.

CHANCLU A., GEORGETON L., FREDOUILLE C. & BONASTRE J.-F. (2020). Ptsvox : une base de données pour la comparaison de voix dans le cadre judiciaire. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, p. 73–81 : ATALA ; AFCP.

CLARKE W. M. (1975). The measurement of the oral and nasal sound pressure levels of speech. *Journal of Phonetics*, **3**(4), 257–262.

CLUMECK H. (1976). Patterns of soft palate movements in six languages. *Journal of phonetics*, **4**(4), 337–351.

CROFT C. B., SHPRINTZEN R. J. & RAKOFF S. J. (1981). Patterns of velopharyngeal valving in normal and cleft palate subjects : A multi-view videofluoroscopic and nasendoscopic study. *The Laryngoscope*, **91**(2), 265–271.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DURAND M. (1953). De la formation des voyelles nasales. *Studia Linguistica*, **7**(1-2), 33–53.

ELMERICH A., AMELOT A., MAEDA S., LAPRIE Y., PAPON J. F. & CREVIER-BUCHMAN L. (2020). F1 and f2 measurements for french oral vowel with a new pneumotachograph mask. In *ISSP 2020-12th International Seminar on Speech Production*.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). The ester evaluation campaign for the rich transcription of french broadcast news. In *LREC*.

HONDA K. & MAEDA S. (2008). Glottal-opening and airflow pattern during production of voiceless fricatives : a new non-invasive instrumentation. *The Journal of the Acoustical Society of America*, **123**(5), 3738–3738.

IOFFE S. & SZEGEDY C. (2015). Batch normalization : Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, p. 448–456 : pmlr.

KAHN J., AUDIBERT N., BONASTRE J.-F. & ROSSATO S. (2011). Inter and intra-speaker variability in french : An analysis of oral vowels and its implication for automatic speaker verification. In *ICPhS*, p. 1002–1005.

KRAKOW R. A. (1993). Nonsegmental influences on velum movement patterns : Syllables, sentences, stress, and speaking rate. In *Nasals, nasalization, and the velum*, p. 87–116. Elsevier.

MOHAMMED M. A., ABDULKAREEM K. H., MOSTAFA S. A., KHANAPI ABD GHANI M., MAASHI M. S., GARCIA-ZAPIRAIN B., OLEAGORDIA I., ALHAKAMI H. & AL-DHIEF F. T. (2020). Voice pathology detection and classification using convolutional neural network model. *Applied Sciences*, **10**(11), 3723.

PELLEGRINI T. & MOUYSSET S. (2016). Inferring phonemic classes from cnn activation maps using clustering techniques. In *Annual conference Interspeech (INTERSPEECH 2016)*, p. pp–1290.

SKOLNICK M. L., MCCALL G. N. & BARNES M. (1973). The sphincteric mechanism of velopharyngeal closure. *The Cleft Palate Journal*, **10**(3), 286–305.

STYLER W. (2017). On the acoustical features of vowel nasality in english and french. *The Journal of the Acoustical Society of America*, **142**(4), 2469–2482.

TEAM R. D. C. (2009). A language and environment for statistical computing. *http ://www. R-project. org*.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**(3), 201–212.

VAISSIÈRE J. (1988). Prediction of velum movement from phonological specifications. *Phonetica*, **45**(2-4), 122–139.

WANG X., TANG M., YANG S., YIN H., HUANG H. & HE L. (2019). Automatic hypernasality detection in cleft palate speech using cnn. *Circuits, Systems, and Signal Processing*, **38**, 3521–3547.

ZERLING J.-P. (1984). Phénomènes de nasalité et de nasalisation vocalique : Étude cinéradiographique pour deux locuteurs. *Travaux de l'Institut de Phonétique de Strasbourg Strasbourg*, (16), 241–266.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

DrBERT: Un modèle robuste pré-entraîné en français pour les domaines biomédical et clinique

Yanis Labrak[1,4], Adrien Bazoge[2,3], Richard Dufour[2], Mickael Rouvier[1], Emmanuel Morin[2], Béatrice Daille[2] and Pierre-Antoine Gourraud[3]
(1) LIA - Avignon Univserité (2) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France (3) Nantes Université, Clinique des données, CHU de Nantes (4) Zenidoc
{prenom.nom}@univ-avignon.fr, {prenom.nom}@univ-nantes.fr

RÉSUMÉ

Ces dernières années, les modèles de langage pré-entraînés ont obtenu les meilleures performances sur un large éventail de tâches de traitement automatique du langage naturel (TALN). Alors que les premiers modèles ont été entraînés sur des données issues de domaines généraux, des modèles spécialisés sont apparus pour traiter plus efficacement des domaines spécifiques. Dans cet article, nous proposons une étude originale de modèles de langue dans le domaine médical en français. Nous comparons pour la première fois les performances de modèles entraînés sur des données publiques issues du web et sur des données privées issues d'établissements de santé. Nous évaluons également différentes stratégies d'apprentissage sur un ensemble de tâches biomédicales. Enfin, nous publions les premiers modèles spécialisés pour le domaine biomédical en français, appelés DrBERT, ainsi que le plus grand corpus de données médicales sous licence libre sur lequel ces modèles sont entraînés.

ABSTRACT

DrBERT : A Robust Pre-trained Model in French for Biomedical and Clinical domains.

In recent years, pre-trained language models (PLMs) achieve the best performance on a wide range of natural language processing tasks. While the first models were trained on general domain data, specialized ones have emerged to more effectively treat specific domains. In this paper, we propose an original study of PLMs in the medical domain on French language. We compare, for the first time, the performance of PLMs trained on both public data from the web and private data from healthcare establishments. We also evaluate different learning strategies on a set of biomedical tasks. Finally, we release the first specialized PLMs for the biomedical field in French, called DrBERT, as well as the largest corpus of medical data under free license on which these models are trained.

MOTS-CLÉS : BERT ; RoBERTa ; Transformers ; Biomédical ; Clinique ; Modèle de langue.

KEYWORDS: BERT ; RoBERTa ; Transformers ; Biomedical ; Clinical ; Language Model.

1 Introduction

Au cours des dernières années, les modèles de langage pré-entraînés ont permis d'améliorer considérablement les performances de nombreuses tâches du traitement automatique du langage naturel (TALN). Les modèles récents, tels que BERT (Devlin *et al.*, 2019) ou RoBERTa (Liu *et al.*, 2019), tirent profit des énormes quantités de données non étiquetées grâce à des approches non-supervisées fondées sur l'architecture Transformers (Vaswani *et al.*, 2017). La plupart de ces modèles sont souvent pré-entraînés sur des corpus de domaines généraux et une étape supplémentaire de réglage fin

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 109–120.

(*fine-tuning*) peut être appliquée pour utiliser ces modèles sur une tâche cible (Devlin *et al.*, 2019).

L'adaptation des modèles de langue à un domaine de spécialité suit généralement deux stratégies. La première consiste à entraîner à partir de zéro (*from scratch*) un nouveau modèle en utilisant uniquement les données de la spécialité cible. La seconde approche, appelée pré-entraînement continué (Howard & Ruder, 2018), poursuit l'entraînement de modèles déjà pré-entraînés, permettant de passer d'un modèle générique à un modèle spécialisé. Bien que des études aient montré que la première stratégie offre généralement de meilleures performances (Lee *et al.*, 2019), la seconde nécessite un nombre de ressources moins important (Chalkidis *et al.*, 2020; El Boukkouri *et al.*, 2022) que ce soit en termes de ressources informatiques ou de quantité de données.

Récemment, de multiples modèles de langue ont été développés pour les domaines biomédical et clinique, principalement pour l'anglais. Les modèles BioBERT (Lee *et al.*, 2019), BlueBERT (Peng *et al.*, 2019) et ClinicalBERT (Huang *et al.*, 2019) ont utilisé le pré-entraînement continué à partir d'un modèle BERT générique. D'autres modèles, comme SciBERT (Beltagy *et al.*, 2019) et PubMedBERT (Gu *et al.*, 2021), ont été pré-entraînés à partir de zéro. Dans les langues autres que l'anglais, les modèles s'appuyant sur BERT sont plus rares et reposent principalement sur un pré-entraînement continué. Pour le français, il n'existe, à notre connaissance, aucun modèle disponible publiquement pour le domaine biomédical.

Dans cet article, nous décrivons et diffusons librement DrBERT, les premiers modèles spécialisés dans le domaine biomédical pour le français, ainsi que le corpus qui a permis leur entraînement. Nous proposons également une étude originale évaluant les différentes stratégies de pré-entraînement de modèles de langue pour le domaine médical, en comparant les modèles DrBERT avec un modèle pré-entrainé sur des données cliniques privées, nommé ChuBERT. Nos contributions sont :

— Un nouveau *benchmark* agrégeant un ensemble de tâches de TALN dans le domaine médical en français, permettant d'évaluer les modèles de langue au niveau syntaxique et sémantique.
— Un corpus de données textuelles, nommé NACHOS, rassemblant plusieurs sources biomédicales en ligne.
— La construction et l'évaluation des premiers modèles de langues libres en français pour le domaine biomédical fondés sur l'architecture RoBERTa, appelés DrBERT, comprenant l'analyse de différentes stratégies de pré-entraînement.
— Un ensemble de modèles utilisant des données publiques et privées entraînés sur des tailles de données comparables. Ces modèles ont ensuite été comparés en évaluant leurs performances sur un large éventail de tâches, tant publiques que privées.
— La distribution des modèles de langue publics sous la licence open-source MIT ainsi que le corpus NACHOS sous la licence CC0 1.0.

2 Corpus de pré-entraînement

Les travaux précédents dans le domaine biomédical (Gu *et al.*, 2021) sur les modèles de langue ont souligné l'importance de faire correspondre les sources de données utilisées lors de l'entraînement avec les tâches cibles. En raison de leur nature sensible, les données cliniques sont extrêmement difficiles à obtenir. La collecte massive de données web relatives à ce domaine apparaît comme une solution permettant de pallier ce manque.

Le tableau 1 donne un aperçu général des deux corpus collectés. Les données publiques issues du web ont permis la constitution d'un corpus, appelé NACHOS$_{large}$, à partir de sources disponibles ouvertement en ligne contenant 7,4 Go de données. Le jeu de données privées, appelé XBDW$_{small}$,

Corpus	Taille	#mots	#phrases
NACHOS$_{large}$ (public)	7,4 Go	1,1 G	54,2 M
NACHOS$_{small}$ (public)	4 Go	646 M	25.3 M
XBDW$_{small}$ [1] (privé)	4 Go	655 M	43.1 M
XBDW$_{mixed}$ (public+privé)	4+4 Go	1,3 G	68,4 M

TABLE 1 – Aperçu des corpus de données publiques (NACHOS) et privées (XBDW) collectées.

contient 4 Go de données provenant du Centre Hospitalier Universitaire (CHU) de XXX [2]. Afin d'effectuer des expériences comparables, nous avons extrait un sous-corpus NACHOS (NACHOS$_{small}$) de la même taille que les données privées.

Corpus public - NACHOS opeN crAwled frenCh Healthcare cOrpuS (NACHOS) est un ensemble de données textuelles médicales françaises distribué de façon libre et obtenu à partir de la collecte massive d'une variété de sources textuelles autour du sujet médical sur le web. Il se compose de plus d'un milliard de mots, tirés de 24 sites web francophones de haute qualité. Le corpus comprend un large éventail d'informations médicales : descriptions de maladies, fiches d'information sur les traitements et les médicaments, conseils généraux sur la santé, rapports de réunions scientifiques officielles, cas cliniques anonymisés, littérature scientifique, thèses, cours de médecine universitaire, ainsi que de nombreuses données obtenues à partir de sources textuelles brutes, de *scrapping web* et de Reconnaissance Optique de Caractères (ROC) comme présenté dans le Tableau 6 en annexe. Nous avons utilisé des heuristiques pour découper les textes en phrases et filtrer les phrases courtes ou de mauvaise qualité comme celles obtenues par ROC. Ensuite, nous les classons par langues en utilisant notre propre classificateur entraîné sur les corpus multilingues Opus EMEA (Tiedemann & Nygaard, 2004) et MASSIVE (FitzGerald *et al.*, 2022) pour ne garder que les phrases en français. Pour la version 4 Go de NACHOS (NACHOS$_{small}$), nous avons mélangé l'ensemble du corpus et sélectionné au hasard 25,3 millions de phrases afin de maximiser l'homogénéité des sources de données.

Corpus privé - XBDW Le corpus privé, appelé XXX Biomedical Data Warehouse (XBDW), a été obtenu en utilisant l'entrepôt de données du CHU de XXX. Cet entrepôt de données comprend différentes dimensions de données relatives aux patients : socio-démographiques, prescriptions médicamenteuses et autres informations associées au séjour hospitalier (diagnostic, biologie, imagerie, etc.). L'autorisation de mise en œuvre et d'exploitation de l'entrepôt de données XBDW a été accordée en XXXX [3] par la CNIL (Commission Nationale de l'Informatique et des Libertés) ; autorisation N° XXX [4]. Pour ce travail, un échantillon de 1,7 millions de comptes-rendus hospitaliers désidentifiés a été sélectionné aléatoirement et extrait de l'entrepôt de données. Les comptes-rendus proviennent de différents services hospitaliers, tels que les urgences, la gynécologie et la cardiologie. Ce corpus contient 655 millions de mots, issus de 43,1 millions de phrases, pour une taille totale d'environ 4 Go.

3 Pré-entraînement des modèles

3.1 Impact des données

L'un des problèmes consiste à identifier la quantité de données nécessaires pour créer un modèle performant et capable de rivaliser avec les modèles pré-entraînés sur des domaines généraux. Des études récentes, comme celles de Martin *et al.* (2020) et Zhang *et al.* (2021), discutent de l'impact de

2. Nom de la ville masqué à des fins d'anonymisation lors de la relecture par les pairs.
3. Année cachée pour le double aveugle.
4. Numéro d'autorisation caché pour le double aveugle.

la taille des données de pré-entraînement sur la performance des modèles et montrent que certaines tâches bénéficient d'une quantité moindre de données. Dans le domaine médical, aucune étude n'a été menée pour comparer l'impact de la variation de la quantité de données sur le pré-entraînement, ou pour évaluer l'impact de la qualité des données en fonction de leur source de collecte.

Nous proposons donc d'évaluer l'impact des différentes sources de données en comparant NACHOS$_{small}$ et XBDW$_{small}$ entre elles comme décrit dans la section 2. De plus, nous proposons de comparer les résultats obtenus avec ceux d'un modèle pré-entraîné sur une quantité de données plus grande (NACHOS$_{large}$) afin d'étudier si le fait de disposer de presque deux fois plus de données permet d'améliorer les performances. Pour finir, nous avons évalué une combinaison des corpus public (NACHOS$_{small}$) et privé (XBDW$_{small}$) pour un total de 8 Go (XBDW$_{mixed}$), afin de démontrer si la combinaison de données de qualité variable permettent des représentations complémentaires.

3.2 Stratégies de pré-entraînement

En plus de l'analyse de la taille et des sources de données, nous cherchons également à évaluer trois stratégies de pré-entraînement des modèles de langue pour le domaine médical :

— Pré-entraînement du modèle à partir de zéro, incluant une tokenization spécifique des mots à partir de nos données d'apprentissage.
— Poursuivre le pré-entraînement d'un modèle de langue général pour le français, ici CamemBERT, sur nos données du domaine médical, tout en conservant le tokenizer initial (*i.e.* celui de CamemBERT).
— Poursuivre le pré-entraînement d'un modèle de langage spécifique au domaine médical anglais, appelé PubMedBERT, sur nos données en français, en conservant le tokenizer initial.

Concernant la dernière stratégie, l'objectif est de comparer les performances d'un modèle médical anglais pré-entraîné sur nos données médicales françaises, à un autre modèle basé sur un modèle générique français. En effet, ces deux langues partagent une terminologie de spécialité commune.

Modèle	Architecture	Stratégie de pré-entraînement	Corpus
DrBERT	RoBERTa	À partir de zéro	NACHOS$_{large}$
DrBERT	RoBERTa	À partir de zéro	NACHOS$_{small}$
ChuBERT	RoBERTa	À partir de zéro	XBDW$_{small}$
ChuBERT	RoBERTa	À partir de zéro	XBDW$_{mixed}$
CamemBERT	RoBERTa	Pré-entraînement continué	NACHOS$_{small}$
PubMedBERT	BERT	Pré-entraînement continué	NACHOS$_{small}$
CamemBERT	RoBERTa	Pré-entraînement continué	XBDW$_{small}$

TABLE 2 – Liste des configurations des modèles pré-entraînés.

Le tableau 2 résume toutes les configurations évaluées dans cet article, intégrant à la fois l'étude de la taille des données et les stratégies de pré-entraînement.

4 Jeux de données et tâches d'évaluation

Pour évaluer les différentes configurations de pré-entraînement de nos modèles, un ensemble de tâches biomédicales est nécessaire. Si un tel *benchmark* spécifique au domaine existe pour l'anglais (BLURB (Gu *et al.*, 2021)), il n'en existe aucun pour le français. Dans cette section, nous décrivons un *benchmark* original, résumé dans le tableau 3, intégrant diverses tâches biomédicales de TALN pour le français. Parmi celles-ci, certaines proviennent de corpus publics, permettant la réplication de nos expériences. D'autres tâches proviennent de corpus privés et ne peuvent être partagées. Cependant,

Thématique / Corpus	Tâche	Métrique	Entraînement	Validation	Test
Corpus Publics					
ESSAIS (Dalloux *et al.*, 2021)	POS Tagging	Macro F1	9 693	2 077	2 078
CAS : Corpus de cas cliniques (Grabar *et al.*, 2018)	POS Tagging	Macro F1	5 306	1 137	1 137
MUSCA-DET - Déterminants sociaux de santé (Task 1)	REN imbriquée	Macro F1	19 861	2 207	5 518
MUSCA-DET - Déterminants sociaux de santé (Task 2)	Classification Multi-label	Macro F1	19 861	2 207	5 518
QUAERO French Medical Corpus - EMEA (Névéol *et al.*, 2014)	REN imbriquée	Weighted F1	11	12	15
QUAERO French Medical Corpus - MEDLINE (Névéol *et al.*, 2014)	REN imbriquée	Weighted F1	833	832	833
FrenchMedMCQA (Labrak *et al.*, 2022)	MCQA	EMR / Hamming Score	2 171	312	622
Corpus Privés					
Structuration de l'insuffisance cardiaque aiguë dans les comptes-rendus	REN	Macro F1	2 527	281	703
Classification de l'insuffisance cardiaque aiguë	Classification Binaire	Macro F1	1 179	132	328
Tri des comptes-rendus par spécialité	Classification Multi-classe	Macro F1	4 413	1 470	1 473
Structuration des prescriptions dans les comptes-rendus	REN	Macro F1	61	15	26

TABLE 3 – Corpus, tâches et métrique utilisés pour évaluer les modèles de langue.

elles sont utiles pour évaluer nos modèles avec plus de précision et ainsi observer leurs capacités de généralisation.

5 Résultats et discussions

Comme décrit précédemment, nous évaluons les performances de nos modèles de langue pré-entraînés sur un ensemble de tâches publiques et privées liées au domaine biomédical. Nous proposons d'analyser les résultats en fonction des différentes stratégies de pré-entraînement (section 5.1) puis de discuter de l'impact des données de pré-entraînement, que ce soit en termes de taille ou de nature (section 5.2).

	aHF NER			aHF classification			NER Medical Report			Specialities Classification		
	P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
CamemBERT OSCAR 138 GB	40,89	35,22	35,13	81,90	79,12	80,13	87,98	91,66	89,35	99,32	99,09	99,20
CamemBERT OSCAR 4 GB	46,32	43,17	42,66	81,49	81,42	81,41	87,79	90,74	88,78	99,53	99,69	99,61
CamemBERT CCNET 4 GB	47,25	42,2	43,11	82,02	79,30	79,98	87,61	92,28	89,34	99,54	99,55	99,55
DrBERT NACHOS$_{large}$	55,29	46,66	48,22	81,33	81,25	81,25	87,99	**92,80**	**89,83**	99,82	**99,90**	**99,86**
DrBERT NACHOS$_{small}$	54,55	43,39	45,93	79,85	80,10	79,87	87,57	92,76	89,44	**99,85**	99,85	99,85
ChuBERT XBDW$_{small}$	**56,92**	47,46	49,01	81,03	**82,67**	81,56	87,76	92,63	89,58	99,76	**99,90**	99,83
ChuBERT XBDW$_{mixed}$	54,62	47,81	**49,14**	82,23	81,71	**81,98**	87,42	92,36	89,30	99,81	99,82	99,81
CamemBERT NACHOS$_{small}$	22,02	16,67	16,08	74,86	69,82	69,80	65,72	68,49	66,74	99,44	99,67	99,54
PubMedBERT NACHOS$_{small}$	53,44	**48,21**	48,72	**83,06**	80,39	81,40	87,35	92,69	89,36	99,52	99,58	99,55
CamemBERT XBDW$_{small}$	25,44	19,33	19,12	79,50	74,74	76,02	68,80	71,23	69,64	99,60	99,57	99,58

TABLE 4 – Performance sur nos tâches biomédicales privées. Le meilleur modèle est en gras et le second est souligné. Pour les modèles CamemBERT NACHOS$_{small}$ et CamemBERT XBDW$_{small}$, le modèle CamemBERT OSCAR 138 GB a été utilisé comme abse initiale pour le pré-entraînement continué.

Tous les modèles ont été réglés finement (*fine-tuned*) de la même façon pour toutes les tâches. Tous les résultats rapportés sont la moyenne des scores de quatre exécutions. Les performances obtenues sur les tâches biomédicales sont présentées dans les tableaux 4 et 5 pour les tâches privées et publiques respectivement. Pour des raisons de lisibilité, la première partie de chaque tableau présente les résultats des modèles déjà existants, la deuxième partie nos modèles spécialisés entraînés à partir de zéro, et la dernière partie nos modèles utilisant un pré-entraînement continué.

5.1 Impact des stratégies de pré-entraînement

Comme montré dans les tableaux 4 et 5, les modèles pré-entraînés à partir de zéro (DrBERT NACHOS et ChuBERT XBDW) obtiennent les meilleurs scores F1 sur toutes les tâches privées et sur la majorité

	MUSCA-DET T1			MUSCA-DET T2			ESSAI POS			CAS POS			FrenchMedMCQA		QUAERO-EMEA			QUAERO-MEDLINE		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*Hamming*	*EMR*	*P*	*R*	*F1*	*P*	*R*	*F1*
CamemBERT OSCAR 138 GB	89,04	88,59	88,54	89,87	87,12	88,20	81,57	81,01	81,10	96,37	94,53	95,22	36,24	**16,55**	90,57	91,06	90,71	76,58	78,67	77,41
CamemBERT OSCAR 4 GB	86,09	85,45	85,43	92,68	90,34	91,27	84,01	83,51	83,69	98,15	95,34	96,42	35,75	15,37	90,75	91,16	90,83	78,55	79,33	78,76
CamemBERT CCNET 4 GB	91,12	89,91	90,33	93,10	90,42	91,38	85,60	85,63	85,42	**98,19**	**96,75**	**97,33**	34,71	14,41	90,31	90,59	90,33	78,06	78,11	77,61
DrBERT NACHOS$_{large}$	92,10	90,27	91,04	**94,97**	90,41	92,24	**90,96**	**89,19**	**89,75**	97,37	94,49	95,65	36,66	15,32	**91,93**	**92,52**	**92,09**	77,85	78,54	77,88
DrBERT NACHOS$_{small}$	93,35	90,62	91,77	91,31	86,60	88,57	90,12	88,37	88,76	97,04	94,88	95,70	**37,37**	13,34	91,54	92,00	91,66	77,91	79,34	78,18
ChuBERT XBDW$_{small}$	**94,88**	90,79	92,23	94,77	90,27	92,17	88,53	87,73	87,71	97,00	94,65	95,61	35,16	14,79	88,11	88,78	88,15	75,05	76,57	74,94
ChuBERT XBDW$_{mixed}$	94,39	**91,93**	**92,73**	94,22	90,02	91,71	86,36	85,50	85,73	97,77	95,30	96,35	34,58	12,21	90,36	90,94	90,52	78,61	79,32	78,63
CamemBERT NACHOS$_{small}$	81,44	81,39	80,96	79,74	78,08	78,70	80,59	79,88	80,04	95,64	91,57	92,46	32,87	13,76	67,56	77,48	71,10	55,45	62,34	57,43
PubMedBERT NACHOS$_{small}$	92,51	91,49	91,53	94,95	**92,55**	**93,62**	84,73	83,80	83,85	97,82	96,12	96,81	35,88	15,21	90,97	91,27	91,03	**82,03**	**81,71**	**81,73**
CamemBERT XBDW$_{small}$	82,35	81,59	81,57	78,14	76,38	77,12	79,44	79,79	79,25	95,98	92,11	93,18	27,73	11,89	53,44	73,11	61,75	48,71	61,33	53,05

TABLE 5 – Performance sur les tâches biomédicales publiques. Le meilleur modèle est en gras et le second est souligné. Pour les modèles CamemBERT NACHOS$_{small}$ et CamemBERT XBDW$_{small}$, le modèle CamemBERT OSCAR 138 GB a été utilisé comme abse initiale pour le pré-entraînement continué.

des tâches publiques (5 tâches sur 7). Les deux tâches publiques restantes (MUSCA-DET T2 et Quaero-Medline) sont mieux traitées avec PubMedBERT NACHOS$_{small}$, un modèle pré-entraîné deux fois, une première fois sur des données biomédicales anglaises, puis sur nos données biomédicales françaises (NACHOS$_{small}$). Nous observons également que le pré-entraînement continué à partir des modèles génériques français (CamemBERT NACHOS$_{small}$ ou CamemBERT XBDW$_{small}$) est moins performant que le pré-entraînement à partir de zéro. Enfin, les modèles état de l'art pré-entraînés sur des données génériques (CamemBERT OSCAR) restent compétitifs dans un certain nombre de tâches publiques biomédicales (CAS POS, FrenchMCQA ou MUSCA-DET T2), mais rencontrent plus de difficultés sur les tâches privées. Cela met en évidence la difficulté des tâches privées lorsque les données de pré-entraînement sont moins spécialisées.

5.2 Impact des données

En ce qui concerne la quantité de données utilisées pour le pré-entraînement des modèles ($small$ vs. $large$ ou $mixed$), les résultats montrent que plus les quantités de données sont grandes, plus les modèles sont performants, quelle que soit la stratégie de pré-entraînement ou la source de données (privée ou publique). Nous remarquons une nette domination des modèles pré-entraînés sur des sources web, notamment OSCAR et NACHOS, lorsqu'ils sont appliqués à des tâches publiques. En effet, les modèles reposant sur des données privées XBDW n'obtiennent les meilleures performances (en termes de score F1) que pour la tâche MUSCA-DET T1. Cette tendance n'est pas tout à fait observée sur les tâches privées, où les modèles fondés sur XBDW obtiennent des performances similaires, voire meilleures, lorsqu'ils sont mélangés avec des données biomédicales publiques (ChuBERT XBDW$_{mixed}$), comme le montre le tableau 4. Nous pensons que cette divergence est principalement due à la nature différente des données traitées. Enfin, nous observons que les modèles issus d'un pré-entraînement continué en partant d'un modèle spécialisé anglais (ici PubMedBERT) ont des performances supérieures à celles des modèles fondés sur CamemBERT, corroborant en partie notre hypothèse sur l'efficacité du transfert de connaissances inter-langues.

6 Conclusion

Dans ce travail, nous avons introduit les premiers modèles de langage français pour le domaine biomédical et clinique fondés sur l'architecture RoBERTa. Nous proposons aussi une étude comparative sur une collection de tâches biomédicales diverses provenant de sources privées et publiques. Nos modèles open-source DrBERT ont établi des performances à l'état de l'art dans la quasi-totalité des tâches biomédicales, surpassant le modèle généraliste français (CamemBERT). De plus, nous avons démontré que les pré-entraînements sur des ressources spécialisées de taille limitées (4 Go) obtenues sur le web permettent de très souvent dépasser les modèles entraînés avec des données spécialisées provenant de comptes-rendus médicaux.

Les modèles pré-entraînés ainsi que les scripts de pré-apprentissage [5] ont été publiés publiquement en ligne sous une licence open source MIT. Pour ce qui est du corpus NACHOS, l'objectif principal est de promouvoir le développement d'outils de TALN robustes par la communauté, nous avons donc décidé de rendre les corpus accessibles pour la recherche académique seulement.

Remerciements

Ce travail a été réalisé grâce aux ressources de GENCI-IDRIS (Grant 2022-AD011013061R1 and 2022-AD011013715) et du CCIPL (Centre de Calcul Intensif des Pays de la Loire). Ce travail a été soutenu financièrement par l'ANR AIBy4 (ANR-20-THIA-0011) and Zenidoc.

Références

BELTAGY I., LO K. & COHAN A. (2019). SciBERT : A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3615–3620, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1371.

CHALKIDIS I., FERGADIOTIS M., MALAKASIOTIS P., ALETRAS N. & ANDROUTSOPOULOS I. (2020). LEGAL-BERT : The muppets straight out of law school. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 2898–2904, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.findings-emnlp.261.

DALLOUX C., CLAVEAU V., GRABAR N., OLIVEIRA L. E. S., MORO C. M. C., GUMIEL Y. B. & CARVALHO D. R. (2021). Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, **27**(2), 181–201. DOI : 10.1017/S1351324920000352.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

5. https://drbert.univ-avignon.fr/

EL BOUKKOURI H., FERRET O., LAVERGNE T. & ZWEIGENBAUM P. (2022). Re-train or Train from Scratch ? Comparing Pre-training Strategies of BERT in the Medical Domain. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, p. 2626–2633, Marseille, France : European Language Resources Association.

FITZGERALD J., HENCH C., PERIS C., MACKIE S., ROTTMANN K., SANCHEZ A., NASH A., URBACH L., KAKARALA V., SINGH R., RANGANATH S., CRIST L., BRITAN M., LEEUWIS W., TUR G. & NATARAJAN P. (2022). Massive : A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. DOI : 10.48550/ARXIV.2204.08582.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French Corpus with Clinical Cases. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, p. 1–7, Brussels, Belgium. HAL : hal-01937096.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, **3**(1), 1–23. DOI : 10.1145/3458754.

HOWARD J. & RUDER S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 328–339, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1031.

HUANG K., ALTOSAAR J. & RANGANATH R. (2019). Clinicalbert : Modeling clinical notes and predicting hospital readmission. DOI : 10.48550/ARXIV.1904.05342.

LABRAK Y., BAZOGE A., DUFOUR R., DAILLE B., GOURRAUD P.-A., MORIN E. & ROUVIER M. (2022). FrenchMedMCQA : A French Multiple-Choice Question Answering Dataset for Medical domain. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, Abou Dhabi, United Arab Emirates. HAL : hal-03824241.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2019). BioBERT : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. DOI : 10.1093/bioinformatics/btz682.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. *CoRR*, **abs/1907.11692**.

LUCCIONI A. S., VIGUIER S. & LIGOZAT A.-L. (2022). Estimating the carbon footprint of bloom, a 176b parameter language model.

MARTIN L., MULLER B., SUÁ REZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 7203—-7219 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. In *Proc of BioTextMining Work*, p. 24–30.

PENG Y., YAN S. & LU Z. (2019). Transfer learning in biomedical natural language processing : An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, p. 58–65.

TIEDEMANN J. & NYGAARD L. (2004). The OPUS corpus - parallel and free : http://logos.uio.no/opus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal : European Language Resources Association (ELRA).

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems (NIPS)*, volume 30 : Curran Associates, Inc.

ZHANG Y., WARSTADT A., LI X. & BOWMAN S. R. (2021). When do you need billions of words of pretraining data ? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1112–1125, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.90.

A Annexes

A.1 Sources NACHOS

Ressource	# mots
HAL	638 508 261
Haute Autorité de Santé (HAS)	113 394 539
Drug leaflets	74 770 229
Scrapping Web Médical	64 904 334
ANSES SAISINE	51 372 932
Base de données publique des médicaments (BDPM)	48 302 695
ISTEX	44 124 422
CRTT	26 210 756
WMT-16	10 282 494
EMEA-V3	6 601 617
Wikipedia Science de la vie	4 671 944
ANSES RCP	2 953 045
Cerimes	1 717 552
LiSSa	235 838
DEFT-2020	231 396
CLEAR	225 898
CNEDiMTS	175 416
QUAERO French Medical Corpus	72 031
ANSM Registre des essais cliniques	47 678
ECDC	44 482
QualiScope	12 718
WMT-18-Medline	7 673
Total	**1,088,867,950**

TABLE 6 – Sources de données comprises dans le corpus NACHOS.

A.2 Impact écologique

Une quantité considérable de ressources de calcul a été utilisée pour mener cette étude, puisqu'environ 18 000 heures de calcul GPU ont été utilisées pour créer les 7 modèles présentés ici, ainsi qu'environ 7 500 heures de GPU pour le débogage en raison de problèmes techniques liés aux configurations des modèles et à la mauvaise performance, pour un total de 25 500 heures. Le coût environnemental total, selon la documentation du supercalculateur Jean Zay [6] équivaut à 6 604 500 Wh soit 376,45 kg CO2eq

6. http://www.idris.fr/media/jean-zay/jean-zay-conso-heure-calcul.pdf

basé sur l'intensité carbone du réseau énergétique mentionné par l'étude de coût environnemental de BLOOM, qui elle aussi a été réalisée sur Jean Zay (Luccioni *et al.*, 2022).

A.3 Pré-traitement des corpus d'évaluation

L'extraction d'entités imbriquées n'est pas directement réalisable avec un modèle BERT sans adapter le corpus. Parmi les corpus d'évaluation, deux corpus traitent la tâche de reconnaissance d'entités nommées imbriquées : QUAERO et MUSCA-DET. Pour simplifier le processus d'évaluation de ces corpus, nous trions les étiquettes imbriquées par ordre alphabétique et les concaténons en une seule pour transformer la tâche en un format utilisable pour la classification de tokens avec les architectures basées sur BERT.

A.4 Évaluation généraliste sur le français

Le tableau 7 donne les résultats obtenus par tous les modèles sur les tâches généralistes. Ces tâches proviennent de Martin *et al.* (2020) et ont été utilisées pour évaluer les différents modèles CamemBERT. Les quatre premières sont des tâches d'étiquetage morphosyntaxique POS (GSD, SEQUOIA, SPOKEN et PARTUT) et la dernière est une tâche d'inférence en langage naturel (XNLI).

Tous les résultats de nos modèles diminuent les performances sur toutes les tâches. La baisse la plus importante concerne la tâche d'inférence en langage naturel, avec une performance de ChuBERT NBDW$_{small}$ presque 13% inférieure à celle de CamemBERT 138 Go. Nous observons également que les modèles spécialisés en anglais sont aussi performants que nos modèles biomédicaux en français. Il semble assez clair d'après les observations précédentes que les modèles spécialisés sont difficiles à généraliser à d'autres tâches, mais que des informations spécialisées capturées dans une langue pourraient être transférées dans une autre langue.

	GSD	SEQUOIA	SPOKEN	PARTUT	XNLI
CamemBERT OSCAR 138 Go	**98.28**	98.68	<u>97.26</u>	97.70	**81.94**
CamemBERT OSCAR 4 Go	98.14	**99.18**	**97.57**	<u>97.86</u>	<u>81.76</u>
CamemBERT CCNET 4 Go	<u>98.18</u>	<u>98.92</u>	97.20	**97.92**	81.26
PubMedBERT	96.48	96.49	90.00	93.97	73.79
BERT clinique	96.49	96.31	89.60	93.17	70.57
BioBERT v1.1	97.32	96.54	91.81	94.52	71.54
DrBERT NACHOS$_{large}$	96.94	98.05	95.92	96.54	72.18
DrBERT NACHOS$_{small}$	97.17	98.21	96.38	96.45	72.86
ChuBERT NBDW$_{small}$	96.45	97.38	94.90	95.83	69.00
ChuBERT NBDW$_{mixed}$	97.18	98.10	96.43	96.33	72.32
CamemBERT NACHOS$_{small}$	97.63	96.90	91.12	94.00	71.26
PubMedBERT NACHOS$_{small}$	97.41	98.71	95.54	97.01	77.35
CamemBERT NBDW$_{small}$	97.55	96.26	89.17	91.34	72.73

TABLE 7 – Performance sur les tâches généralistes. Le meilleur modèle est en gras et le deuxième est souligné.

A.5 Intersection des vocabulaires

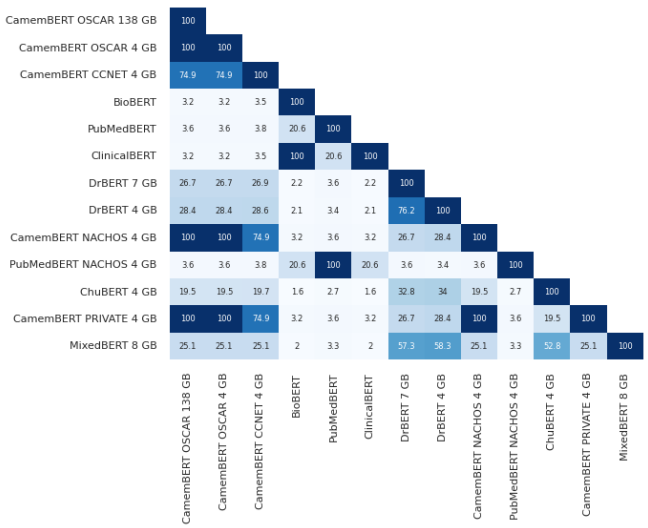
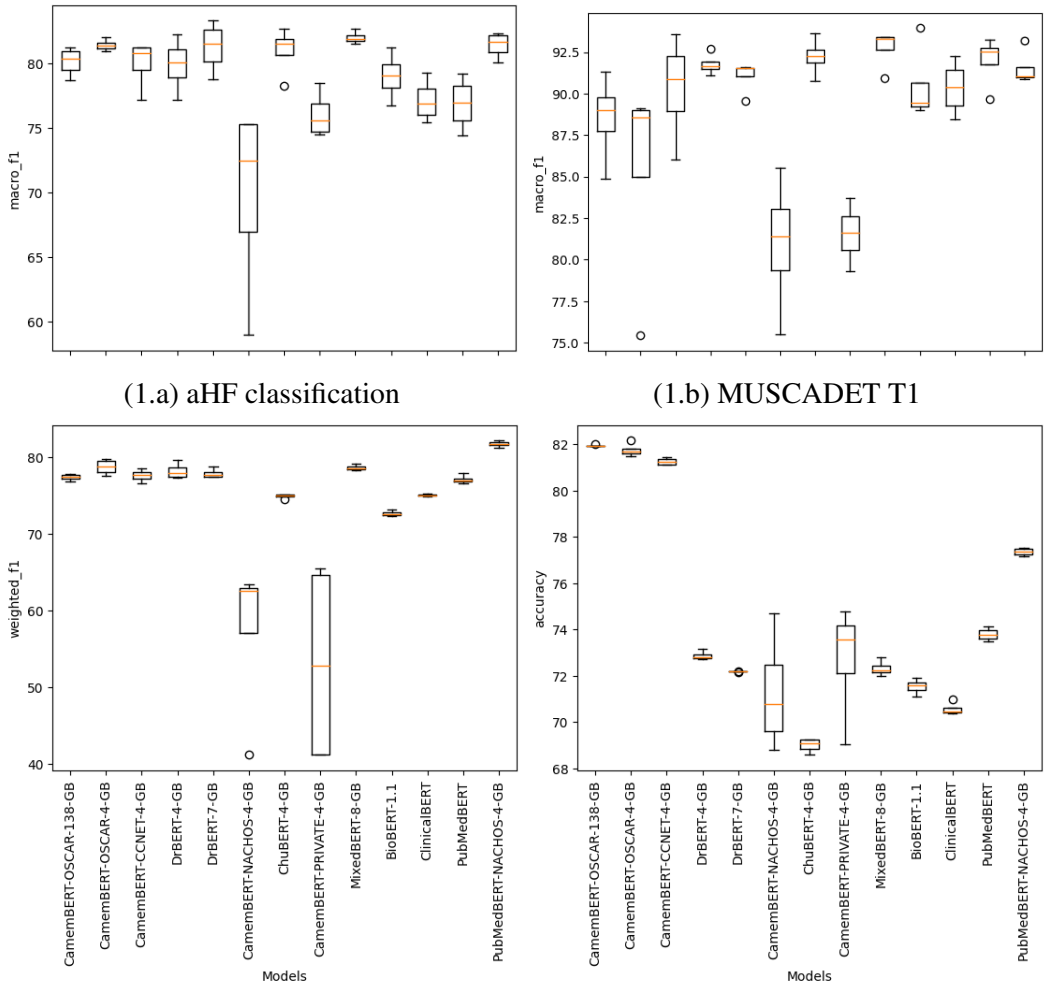


FIGURE 1 – Matrice d'intersection des vocabulaires.

Comme nous pouvons le voir sur la figure 1, malgré des performances similaires, certains modèles ne partagent pas beaucoup de vocabulaire mutuel.

A.6 Stabilité des modèles

Nous observons lors de la phase d'évaluation que la plupart des modèles basés sur la stratégie de pré-formation continue de CamemBERT OSCAR 138 Go souffrent d'une mauvaise stabilité lors de l'affinage, ce qui se traduit par une fluctuation des performances entre les exécutions.



(1.a) aHF classification

(1.b) MUSCADET T1

(2.c) QUAERO MEDLINE

(2.d) XNLI

FIGURE 2 – Boîte à moustaches pour chaque modèle.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Classification automatique de données déséquilibrées et bruitées : application aux exercices de manuels scolaires

Elise Lincker[1] Camille Guinaudeau[2,3] Olivier Pons[1] Jérôme Dupire[1]
Céline Hudelot[4] Vincent Mousseau[4] Isabelle Barbet[1] Caroline Huron[5,6]
(1) Cedric, CNAM, Paris, France
(2) Japanese French Laboratory for Informatics, CNRS, NII, Tokyo, Japon
(3) Université Paris-Saclay, Orsay, France
(4) MICS, CentraleSupélec, Université Paris-Saclay, Orsay, France
(5) SEED, Inserm, Université Paris Cité, Paris, France
(6) Learning Planet Institute, Paris, France
elise.lincker@lecnam.net, guinaudeau@nii.ac.jp, olivier.pons@lecnam.net, jerome.dupire@lecnam.net, celine.hudelot@centralesupelec.fr, vincent.mousseau@centralesupelec.fr, isabelle.barbet@lecnam.net, caroline.huron@cri-paris.org

RÉSUMÉ

Pour faciliter l'inclusion scolaire, il est indispensable de pouvoir adapter de manière automatique les manuels scolaires afin de les rendre accessibles aux enfants dyspraxiques. Dans ce contexte, nous proposons une tâche de classification des exercices selon leur type d'adaptation à la dyspraxie. Nous introduisons un corpus d'exercices extraits de manuels de français de niveau élémentaire, qui soulève certains défis de par sa petite taille et son contenu déséquilibré et bruité. Afin de tirer profit des modalités textuelles, structurelles et visuelles présentes dans nos données, nous combinons des modèles état de l'art par des stratégies de fusion précoce et tardive. Notre approche atteint une exactitude globale de 0.802. Toutefois, les expériences témoignent de la difficulté de la tâche, particulièrement pour les classes minoritaires, pour lesquelles l'exactitude tombe à 0.583.

ABSTRACT

Noisy and Unbalanced Multimodal Document Classification: Textbook Exercises as a Use Case

In order to foster inclusive education, automatic systems that can adapt textbooks to make them accessible to children with Developmental Coordination Disorder (DCD) are necessary. In this context, we propose a task to classify exercises according to their DCD's adaptation. We introduce an exercises dataset automatically extracted from French textbooks, with two major difficulties : a small size and an unbalanced and noisy data. To set a baseline on the dataset, we use state-of-the-art models combined through early and late fusion techniques to take advantage of text and vision/layout modalities. Our approach achieves an overall accuracy of 0.802. However, the experiments show the difficulty of the task, especially for minority classes, where the accuracy drops to 0.583.

MOTS-CLÉS : adaptation de manuels scolaires, classification multimodale, données bruitées, données déséquilibrées.

KEYWORDS: textbook adaptation, multimodal document classification, noisy data, unbalanced data.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 121–130.

1 Introduction et travaux connexes

La dyspraxie est un trouble développemental de la coordination affectant 5% des enfants, qui interfère avec la réussite scolaire et les activités de la vie quotidienne. Plus précisément, les enfants dyspraxiques n'automatisent pas l'écriture manuscrite et leurs troubles des mouvements oculaires peuvent les empêcher de lire un texte si sa présentation n'est pas adaptée. Ainsi, pour que les enfants dyspraxiques réussissent à l'école, les manuels utilisés en classe doivent tenir compte de leurs difficultés d'écriture et d'organisation du regard. Une adaptation au format numérique permet de contourner le déficit d'écriture manuscrite sans modifier le contenu des exercices et leur objectif pédagogique. La Figure 1 montre un exemple d'exercice de type « *choix multiple* » et son adaptation, permettant aux enfants de compléter la phrase en *cliquant* sur la bonne réponse. Des associations commencent à produire de tels manuels numériques adaptés, en effectuant toutes les transformations à la main. Malheureusement, étant donné la grande diversité des collections et le renouvellement des programmes d'enseignement, ces adaptations artisanales ne permettent pas de répondre aux besoins. D'autre part, les manuels scolaires sont peu explorés en Traitement Automatique du Langage Naturel (TALN), et la plupart des études existantes portent sur l'analyse du contenu linguistique (Green, 2019; Lucy *et al.*, 2020) ou la génération de questions (Ch & Saha, 2022; Ghosh, 2022), généralement dans des manuels de niveau universitaire. À notre connaissance, aucune ne traite des tâches de classification ou de formatage du contenu.

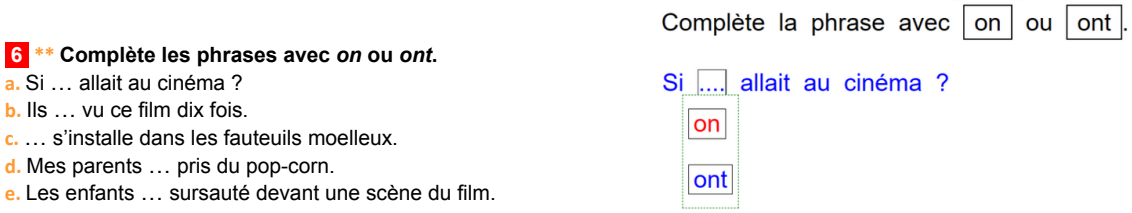


FIGURE 1 – Exercice à trous de type choix multiple et son adaptation.

Dans ce contexte, nous proposons une première étape vers l'automatisation de l'adaptation des manuels scolaires, avec la classification des exercices en fonction du type d'adaptation à la dyspraxie. Nous construisons un corpus d'exercices de manuels scolaires de français de niveau élémentaire annotés manuellement avec des étiquettes de type d'adaptation. Ce jeu de données reflète les difficultés de la tâche. Tout d'abord, le jeu de données est non seulement déséquilibré, certains types d'adaptation étant beaucoup plus fréquents que d'autres, mais aussi bruité, car il peut contenir des phrases agrammaticales ou incomplètes ainsi que des erreurs d'extraction. Ensuite, la classification concerne l'objectif pédagogique des exercices, qui peut être porté de manières très différentes. Enfin, les droits de propriété intellectuelle restreignent l'accès à un nombre limité de manuels scolaires, et conduisent par conséquent à un ensemble de données relativement petit. Notre approche de classification des exercices repose sur des stratégies de fusion précoce et tardive, afin de prendre en compte les informations sémantiques ainsi que structurelles et visuelles des documents.

Deux modèles de langue français basés sur RoBERTa (Liu *et al.*, 2019) – CamemBERT (Martin *et al.*, 2020), entraîné sur la partie française d'OSCAR (Suárez *et al.*, 2019) et FlauBERT (Le *et al.*, 2020), entraîné sur 24 corpus de styles variés collectés sur Internet – présentent des résultats similaires pour la classification. Cependant, ces modèles entraînés sur du texte sans erreur peuvent être impactés négativement par les erreurs présentes dans notre corpus, comme l'ont montré les études de Huang & Chen (2020) et Jiang *et al.* (2021). Alors que de nombreuses approches de classification reposent sur

le texte comme modalité unique, certaines études récentes se concentrent sur l'analyse de la mise en page. Ainsi, LayoutLM (Xu *et al.*, 2020) reprend l'architecture de BERT en ajoutant des plongements visuels et de position 2-D. Deux versions améliorées (Xu *et al.*, 2021b; Huang *et al.*, 2022) ont été récemment proposées, ainsi qu'une extension multilingue (Xu *et al.*, 2021a). BROS (Hong *et al.*, 2022) propose une méthode d'encodage spatial utilisant les positions relatives entre les blocs et DocFormer (Appalaraju *et al.*, 2021) introduit un mécanisme d'attention multimodal permettant le partage d'informations entre les modalités. Enfin, TILT (Powalski *et al.*, 2021) combine des caractéristiques convolutionnelles avec l'architecture T5 (Raffel *et al.*, 2020). Cependant, la plupart des modèles sont pré-entraînés et fine-tunés sur des documents monolingues, généralement en anglais. Pour dépasser cette limitation, LiLT (Wang *et al.*, 2022) permet d'imbriquer n'importe quel modèle pré-entraîné de type RoBERTa dans un module de mise en page. Par ailleurs, ces modèles sont entraînés sur des pages entières correctement formatées. Bien qu'ils puissent être efficaces pour la compréhension de pages de manuels, nous cherchons ici à catégoriser des exercices : des documents courts et très semblables.

Dans cet article, nos principales contributions sont : (i) l'introduction d'une nouvelle tâche de classification, pour l'adaptation automatique des exercices de manuels scolaires pour les enfants dyspraxiques ; (ii) un cadre de classification multimodale pour la tâche de classification des exercices de manuels scolaires ; (iii) des expériences avec différentes architectures multimodales, y compris le modèle LiLT (Wang *et al.*, 2022) récemment proposé.

2 Préparation du corpus

Le corpus est construit à partir de 3 manuels scolaires de français - étude de la langue, de niveau élémentaire (CE1 et CE2), au format PDF. Dans un premier temps, chaque manuel est converti en un document XML au format ALTO par pdfalto [1] couplé à MuPDF [2]. Cette combinaison d'outils OpenSource permet d'extraire les mots avec leur style de police et leurs coordonnées spatiales, ainsi que les images, dans une structure XML. Les mots extraits sont tokénisés et regroupés en segments grâce à des règles sur la taille et le style des polices, les types de caractères (chiffres, symboles, signes de ponctuation) et l'espacement entre les tokens ou les caractères. Une interface d'annotation conçue spécifiquement permet ensuite de réorganiser les segments dans une structure de manuel scolaire. Les segments sont étiquetés en rôles (*titre de chapitre*, *leçon*, *numéro d'exercice*, *consigne*, etc.) de manière semi-automatique sur la base de leur police dominante, puis les blocs d'activité sont reconstitués en utilisant des caractéristiques géométriques et une logique dans l'enchaînement. Par exemple, un numéro d'exercice précédé par un énoncé indique le début d'un nouvel exercice. Cette étape d'extraction aboutit à 2748 exercices, divisés en plusieurs parties : un numéro ou un nom, toujours une consigne, souvent un énoncé, et parfois aussi un exemple ou un conseil.

Toutefois, la complexité de la mise en page des manuels rend la tâche d'extraction difficile : les pages peuvent comporter des tableaux, des listes, des illustrations ou des blocs de texte épars, qui peuvent introduire du bruit dans les données. D'autre part, les notions de *phrase* et *token* ont été étendues en adéquation avec la nature du corpus. Si la plupart des consignes sont grammaticalement et sémantiquement correctes, les énoncés peuvent contenir des mots ou phrases à trous (« c...bat », « Manon a perdu ... chat. »), des choix de type choix multiple (« (son/sont) »), des suites de mots

1. https://github.com/kermitt2/pdfalto
2. https://github.com/ArtifexSoftware/mupdf

concaténés (« cirageâgégéantenfantfantômetomate »), des portions de phrases dispersées (« est une fleur », « la tulipe »), des numéros de liste (« a. », « b. »), etc. Par conséquent, il est plus approprié de faire référence à des segments de texte plutôt qu'à des phrases. De plus, un quart des segments est composé de 1 à 5 *tokens*, tandis que les segments les plus longs comptent jusqu'à 65 *tokens*. Avec 1 à 10 segments et 5 à 91 *tokens* par exercice, la longueur des documents est tout aussi variable mais reste courte par rapport aux jeux de données de référence.

Les exercices extraits sont manuellement annotés par 2 experts de la dyspraxie en 33 catégories. Ces catégories correspondent aux types d'adaptation à la dyspraxie et reflètent l'objectif pédagogique de l'exercice ainsi que le processus d'interaction impliqué dans sa résolution. Par exemple, la Figure 1 illustre un exercice de la classe *choix multiple*, et la Figure 2 en annexe contient 5 exercices d'autres classes. Résultent de cette annotation 2567 exercices qui n'ont qu'une seule étiquette, 146 exercices appartenant à plusieurs classes et 36 exercices à retirer des manuels adaptés, car ils demandent une compétence directement perturbée par le handicap. Pour ce travail, nous ne traitons que les exercices avec une seule étiquette et excluons la classe la moins représentée qui ne contient qu'un seul exercice. Ainsi, notre jeu de données final est composé de 2566 exercices étiquetés avec 32 catégories. Le jeu de données est très déséquilibré : 2 classes sur 32 comptent plus de 300 exercices, tandis que 11 classes en comptent moins de 20. Les 21 classes les plus fournies représentent 95% du corpus.

Le corpus est divisé en 3 sous-ensembles : apprentissage (70%), validation (10%) et test (20%). La proportion des classes d'exercices par manuel et le ratio entre les classes sont conservés. Les numéros et les noms des exercices sont supprimés. Si un exercice contient un exemple ou un conseil, ceux-ci sont concaténés avec la consigne. Le texte est normalisé en minuscules et tous les caractères d'espacement sont réduits à un espace.

3 Méthodologie

Notre objectif consiste à catégoriser les exercices extraits en fonction de leur type d'adaptation à la dyspraxie. Dans un premier temps, nous utilisons CamemBERT. Afin d'adapter le modèle au domaine scolaire, son modèle de langue masqué est fine-tuné sur les textes suivants : les leçons et exercices tirés de 4 manuels de français (les 3 manuels utilisés pour construire notre corpus, hormis les exercices présents dans les sous-corpus de validation et de test, et un 4ème manuel non annoté) ; 1293 *Fantastiques Exercices* de l'association *Le Cartable Fantastique* [3], qui fournit une collection d'exercices accompagnés de leur version interactive adaptée aux enfants dyspraxiques ; les textes de lecture originaux d'Alector (Gala *et al.*, 2020), corpus parallèle de 79 textes de lecture simplifiés au niveau lexical, syntaxique et discursif. Pour la phase d'apprentissage, la consigne et l'énoncé sont concaténés (séparés par le token spécial `<sep>`) avant d'être donnés au modèle. En vue d'exploiter d'autres modalités, nous utilisons LayoutLMv2 (Xu *et al.*, 2021b), qui prend en entrée des plongements textuels, positionnels et visuels. LayoutLMv2 est pré-entraîné sur IIT-CDIP (Lewis *et al.*, 2006) et fine-tuné pour la classification sur son sous-ensemble RVL-CDIP (Harley *et al.*, 2015), composé d'images de documents scannés tels que des lettres ou des formulaires. La plupart des documents utilisent l'anglais comme langue principale, mais IIT-CDIP contient quelques documents dans d'autres langues, dont le français. Si LayoutLMv2 traite des plongements de 3 modalités différentes, il est conçu pour du texte en anglais et peut ne pas bénéficier pleinement des caractéristiques textuelles de notre corpus.

3. https://www.cartablefantastique.fr/

Les 32 classes définies présentant une grande variabilité tant sémantique que structurelle, comme le montrent les exemples de la Figure annexe 2, nous pensons que le texte en français ainsi que la mise en page et l'image sont pertinents pour notre objectif de classification. Des approches de fusion sont mises en œuvre afin d'exploiter chacune de ces modalités. La première solution consiste à appliquer une fusion tardive au niveau des scores des classifieurs CamemBERT et LayoutLMv2. Les scores sont normalisés entre 0 et 1 avec la normalisation $Min - Max$, qui préserve les relations entre les valeurs d'origine, puis fusionnés avec les stratégies de fusion classiques *Moyenne* et *Maximum*. Dans la deuxième solution, nous tirons parti du modèle LiLT (Wang *et al.*, 2022), qui permet d'associer n'importe quel modèle pré-entraîné de type RoBERTa avec un module pré-entraîné sur la structure. Ce module, pré-entraîné sur IIT-CDIP, est combiné à notre version de CamemBERT fine-tunée sur des manuels scolaires et des textes de lecture, afin d'obtenir un modèle de type LayoutLM pour le français scolaire. Finalement, nous appliquons un vote majoritaire sur les prédictions de CamemBERT, LayoutLMv2 et LiLT[CamemBERT], LiLT[CamemBERT] étant le classifieur par défaut.

Enfin, pour faire face au déséquilibre des données, nous envisageons la configuration de la fonction de perte et les méthodes d'échantillonage, dites de *sampling*. Pour les tâches de classification sur des données équilibrées ou déséquilibrées, l'entropie croisée est la fonction de perte la plus largement utilisée. Dans un contexte de classification binaire de données déséquilibrées, la fonction de perte focale (Lin *et al.*, 2017) a été introduite comme une amélioration de l'entropie croisée classique, où les exemples faciles sont dynamiquement sous-pondérés. La perte focale a ensuite été étendue aux problèmes multi-classes et a montré des résultats prometteurs. L'augmentation du paramètre de focalisation γ permet de contrôler le poids des exemples faciles et de focaliser l'attention sur les exemples mal classés. Nous avons expérimenté cette fonction avec différentes valeurs pour γ, allant de 0 (ce qui correspond à l'entropie croisée pondérée) à 5. En ce qui concerne les approches de sampling et compte tenu de la taille de notre corpus, l'*undersampling* entraînerait une perte d'informations. Cependant, nous nous sommes inspirés des techniques d'apprentissage ensembliste pour construire plusieurs sous-ensembles sous-échantillonnés en y répartissant les exercices des classes majoritaires, puis en fusionnant les sorties des différents modèles formés sur ces sous-ensembles sous-échantillonnés.

Nos modèles utilisent l'architecture BASE et la longueur des séquences d'entrée est fixée à 256. Pour les expériences finales, la taille du batch est paramétrée à 16, le nombre d'époques entre 30 et 40 et le taux d'apprentissage initial entre 1e-5 et 1e-4. Nous utilisons l'optimiseur Adam et la fonction de perte d'entropie croisée pondérée par les effectifs des classes. Les résultats sur le corpus de test sont obtenus avec les modèles fine-tunés qui donnent les meilleurs résultats sur le corpus de validation.

4 Résultats et discussion

Le Tableau 1 présente les scores d'exactitude et de macro F-mesure pour la classification des exercices. La meilleure performance est atteinte par LiLT combiné à CamemBERT suivi d'une fusion tardive des 3 modèles : l'exactitude est alors de 0.802, ce qui indique que les 3 modèles sont complémentaires. Pour la majorité des classes, LayoutLMv2 est presque aussi performant que CamemBERT, bien qu'il capture moins bien les informations sémantiques qu'un modèle français. Cela met en évidence la pertinence des modalités structurelles et visuelles. Ce n'est toutefois pas suffisant, les performances de LayoutLM étant très faibles pour les classes sous-représentées. L'augmentation des scores avec les stratégies de fusion tardive et précoce est statistiquement significative par rapport à LayoutLMv2.

Modèle	Exactitude			Macro F1
	Total	Maj.	Min.	Total
Baseline Classe Majoritaire	0.147			0.008
CamemBERT^{T}	0.775	0.788	0.500	0.663
LayoutLMv2^{T+L+I}	0.708	0.722	0.250	0.487
CamemBERT + LayoutLMv2 (Fusion Maximum)	0.767	0.784	0.417	0.627
CamemBERT + LayoutLMv2 (Fusion Moyenne)	0.782	0.796	0.500	0.664
LiLT^{T+L}[CamemBERT]	0.786	0.796	**0.583**	0.696
CamemBERT + LayoutLMv2 + LiLT[CamemBERT]	**0.802**	**0.813**	**0.583**	**0.714**

TABLE 1 – Résultats sur l'ensemble des données de test (Total), les 21 classes majoritaires (Maj.) et les 11 classes minoritaires (Min.). Pour chaque modèle, on rappelle s'il a été pré-entraîné sur le texte (T), la mise en page (L) ou l'image (I).

	$\mathbf{B^-L^-}$	$\mathbf{B^-L^+}$	$\mathbf{B^+L^-}$	$\mathbf{B^+L^+}$
# exercices	79	74	36	321
LiLT	16 (20%)	61 (82%)	20 (56%)	304 (95%)
Fusion Maximum	0	54 (73%)	16 (44%)	321 (100%)
Fusion Moyenne	0	59 (80%)	19 (53%)	321 (100%)

TABLE 2 – Comparaison des classifications pour les 3 stratégies de fusion sur les exercices correctement ($^+$) et incorrectement ($^-$) classifiés par CamemBERT (B) et LayoutLMv2 (L).

Par rapport à CamemBERT, la fusion via LiLT et la fusion tardive *Moyenne* améliorent légèrement l'exactitude globale. Si cette amélioration ne semble pas significative, les scores sur les classes minoritaires révèlent un écart plus important entre les classifieurs CamemBERT et LiLT[CamemBERT]. Les méthodes de fusion surpassent les modèles individuels, soulignant ainsi l'importance des trois modalités pour le traitement des données issues de manuels scolaires. La combinaison du français scolaire et de la mise en page est particulièrement efficace avec LiLT. En outre, des expériences complémentaires sur CamemBERT confirment l'impact positif du fine-tuning du modèle de langue masqué sur un corpus scolaire, l'exactitude augmentant de 0,747 à 0,775.

Cependant, l'application de la perte focale pour faire face au déséquilibre des données s'avère inefficace. Selon le paramétrage de γ, elle conduit à des scores égaux ou inférieurs à ceux obtenus avec l'entropie croisée pondérée. Par ailleurs, le sous-échantillonnage d'un très petit ensemble de données n'est pas efficace car des informations sont perdues. Au mieux, nous obtenons une précision de 0,730 en utilisant des sous-ensembles sous-échantillonnés.

Enfin, des expériences supplémentaires ont été menées pour évaluer la généralisation des modèles intra- et inter-collection de manuels scolaires. De nouveaux modèles ont été entraînés sur des exercices issus de deux manuels de même collection, tandis que le troisième manuel, d'une collection différente, a été utilisé à des fins d'évaluation uniquement. Les résultats révèlent que la capacité de généralisation des modèles est plus élevée à partir des caractéristiques textuelles que positionnelles. En effet, LayoutLMv2 ne parvient pas à généraliser efficacement entre les collections et requiert une quantité de données plus importante que CamemBERT pour obtenir des résultats satisfaisants. La fusion

précoce avec LiLT continue de surpasser les approches à modèle unique et démontre de bonnes capacités de généralisation sur des collections distinctes.

Le Tableau 2 permet la comparaison des méthodes de fusion. Les 3 méthodes peuvent améliorer les prédictions. Bien que LiLT ne saisisse pas la totalité des exercices correctement prédits par CamemBERT et LayoutLMv2 individuellement, il corrige 20% des exercices mal classés par les deux classifieurs. Les scores obtenus avec les stratégies de fusion tardive démontrent que CamemBERT est davantage confiant [4] et fiable que LayoutLMv2. En revanche, LayoutLMv2 gère mieux les exercices où la mise en page prévaut sur le contenu sémantique. Il catégorise correctement 36 exercices que CamemBERT n'a pas su catégoriser, et environ la moitié de ces prédictions sont conservées par les stratégies de fusion *Moyenne* et *Maximum*. Enfin, les classes minoritaires sont les plus difficiles à traiter. Les scores détaillés au niveau des classes soulignent les difficultés posées par le déséquilibre des données. Cela reste un problème délicat qui, pour notre objectif de classification, nécessite une augmentation des données. Cette tâche se révèle complexe en raison de la quantité de données et de la spécificité du langage des manuels, en particulier dans les consignes.

5 Conclusion

Pour favoriser l'inclusion scolaire et avec un objectif à long terme d'adapter automatiquement des manuels scolaires complets pour les rendre accessibles aux enfants dyspraxiques, nous avons introduit dans cet article une nouvelle tâche : la classification des exercices de manuels scolaires en fonction de leur type d'adaptation. Nous avons mené une étude comparative de méthodes de classification neuronales sur notre propre jeu de données composé de 2566 exercices de manuels de français.

Nous avons proposé différentes approches basées sur trois modèles pré-entraînés ayant fait leurs preuves sur de nombreuses tâches de TALN et des corpus de référence. Nous avons cherché à exploiter différentes modalités et avons finalement obtenu un score d'exactitude de 0,802 en utilisant des méthodes de fusion. Les expériences ont démontré l'importance de la mise en page et de l'image en plus du texte dans la compréhension des manuels.

Afin d'améliorer ces résultats prometteurs, nos travaux futurs se concentreront sur l'étape d'extraction et le nettoyage des données. Par ailleurs, si l'exactitude globale est encourageante, les résultats pour les classes minoritaires doivent encore être améliorés. Nous prévoyons de générer artificiellement des exercices pour résoudre les problèmes de petite taille et de déséquilibre entre les classes.

Remerciements

Les auteurs remercient les relecteurs anonymes pour leurs commentaires constructifs, ainsi que Guillaume Faure pour le développement de la plateforme d'extraction et d'annotation. Ce travail a été réalisé dans le cadre du projet MALIN (MAnuels scoLaires INclusifs) sous le financement ANR-21-CE38-0014. Il a bénéficié d'un accès au cluster de calcul Lab-IA.

4. La différence entre le score le plus élevé et le suivant est plus importante.

Références

APPALARAJU S., JASANI B., KOTA B. U., XIE Y. & MANMATHA R. (2021). Docformer : End-to-end transformer for document understanding. In *Proceedings of the 18th IEEE International Conference on Computer Vision*.

CH D. R. & SAHA S. K. (2022). Generation of multiple-choice questions from textbook contents of school-level subjects. *IEEE Transactions on Learning Technologies*.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

GALA N., TACK A., JAVOUREY-DREVET L., FRANÇOIS T. & ZIEGLER J. C. (2020). Alector : A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the 12th Language Resources and Evaluation for Language Technologies*.

GHOSH K. (2022). Remediating textbook deficiencies by leveraging community question answers. *Education and Information Technologies*.

GREEN C. (2019). A multilevel description of textbook linguistic complexity across disciplines : Leveraging NLP to support disciplinary literacy. *Linguistics and Education*.

HARLEY A. W., UFKES A. & DERPANIS K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the 13th International Conference on Document Analysis and Recognition*.

HONG T., KIM D., JI M., HWANG W., NAM D. & PARK S. (2022). Bros : A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

HUANG C.-W. & CHEN Y.-N. (2020). Learning ASR-robust contextualized embeddings for spoken language understanding. In *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing*.

HUANG Y., LV T., CUI L., LU Y. & WEI F. (2022). LayoutLMv3 : Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*.

JIANG M., HU Y., WORTHEY G., DUBNICEK R. C., UNDERWOOD T. & DOWNIE J. S. (2021). Impact of OCR quality on BERT embeddings in the domain classification of book excerpts. In *Proceedings of the Conference on Computational Humanities Research*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised language model pre-training for french. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

LEWIS D., AGAM G., ARGAMON S., FRIEDER O., GROSSMAN D. & HEARD J. (2006). Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*.

LIN T.-Y., GOYAL P., GIRSHICK R., HE K. & DOLLÁR P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). RoBERTa : A robustly optimized BERT pretraining approach. *arXiv preprint arXiv :1907.11692*.

LUCY L., DEMSZKY D., BROMLEY P. & JURAFSKY D. (2020). Content analysis of textbooks via natural language processing : Findings on gender, race, and ethnicity in texas US history textbooks. *AERA Open*.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE , SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

POWALSKI R., BORCHMANN L., JURKIEWICZ D., DWOJAK T., PIETRUSZKA M. & PALKA G. (2021). Going full-TILT boogie on document understanding with text-image-layout transformer. In *Proceedings of 16th International Conference on Document Analysis and Recognition*.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*.

SUÁREZ P. J. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora*.

WANG J., JIN L. & DING K. (2022). LiLT : A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

XU Y., LI M., CUI L., HUANG S., WEI F. & ZHOU M. (2020). LayoutLM : Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

XU Y., LV T., CUI L., WANG G., LU Y., FLORENCIO D., ZHANG C. & WEI F. (2021a). LayoutXLM : Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv :2104.08836*.

XU Y., XU Y., LV T., CUI L., WEI F., WANG G., LU Y., FLORENCIO D., ZHANG C., CHE W., ZHANG M. & ZHOU L. (2021b). LayoutLMv2 : Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Annexes

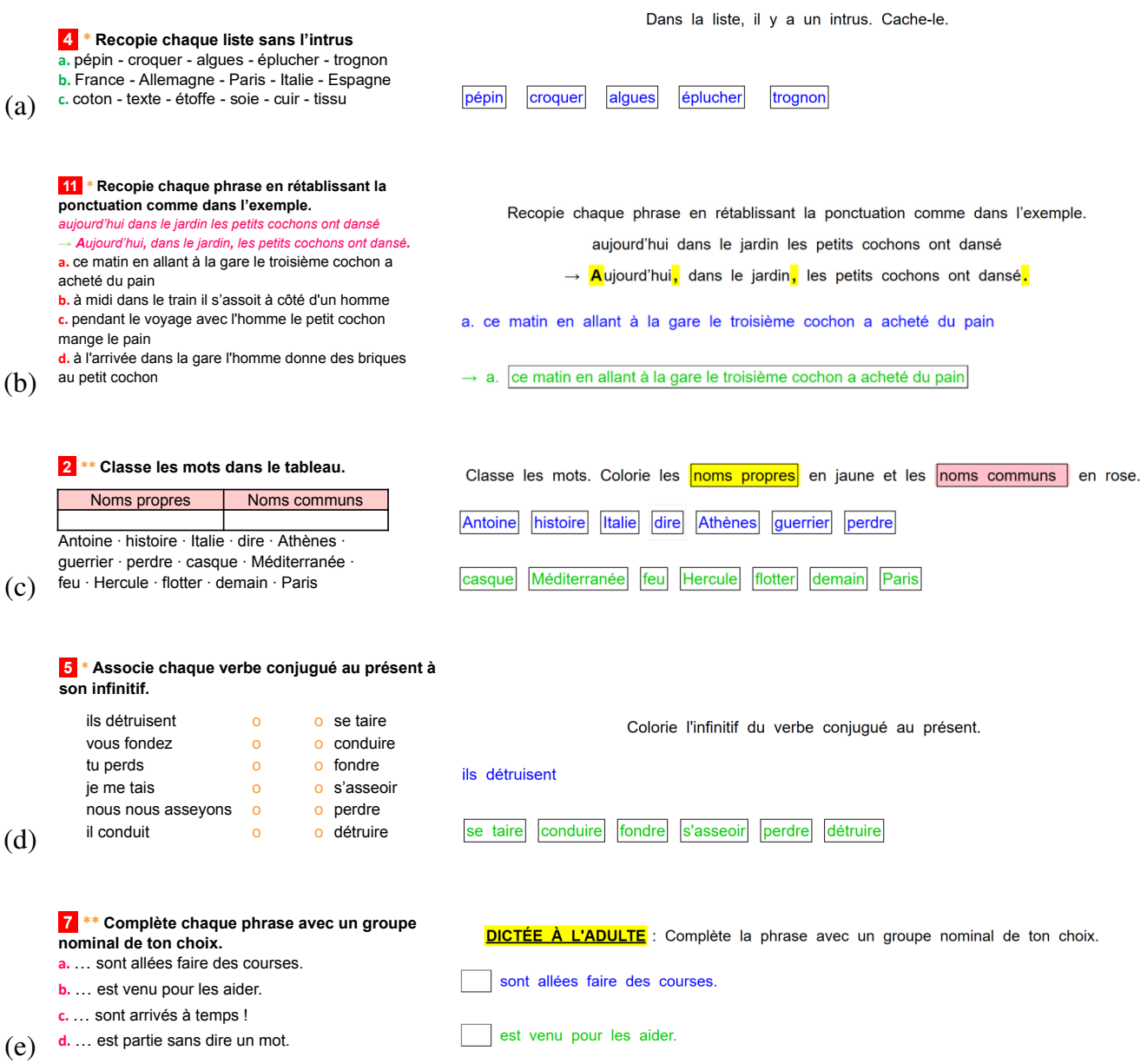


FIGURE 2 – Exemples d'exercices et exercices adaptés pour les classes (a) CacheIntrus, (b) EditePhrase, (c) Classe, (d) Associe, (e) RemplirAuClavier

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Détection de faux tickets de caisse à l'aide d'entités et de relations basées sur une ontologie de domaine

Beatriz Martínez Tornés[1] Emanuela Boros[1] Petra Gomez-Krämer[1]
Antoine Doucet[1] Jean-Marc Ogier[1]
(1) La Rochelle Université, L3i, F-17000, La Rochelle, France
{prénom.nom}@univ-lr.fr

RÉSUMÉ

Dans cet article, nous nous attaquons à la tâche de détection de fraude de documents. Nous considérons que cette tâche peut être abordée avec des techniques de traitement automatique du langage naturel (TALN). Nous utilisons une approche basée sur la régression, en tirant parti d'un modèle de langage pré-entraîné afin de représenter le contenu textuel, et en enrichissant la représentation avec des entités et des relations basées sur une ontologie spécifique au domaine. Nous émulons une approche basée sur les entités en comparant différents types d'entrée : texte brut, entités extraites et une reformulation du contenu du document basée sur des triplets. Pour notre configuration expérimentale, nous utilisons le seul ensemble de données librement disponible de faux tickets de caisse, et nous fournissons une analyse approfondie de nos résultats. Ils montrent des corrélations intéressantes entre les types de relations ontologiques, les types d'entités (produit, entreprise, etc.) et la performance d'un modèle de langage basé sur la régression qui pourrait aider à étudier le transfert d'apprentissage à partir de méthodes de traitement du langage naturel (TALN) pour améliorer la performance des systèmes de détection de fraude existants.

ABSTRACT

Detecting Forged Receipts with Domain-specific Ontologies

In this paper, we tackle the task of document fraud detection. We consider that this task can be addressed with natural language processing techniques. We treat it as a regression-based approach, by taking advantage of a pre-trained language model in order to represent the textual content, and by enriching the representation with domain-specific ontology-based entities and relations. We emulate an entity-based approach by comparing different types of input : raw text, extracted entities and a triple-based reformulation of the document content. For our experimental setup, we utilize the single freely available dataset of forged receipts, and we provide a deep analysis of our results in regard to the efficiency of our methods. Our findings show interesting correlations between the types of ontology relations, types of entities (product, company, etc.) and the performance of a regression-based language model that could help to study the transfer learning from natural language processing (NLP) methods to boost the performance of existing fraud detection systems.

MOTS-CLÉS : Détection de fraude de documents, modèle de langue.

KEYWORDS: Document fraud detection, language model.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 131–139.

1 Introduction

La falsification de documents est un problème très répandu, alors que la numérisation des documents permet un échange plus facile pour les entreprises et les administrations. Si l'on ajoute à cela la disponibilité de logiciels de traitement d'images et d'édition de documents ainsi que de scanners et d'imprimantes peu coûteux, les documents courent de nombreux risques d'être altérés ou contrefaits (Gomez-Kämer, 2021). La contrefaçon est la production d'un document authentique par imitation et le faux est l'altération d'un ou plusieurs éléments d'un document authentique. L'un des principaux défis de la détection de la fraude de documents est le manque de données annotées librement disponibles, car de nombreuses études autour de la fraude ne tiennent pas compte des documents réels et se concentrent sur les transactions (comme la fraude à la carte de crédit, la fraude à l'assurance ou même la fraude financière) (Behera & Panigrahi, 2015; Kowshalya & Nandhini, 2018; Rizki *et al.*, 2017).

La collecte de vrais faux documents est également difficile, car les vrais fraudeurs ne partageraient pas leur travail, et les entreprises ou administrations sont réticentes à révéler leurs failles de sécurité et ne peuvent pas partager des informations sensibles (Sidere *et al.*, 2017; Mishra & Ghorpade, 2018; Vidros *et al.*, 2017). De plus, le défi de travailler avec un corpus de documents administratifs potentiellement frauduleux est la rareté de la fraude ainsi que l'expertise humaine nécessaire pour repérer les documents frauduleux (Benchaji *et al.*, 2018; Li *et al.*, 2016; Carta *et al.*, 2019). S'intéresser aux documents réels réellement échangés par les entreprises ou les administrations est important pour que les méthodes de détection de fraude développées soient utilisables dans des contextes réels et pour que la cohérence des documents authentiques soit assurée. Cependant, ce type de document administratif contient des informations privées sensibles et n'est généralement pas mis à la disposition de la recherche (Behera & Panigrahi, 2015).

Ensuite, la plupart des recherches en matière de fraude de documents se concentrent sur l'analyse d'images de documents, car la plupart d'entre eux sont numérisés et échangés sous forme d'images par les entreprises et les administrations. La détection de falsification de documents est donc souvent définie comme une tâche de vision par ordinateur (Bertrand *et al.*, 2015; Cozzolino & Verdoliva, 2018; Fridrich & Kodovsky, 2012; Cozzolino *et al.*, 2014). L'image d'un document peut être modifiée de différentes manières à l'aide d'un logiciel d'édition d'images. La modification peut se faire dans le document numérique original ou dans la version imprimée et numérisée du document (Cruz *et al.*, 2018; James *et al.*, 2020). À cet égard, la compétition (Artaud *et al.*, 2018) a été, à notre connaissance, la seule tentative visant à encourager l'utilisation de méthodes de vision par ordinateur et de TALN pour la détection de documents falsifiés, en fournissant un corpus parallèle (image/texte) de tickets falsifiés librement accessibles. Toutefois, le nombre de participants était faible (cinq soumissions) et seule l'une d'entre elles intégrait des caractéristiques de contenu textuel sous la forme de modules de vérification basés sur des règles (l'examen des incohérences dans les prix des articles et le total à payer).

Nous considérons donc que le TALN pourrait être utilisé pour améliorer les performances de la détection des documents frauduleux en traitant les incohérences du faux lui-même (Artaud *et al.*, 2018). Ainsi, alors que les méthodes de vision par ordinateur s'appuient sur la recherche d'imperfections, soit en visant à détecter les irrégularités qui auraient pu se produire au cours du processus de modification (Bertrand *et al.*, 2013), soit en se concentrant sur l'identification de l'imprimante, afin de vérifier si le document a été imprimé par l'imprimante d'origine (Elkasrawi & Shafait, 2014; Mikkilineni *et al.*, 2005), les méthodes NLP pourraient combler le fossé entre les incohérences de l'image et du texte.

2 Notre approche

Nous basons notre modèle de détection des fraudes sur le modèle pré-entraîné CamemBERT (Martin *et al.*, 2020) qui est un modèle linguistique pré-entraîné de pointe pour le français basé sur le modèle RoBERTa (Liu *et al.*, 2019).

CamemBERT (Martin *et al.*, 2020) est une pile de couches Transformer (Vaswani *et al.*, 2017), où un bloc Transformer (encodeur) est une architecture d'apprentissage profond basée sur des mécanismes d'attention multitêtes avec des encastrements de position sinusoïdale. Comme indiqué précédemment, nous traitons la tâche de détection des fraudes comme une tâche de régression et un score numérique $s_x \in [0, 1]$ est donc attribué à l'exemple d'entrée $\{x_i\}_{i=1}^{l}$ pour quantifier son niveau de falsification, qui est défini comme $s_x = \sigma(f(\{x_i\})$ où σ est la fonction sigmoïde $\sigma(z) = \frac{1}{1+e^{-z}}$ qui renvoie un score numérique $s_x \in [0, 1]$. Enfin, les valeurs prédites sont seuillées à 0,5.

2.1 Jeu de données de faux tickets de caisse

Le jeu de données librement disponible Find it ! [1] (Artaud *et al.*, 2017, 2018) que nous utilisons est composé de 998 images de tickets français et de leurs transcriptions associées. Ces transcriptions sont issues d'une reconnaissance optique de caractères et ont été corrigées manuellement de manière participative par les créateurs de Find it !. Ce jeu de données a été collecté pour fournir un corpus parallèle image/texte et un point de référence pour évaluer les méthodes de détection de fraude basées sur le texte. Les faux tickets sont le résultat d'ateliers de fraude, au cours desquels les participants ont reçu un ordinateur standard avec plusieurs logiciels d'édition d'images pour modifier manuellement les images et les transcriptions associées des tickets. Ainsi, le jeu de données contient des faux réalistes, correspondant à des situations réelles telles que des demandes de remboursement frauduleuses effectuées en modifiant le prix d'un article (Figure 1 (b)), son nom, le moyen de paiement, etc. La falsification peut également viser une extension indue de la garantie en modifiant la date. D'autres falsifications peuvent impliquer l'entreprise émettrice dans le but de blanchir de l'argent. Les tickets ont été collectés localement dans le laboratoire de recherche où ils ont été développés, ce qui se traduit par une fréquence élevée de magasins à proximité. Bien que cela puisse être considéré comme un biais, nous estimons que cela reste proche d'un cas d'application réel, dans lequel une entreprise stocke les documents/factures qu'elle émet. Le jeu de données de 998 documents est divisé en 498 documents pour l'entraînement et 500 pour le test, chacun comportant 30 faux documents. Ainsi, les données sont déséquilibrées, selon une distribution réaliste. En effet, il y a typiquement moins de 5% de documents falsifiés dans les flux de documents, une distribution similaire aux valeurs aberrantes (Artaud *et al.*, 2018; Nadim *et al.*, 2019).

2.2 Prétraitements et choix de l'entrée

Afin de mieux explorer la nature spécifique semi-structurée des tickets, nous avons expérimenté avec quatre types d'entrées (présentées avec plus de détail dans la version longue de l'article (Martínez Tornés *et al.*, 2023)) :

1. **Texte** : le texte brut d'un ticket sans aucun prétraitement ;

1. http://findit.univ-lr.fr/download-the-dataset/

2. **Entités** : nous extrayons les entités présentes sur la base d'une ontologie de ticket de caisse et les concaténons avec un séparateur d'espace (par exemple« Carrefour ») comme décrit ci-dessous ;
3. **Texte + Entités** : nous enrichissons le texte du ticket en introduisant des marqueurs spéciaux pour chaque type d'entité (tel que, entreprise, produit, etc.) et remplaçons chaque entité dans le texte par son libellé entouré des marqueurs de son type d'entité (Boros *et al.*, 2021) ;
4. **Triplets** : basés sur la même ontologie, mais en extrayant également les relations sémantiques.

L'ontologie a été peuplée automatiquement avec des expressions régulières créées manuellement et basées sur les régularités des tickets de caisse. Par exemple, les produits et leurs prix ont été extraits des lignes du document terminées par le symbole « € », ou l'utilisant comme séparateur décimal, à l'exclusion des lignes qui rapportent le total ou le paiement. Le processus d'extraction a été exécuté comme une machine à états finis pour s'adapter à des structures plus variées, telles que des prix et des produits non alignés. L'ontologie a été peuplée dynamiquement à l'aide de la bibliothèque Python Owlready2 [2]. L'ontologie (Artaud, 2019) comporte des classes décrivant les informations présentes dans les tickets, telles qu'Entreprise, Adresse, Produit, etc. ainsi que des propriétés d'objet (relations entre ces classes) telles que a_adresse, contient. Nous notons que le ticket est une entité en soi, représentée par l'étiquette de son ID (une valeur numérique). L'ontologie définit également des propriétés de données, qui associent une instance à une valeur, comme *a_date*, *a_heure*, *a_montant_total*, *a_prix_total*, *a_nombre_darticles*, etc. [3]

Extraction d'entités Nous considérons qu'une entité extraite correspond à une instance d'une classe définie dans l'ontologie, qui définit son type. Nous avons annoté chaque entité fraudée, c'est-à-dire chaque entité ayant été modifiée pendant la constitution du dataset (altération, suppression ou ajout). Les modifications ne sont pas comptabilisées en elles-mêmes, seules les entités modifiées le sont : par exemple, une date « 11/02/2017 » altérée en « 10/02/2016 » compte pour une entité modifiée, même si elle a subi deux modifications graphiques. Le nombre d'entités modifiées est présenté dans le tableau 1 (a).

Type d'entité	Train	Test
Entreprise	7	5
Produit	33	44
Paiement	30	37
Métadonnée	18	17
Total	88	103

FIGURE 1 – **(a)** Distribution des entités modifiés. **(b)** Exemple de fraude sur le prix sur le ticket de droite.

La plupart des entités modifiées impliquent des montants d'argent (entités de produit et de paiement), même si ceux-ci ne sont pas toujours modifiés de manière cohérente : la Figure 1 (b) présente un

2. https://owlready2.readthedocs.io/en/v0.37/
3. Pour plus de détails et d'exemples, (Martínez Tornés *et al.*, 2023)

ticket comportant trois fraudes (le prix du produit, le total à payer et le montant payé) qui ne comporte pas d'incohérence numérique.

Extraction des triplets Afin d'aller au-delà des entités extraites et de fournir plus d'informations sur les relations entre les entités, nous avons choisi de les incorporer. Notre objectif était de mettre en évidence la structure sous-jacente des documents en énonçant explicitement les relations entre les entités. Nous avons veillé à supprimer les relations inverses, par exemple *a_fax* et *est_fax_de* en ne conservant qu'une seule de chaque paire. Nous avons également inclus les relations attributives, c'est-à-dire les propriétés des données, qui associent une entité à une valeur (numérique, date ou heure). Nous nous sommes appuyés sur ces triplets pour normaliser le contenu des tickets, en proposant une entrée (4) composée des triplets extraits en remplacement du texte extrait des tickets.

3 Expériences

Nous comparons notre modèle à deux méthodes *baseline*. Tout d'abord, nous considérons un *vérificateur d'incohérence numérique*, en simulant manuellement un vérificateur qui ne prend en compte que les incohérences numériques simples, sans s'appuyer sur des connaissances externes. Nous considérons comme une incohérence numérique simple tout écart entre le total et la somme des prix, entre le total et le total payé, ou entre la quantité, le prix unitaire et le prix du produit. Deuxièmement, nous considérons un classifieur de régression par machine à vecteur de support (SVM) avec des hyperparamètres par défaut comme notre modèle de base appliqué à la représentation fréquence des termes-fréquence inverse des documents (TF-IDF) des unigrammes et des bigrammes extraits des tickets en minuscules.

Nous comparons également nos résultats à deux approches image existantes, proposées dans la compétition Find it ! (Artaud *et al.*, 2018). L'architecture Verdoliva (Cozzolino & Verdoliva, 2020, 2018) est également basée sur un SVM et combine trois approches différentes : un module de détection de fraude par copier-coller, basé sur Cozzolino *et al.* (2015), un module d'extraction (et de comparaison) de signatures de caméra (Cozzolino & Verdoliva, 2020, 2018), et un module de détection de faux basée sur les caractéristiques locales de l'image, proposée à l'origine comme méthode de stéganalyse (Cozzolino *et al.*, 2014). Nous présentons également les résultats proposés par Fabre (Artaud *et al.*, 2018), qui s'appuient sur un modèle pré-entraîné Resnet152 (He *et al.*, 2015) pour la classification.

Résultats Le tableau 1 détaille les résultats de la classification binaire entre les classes « Fraudé » et « Authentique ». La classification étant très déséquilibrée, nous ne présentons que les résultats pour la classe « Fraudé ». Nous remarquons que les méthodes utilisant les *Triplets* comme entrée sont plus performantes que les autres, même dans leur représentation TF-IDF, le rappel est égal à un, ce qui signifie que tous les tickets falsifiés sont retrouvés avec succès.

Dans le cas des *Triplets*, nous n'avons observé que deux vrais tickets mal étiquetés. Pour l'un d'entre eux, le prix total est plutôt flou dans l'image, il a donc été corrigé manuellement dans la transcription avec « , » au lieu de « . » comme séparateur décimal. L'autre ticket authentique mal étiqueté présente un montant total élevé en comparaison avec le reste des tickets (plus de 87 euros). Ces irrégularités pourraient expliquer ces deux erreurs. En ce qui concerne la comparaison avec le

Méthode	P	R	F1
Vérificateur d'incohérence numérique	100.0	46.67	63.34
Approches Image			
Fabre (Artaud *et al.*, 2018)	36.4	93.3	52.3
Verdoliva (Artaud *et al.*, 2018)	90.6	96.7	93.5
Baselines			
SVM (texte)	7.73	53.33	13.50
SVM (entités)	5.24	33.33	9.05
SVM (texte + entités)	5.77	40.00	10.08
SVM (triplets)	**29.41**	**100.0**	**45.45**
Approches CamemBERT			
CamemBERT (texte)	6.61	50.0	11.67
CamemBERT (entités)	8.76	73.33	15.66
CamemBERT (texte + entités)	7.39	63.33	13.24
CamemBERT (triplets)	93.75	**100.0**	**96.77**

TABLE 1 – Résultats de l'évaluation de la tâche de détection de tickets fraudés.

vérificateur d'incohérence numérique, nous avons constaté que notre approche est plus performante, comme en témoigne le rappel plus élevé. Cependant, il est important de noter la définition stricte de l'incohérence que nous avons utilisée : nous ne prenons en compte que les incohérences sur les valeurs numériques. La plupart des falsifications non détectées portent sur des valeurs modifiées de manière cohérente (le prix des articles est en accord avec le total et le montant payé, modification d'une date) et devraient être plus difficiles à repérer sans accès à des informations externes. Pourtant, dans l'ensemble de test, ces falsifications ne sont pas plausibles : par exemple, les tickets dans lesquels seule la date a été modifiée sont en fait attribués à une date impossible ou improbable, comme le 32/01, ou une année postérieure à l'arrêt de la collecte des données.

Résultats par type de relation Nous avons également analysé les résultats de la détection des faux tickets en utilisant un seul type de relation à la fois. Trois relations se sont distinguées par leurs résultats étonnamment parfaits (R=100) : *type*, *est_emis_par* et *a_paiement_total*. Ces résultats soulignent différents biais présents dans les données. Par exemple, près de 50 % des faux tickets ont été émis par Carrefour, qui ne représente que 30 % de l'ensemble des tickets. De plus, l'identifiant associé à chaque ticket de caisse n'est pas entièrement aléatoire, car les tickets de caisse sont au moins triés en fonction de l'entreprise qui les a émis. La relation attributive *a_paiement_total* permet d'obtenir des résultats très efficaces. Il faut s'attendre à un certain biais dans les valeurs numériques modifiées, comme la loi de Benford (Nigrini, 2012) qui décrit la distribution non normale des données numériques naturelles et qui a été utilisée dans la détection des fraudes comptables. Dans les nombres réels (tels que les prix, les nombres de population, etc.), le premier chiffre est susceptible d'être petit. En effet, les auteurs de l'ensemble de données signalent que l'utilisation de la loi de Benford pour rechercher des données numériques anormales permet d'obtenir un rappel de 70 % (Artaud, 2019). Ces résultats sont très encourageants quant à la capacité de notre approche à exploiter des informations statistiques, même sur des valeurs numériques, pour détecter les fraudes.

4 Conclusions

Cet article prouve que les méthodes basées sur le contenu sont capables de relever le défi de la détection de la fraude de documents au même niveau que les méthodes basées sur l'image. Notre objectif initial était d'établir une *baseline* et d'encourager les travaux futurs dans le domaine du NLP pour aborder la détection de la fraude de documents, et les résultats ont dépassé nos attentes. Notre approche basée sur le modèle pré-entraîné de CamemBERT considérant les relations entre les entités pour représenter le contenu des tickets atteint des valeurs de rappel élevées en exploitant efficacement les informations extraites des documents sous la forme de triplets.

Remerciements

Ce travail a été soutenu par l'Agence de l'innovation de défense (AID), ainsi que le projet VERINDOC financé par la Région Nouvelle-Aquitaine.

Références

ARTAUD C. (2019). *Détection des fraudes : de l'image à la sémantique du contenu. Application à la vérification des informations extraites d'un corpus de tickets de caisse*. PhD Thesis, University of La Rochelle.

ARTAUD C., DOUCET A., OGIER J.-M. & POULAIN D'ANDECY V. (2017). Receipt dataset for fraud detection. In *First International Workshop on Computational Document Forensics*.

ARTAUD C., SIDÈRE N., DOUCET A., OGIER J.-M. & POULAIN D'ANDECY V. (2018). Find it ! Fraud detection contest report. In *2018 24th International Conference on Pattern Recognition (ICPR)*, p. 13–18.

BEHERA T. K. & PANIGRAHI S. (2015). Credit card fraud detection : a hybrid approach using fuzzy clustering & neural network. In *2015 Second International Conference on Advances in Computing and Communication Engineering*.

BENCHAJI I., DOUZI S. & EL OUAHIDI B. (2018). Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection. In *International Conference on Advanced Information Technology, Services and Systems*.

BERTRAND R., GOMEZ-KRÄMER P., TERRADES O. R., FRANCO P. & OGIER J.-M. (2013). A system based on intrinsic features for fraudulent document detection. In *2013 12th International Conference on Document Analysis and Recognition*, p. 106–110, Washington, DC, USA.

BERTRAND R., TERRADES O. R., GOMEZ-KRÄMER P., FRANCO P. & OGIER J.-M. (2015). A conditional random field model for font forgery detection. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, p. 576–580.

BOROS E., MORENO J. & DOUCET A. (2021). Event detection with entity markers. In *European Conference on Information Retrieval*, p. 233–240.

CARTA S., FENU G., RECUPERO D. R. & SAIA R. (2019). Fraud detection for e-commerce transactions by employing a prudential multiple consensus model. *Journal of Information Security and Applications*, **46**.

COZZOLINO D., GRAGNANIELLO D. & VERDOLIVA L. (2014). Image forgery detection through residual-based local descriptors and block-matching. In *2014 IEEE International Conference on Image Processing (ICIP)*.

COZZOLINO D., POGGI G. & VERDOLIVA L. (2015). Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, **10**(11).

COZZOLINO D. & VERDOLIVA L. (2018). Camera-based image forgery localization using convolutional neural networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*.

COZZOLINO D. & VERDOLIVA L. (2020). Noiseprint : A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, **15**, 144–159.

CRUZ F., SIDÈRE N., COUSTATY M., POULAIN D'ANDECY V. & OGIER J.-M. (2018). Categorization of document image tampering techniques and how to identify them. In *Pattern Recognition and Information Forensics - ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Revised Selected Papers*, p. 117–124.

ELKASRAWI S. & SHAFAIT F. (2014). Printer identification using supervised learning for document forgery detection. In *2014 11th IAPR International Workshop on Document Analysis Systems*, p. 146–150.

FRIDRICH J. & KODOVSKY J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, **7**(3).

GOMEZ-KÄMER P. (2021). Vérification de l'intégrité des documents. *Sécurité multimédia 2 : Biometrie, protection et chiffrement multimedia*, **2**, 71.

HE K., ZHANG X., REN S. & SUN J. (2015). Deep residual learning for image recognition. DOI : 10.48550/ARXIV.1512.03385.

JAMES H., GUPTA O. & RAVIV D. (2020). Ocr graph features for manipulation detection in documents.

KOWSHALYA G. & NANDHINI M. (2018). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*.

LI Y., YAN C., LIU W. & LI M. (2016). Research and application of random forest model in mining automobile insurance fraud. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *ArXiv*, **abs/1907.11692**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MARTÍNEZ TORNÉS B., BOROS E., GOMEZ-KRÄMER P., DOUCET A. & OGIER J.-M. (2023). Detecting Forged Receipts with Domain-specific Ontology-based Entities & Relations. In *Document Analysis and Recognition – ICDAR 2023*.

MIKKILINENI A. K., CHIANG P.-J., ALI G. N., CHIU G. T., ALLEBACH J. P. & DELP III E. J. (2005). Printer identification based on graylevel co-occurrence features for security and forensic applications. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, p. 430–440 : International Society for Optics and Photonics.

MISHRA A. & GHORPADE C. (2018). Credit card fraud detection on the skewed data using various classification and ensemble techniques. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*.

NADIM A. H., SAYEM I. M., MUTSUDDY A. & CHOWDHURY M. S. (2019). Analysis of machine learning techniques for credit card fraud detection. In *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)*, p. 42–47.

NIGRINI M. J. (2012). *Benford's Law : Applications for forensic accounting, auditing, and fraud detection*, volume 586. John Wiley & Sons.

RIZKI A. A., SURJANDARI I. & WAYASTI R. A. (2017). Data mining application to detect financial fraud in indonesia's public companies. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*.

SIDERE N., CRUZ F., COUSTATY M. & OGIER J.-M. (2017). A dataset for forgery detection and spotting in document images. In *2017 Seventh International Conference on Emerging Security Technologies (EST)*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VIDROS S., KOLIAS C., KAMBOURAKIS G. & AKOGLU L. (2017). Automatic detection of online recruitment frauds : Characteristics, methods, and a public dataset. *Future Internet*, **9**(1).

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Jeu de données de tickets de caisse pour la détection de fraude documentaire

Beatriz Martínez Tornés[1] Théo Taburet[1] Emanuela Boros[1] Kais Rouis[1] Petra Gomez-Krämer[1] Nicolas Sidère[1] Antoine Doucet[1] Vincent Poulain d'Andecy[2]
(1) La Rochelle Université, L3i, F-17000, La Rochelle, France
{prénom.nom}@univ-lr.fr
(2) Yooz, 1 Rue Fleming, 17000 La Rochelle, France
Vincent.PoulaindAndecy@getyooz.com

RÉSUMÉ

L'utilisation généralisée de documents numériques non sécurisés par les entreprises et les administrations comme pièces justificatives les rend vulnérables à la falsification. En outre, les logiciels de retouche d'images et les possibilités qu'ils offrent compliquent les tâches de la détection de fraude d'images numériques. Néanmoins, la recherche dans ce domaine se heurte au manque de données réalistes accessibles au public. Dans cet article, nous proposons un nouveau jeu de données pour la détection des faux tickets contenant 988 images numérisées de tickets et leurs transcriptions, provenant du jeu de données SROIE (scanned receipts OCR and information extraction). 163 images et leurs transcriptions ont subi des modifications frauduleuses réalistes et ont été annotées. Nous décrivons en détail le jeu de données, les falsifications et leurs annotations et fournissons deux *baselines* (basées sur l'image et le texte) sur la tâche de détection de la fraude.

ABSTRACT

Receipt Dataset for Document Forgery Detection

The widespread use of unsecured digital documents by companies and administrations as supporting documents makes them vulnerable to forgeries. Moreover, image editing software and the capabilities they offer complicate the tasks of digital image forensics. Nevertheless, research in this field struggles with the lack of publicly available realistic data. In this paper, we propose a new receipt forgery detection dataset containing 988 scanned images of receipts and their transcriptions, originating from the scanned receipts OCR and information extraction (SROIE) dataset. 163 images and their transcriptions have undergone realistic fraudulent modifications and have been annotated. We describe in detail the dataset, the forgeries and their annotations and provide two baselines (image and text-based) on the fraud detection task.

MOTS-CLÉS : Fraude documentaire, jeu de données, détection de fraude.

KEYWORDS: Document forgery, dataset, fraud detection.

1 Introduction

La détection automatique de fraudes est devenue une tâche inévitable dans les flux de documents des entreprises, car l'acceptation de documents falsifiés peut servir comme support à d'autres types de fraude. Par exemple, un fraudeur peut grâce à des faux documents s'assurer une usurpation d'identité

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 140–147.

ou à l'obtention d'un prêt pour financer des activités criminelles telles que des attaques terroristes. Cependant, les travaux de recherche proposés manquent de généralité, car ils sont très spécifiques à un certain type ou méthode de falsification, et il en va de même pour les ensembles de données disponibles. L'un des principaux défis de la détection de la fraude documentaire est le manque de données annotées librement disponibles. En effet, la collecte de documents frauduleux est entravée par la réticence des fraudeurs à partager leur travail, comme on peut s'y attendre dans toute activité illégale, ainsi que par les contraintes qui pèsent sur les entreprises et les administrations pour partager des données sensibles (Sidere *et al.*, 2017; Mishra & Ghorpade, 2018; Vidros *et al.*, 2017). En outre, de nombreuses études sur la fraude ne se concentrent pas sur les documents eux-mêmes, mais sur les transactions, telles que la fraude à l'assurance, la fraude à la carte de crédit ou la fraude financière (Behera & Panigrahi, 2015; Kowshalya & Nandhini, 2018; Rizki *et al.*, 2017). Nous tentons donc de combler ce fossé entre le manque d'ensembles de données de détection de falsifications disponibles publiquement et l'absence de contenu textuel exploitable, en construisant un nouvel ensemble de données génériques pour la détection de falsifications basé sur des images de documents réels. Nous avons basé l'ensemble de données sur jeu de données existant de tickets scannés (SROIE) qui a été initialement proposé pour des tâches d'extraction d'informations et qui contient des images et du texte. Nous avons altéré les images en utilisant plusieurs méthodes d'altération (copier-coller, imitation de texte, suppression d'informations et modification de pixels) et modifié les transcriptions en conséquence. Nous fournissons les images ainsi que les transcriptions permettant une analyse en texte seul.

2 Construction d'un jeu de données pour la détection de faux tickets

S'intéresser aux documents réellement échangés par les entreprises ou les administrations est essentiel pour que les méthodes de détection de fraude développées soient utilisables dans des contextes réels et que la cohérence des documents authentiques soit assurée. Cependant, ces documents administratifs contiennent des informations privées sensibles et ne sont généralement pas mis à disposition de la recherche (Artaud *et al.*, 2018). Nous considérons la tâche de détection de la fraude sur les tickets, car les tickets ne contiennent pas d'informations sensibles et ont une structure très similaire à celle des factures. Ainsi, des scénarios réalistes peuvent être associés à la falsification de tickets, tels que le remboursement de frais de voyage (gagner un peu d'argent supplémentaire, produits non remboursés), et la preuve d'achat (pour l'assurance, pour la garantie).

2.1 Données du SROIE

L'ensemble de données a été choisi comme point de départ pour créer l'ensemble de données de falsification. Il a été créé à l'origine pour l'OCR de tickets numérisés et l'extraction d'informations (SROIE) dans le cadre d'une compétition ICDAR 2019 et contient 1 000 images de tickets numérisés accompagnées de leurs transcriptions, issues de la vérité terrain de la compétition.

Une caractéristique des tickets scannés originaux du SROIE est que certains ont été modifiés, soit numériquement, soit manuellement. Ces modifications ne sont pas considérées comme des faux. Même si les documents ont été modifiés, ils restent authentiques, car ils n'ont pas été falsifiés. Ces annotations conviennent à notre étude de cas, car la plupart d'entre elles sont des notes spécifiques au

contexte que l'on trouve dans les applications de documents réels. Par exemple, certaines annotations correspondent à des notes laissées sur les tickets, telles que « staff outing » pour décrire la nature de l'événement (figure 2), des chiffres qui peuvent décrire une mission ou un numéro de dossier (figures 1 et 4), des noms ou des marques pour mettre en évidence des informations clés sur le document, telles que le prix dans la figure 3. Nombre de ces annotations peuvent même provenir du processus de collecte de l'ensemble de données et sont difficiles à interpréter sans plus d'informations contextuelles (noms, numéros, etc.).

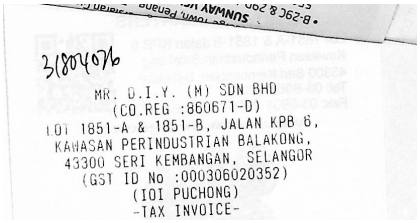


FIGURE 1 – Insertion manuelle de chiffres.

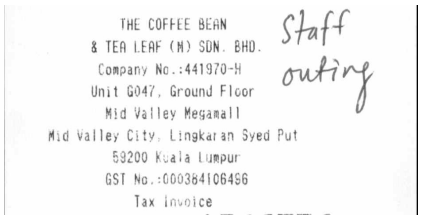


FIGURE 2 – Note sur le ticket.

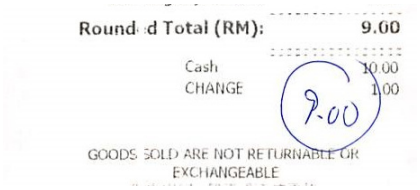


FIGURE 3 – Mise en valeur du total.

FIGURE 4 – Insertion numérique de chiffres.

Ces modifications apportées à des documents authentiques posent le problème difficile de la détection de la fraude, qui consiste à faire la distinction entre une modification frauduleuse et une modification non malveillante. Une modification frauduleuse se caractérise non seulement par la mauvaise intention de son auteur, mais aussi par le fait qu'elle modifie des caractéristiques structurelles ou significatives cruciales du document qui peuvent être utilisées pour déformer le sens du document original.

Afin d'évaluer l'impact de ces modifications, nous avons annoté manuellement les tickets en fonction du type de modifications qu'ils ont subies. Nous définissons une annotation numérique comme un cas particulier récurrent d'une séquence de numéros ou de noms ajoutés numériquement dans plusieurs en-têtes de tickets, comme dans la figure 4. Nous considérons qu'il y a eu une annotation manuelle (figures 1, 2, et 3) s'il y a une note manuscrite de quelque type que ce soit sur le ticket (mots, coches, zones surlignées ou soulignées, etc.), ainsi que des tampons. Nous avons remarqué que les annotations numériques sont reportées dans les transcriptions, alors que les annotations manuelles ne le sont pas. Au total, nous avons compté 54 tickets contenant une annotation numérique, 500 comptant une annotation manuelle et 34 contenant une annotation numérique et une annotation manuelle.

2.2 Campagne de fraude

Afin de fournir des faux aussi réalistes que possible, nous avons organisé plusieurs ateliers de falsification similaires à ceux réalisés par les ensembles de données falsifiées pseudo-réalistes (Artaud *et al.*, 2018; Sidere *et al.*, 2017). Les 19 participants étaient des volontaires, principalement issus du

monde de l'informatique, même si nous avons tenté d'élargir la portée de notre projet à différents niveaux de compétence et d'expertise en matière de documents numériques et de logiciels de retouche d'images. L'objectif n'était pas de créer un ensemble de données d'experts en falsification, mais d'avoir une représentation réaliste de différentes compétences et temps consacré. Les participants n'ont pas reçu de directives spécifiques sur les outils et les techniques à utiliser, afin qu'ils puissent utiliser ce avec quoi ils étaient le plus à l'aise. Cinq logiciels différents ont été utilisés : `aperçu` (15 documents), `paint` (70), `paint3d` (10), `GIMP` (65) et `kolourpaint` (3).

Modification du texte et de l'image Les participants ont reçu des exemples et des scénarios pour commencer, tels que le remboursement de frais de mission, la preuve d'achat pour l'assurance ou pour la garantie (par exemple, date trop ancienne). Il leur a été demandé de modifier l'image ainsi que son fichier texte associé (transcription).

Annotation des fraudes Ensuite, les participants ont été invités à annoter les fraudes qu'ils venaient de réaliser à l'aide de l'annotateur d'images VGG [1]. Ces annotations sont fournies avec l'ensemble de données au format JSON, comme le montre l'exemple suivant :


```
{'filename': 'X51005230616.png', 'size': 835401, 'regions':
[{'shape_attributes': {'name': 'rect', 'x': 27, 'y': 875, 'width': 29,
'height': 43}, 'region_attributes': {'Modified area': {'IMI': True},
'Entity type': 'Product', 'Original area': 'no'}},
{'shape_attributes': {'name': 'rect', 'x': 458, 'y': 883, 'width': 35,
'height': 37},'region_attributes': {'Modified area': {'IMI': True},
'Entity type': 'Product', 'Original area': 'no'}}],
'file_attributes': {'Software used': 'paint', 'Comment': ''}}
```


Le processus a consisté, d'une part, à localiser les zones modifiées en définissant les régions rectangulaires concernées et, d'autre part, à décrire le type de falsification selon la nomenclature proposée dans (Cruz *et al.*, 2019) (copier-coller à partir du même document, copier-coller à partir d'un autre document, suppression d'informations et imitation). Nous incluons un type de falsification supplémentaire (PIX) pour toutes les modifications « à main levée » (James *et al.*, 2020). Nous avons donc proposé les types de falsification suivants :

- **CPI** : Copier et coller à l'intérieur du document, c'est-à-dire copier une partie de l'image (un caractère, un mot entier, une séquence de mots, etc.) et la coller dans la même image ;
- **CPO** : Copier-coller en dehors du document, c'est-à-dire copier une partie de l'image (un caractère, un mot entier, une suite de mots, etc.) et la coller dans un autre document ;
- **IMI** : Zone de texte imitant la police, utilisant un outil d'insertion de texte pour remplacer ou ajouter un texte ;
- **CUT** : Supprimer un ou plusieurs caractères, sans les remplacer ;
- **PIX** : Modification par pixel, pour toutes les modifications effectuées « à main levée » avec un outil de type pinceau pour introduire une modification (par exemple, transformer un caractère en un autre en ajoutant une ligne) ;
- ***Autre*** : Utilisation de filtres ou d'autres éléments (à préciser dans les commentaires).

Annotation des entités modifiées Les participants ont également été invités à identifier le type d'entité altérée pour chaque zone modifiée de la liste suivante :

1. https://www.robots.ox.ac.uk/$\sim$vgg/software/via/

— **Entreprise** : Informations relatives à l'entreprise (adresse, téléphone, nom) ;
— **Produit** : Informations relatives à un produit (nom, prix, suppression ou ajout d'un produit) ;
— **Total/Paiement** : Prix total, mode de paiement ou montant payé ;
— **Metadonnées** : Date, heure.

Post-traitement Toutes les annotations fournies pour les faux tickets sont manuelles. Afin de nous assurer que les zones modifiées ont été correctement annotées par les participants, nous avons corrigé manuellement toutes les annotations. Ces corrections ont été effectuées en comparant les documents falsifiés aux originaux.

2.3 Description du dataset

Le jeu de données contient 988 images PNG avec leurs transcriptions correspondantes dans un format texte. Les données peuvent être téléchargées à l'adresse `http://l3i-share.univ-lr.fr/2023Finditagain/findit2.zip`. Nous proposons une répartition des données entre les ensembles d'entraînement, de validation et de test afin de permettre une comparaison entre les différentes méthodes. La répartition des données est décrite dans le tableau 1, avec les décomptes des faux effectués pendant la campagne de falsification (Section 2.2) ainsi que les annotations présentes dans les tickets authentiques (Section 2.1).

	Entraînement	**Validation**	**Test**	**Total**
Nombre de tickets	577	193	218	988
Nombre de tickets fraudés	94	34	35	163
Nombre de tickets avec une annotation numérique	34	9	11	54
Nombre de tickets avec une annotation manuelle	305	86	109	500

TABLE 1 – Répartition des données.

Au total, 455 zones différents ont été modifiés dans 163 documents. Le tableau 2 détaille le nombre de modifications effectuées par type : une même zone peut avoir été affectée par plus d'un type de modification. En ce qui concerne les entités, la plupart des modifications ont porté sur les informations relatives au total ou au paiement. Le tableau montre également que la technique de falsification la plus utilisée est le CPI.

Type de fraude	**Décompte**	**Type d'entité**	**Décompte**
CPI	353	Total/paiement	234
IMI	36	Produit	95
CUT	36	Métadonnées	82
PIX	33	Entreprise	26
CPO	10	Autre	18

TABLE 2 – Description des zones modifiées.

3 Baselines proposées

Bag-of-words (BoW) & régression logistique Tout d'abord, nous avons choisi ce modèle de classification sur une représentation en sac-de-mots, car il sert généralement de modèle de base et

peut-être utilisé comme référence pour évaluer les résultats et avoir un premier aperçu de la difficulté de la tâche. Nous considérons le modèle le plus couramment utilisé pour une *baseline* simple et rapide : la régression logistique (RL).

ChatGPT Le modèle récent créé par OpenAI [2], proposé en novembre 2022, a suscité une grande attention dans les communautés universitaires et industrielles, et a été rapidement adopté par tout types d'utilisateurs, non seulement en raison de son impressionnante capacité à engager des conversations, mais aussi de sa capacité à répondre aux questions de suivi, à paraphraser, à corriger les fausses déclarations et à décliner les demandes inappropriées (Guo *et al.*, 2023). Nous étions donc curieux de comparer les réponses d'un expert humain et de ChatGPT à la même question (Askell *et al.*, 2021). Nous avons utilisé l'invite suivante [3] :

```
Extract the locations (LOC), products (PROD) and prices (PRI)
from the following receipt and tell me if it's fraudulent:{receipt}
```

Comme les réponses dans le texte libre ne correspondent pas à des résultats de classification binaire, nous les alignons selon deux configurations :

- **Strict** : Seules les réponses exprimant des doutes précis ou des éléments notables concernant le ticket ou sa légitimité ont été classées comme fausses. Pour sept reçus seulement, la réponse indiquait explicitement qu'un élément était « digne d'intérêt », « suspect » ou semblait « frauduleux ».
- **Relâché** : Toutes les réponses qui ne se prononçaient pas explicitement en faveur de la classe authentique on été considérées comme des faux tickets.

Les résultats [4] sont présentés dans la table 3. Le jeu de données étant très déséquilibré, nous ne présentons que les résultats de classification binaire pour la classe « Faux ». L'approche de classification de texte et son très faible rappel montrent à quel point elle est insuffisante : seuls quatre faux tickets ont été correctement étiquetés par le classificateur de texte.

Méthode	Précision	Rappel	F1-score
Classification de texte (BoW + LR)	40.00	11.43	17.78
ChatGPT (strict)	14.69	**88.57**	25.20
ChatGPT (relâché)	18.33	62.86	**28.39**

TABLE 3 – Résultats.

4 Conclusions et perspectives

Cet article présente le jeu de données de tickets librement disponible [5] pour la détection de documents fraudés, contenant à la fois des images et leur transcription de 988 tickets. Il fournit également des annotations sémantiques sur les zones modifiées, ainsi que des détails sur les techniques de falsification utilisées et leurs zones de délimitation. L'ensemble de données peut donc être utilisé pour des tâches de classification et de localisation. Nous pensons que cet ensemble de données peut

2. https://openai.com/blog/chatgpt/
3. L'invite et les réponses ont été traduites de l'anglais par nous.
4. Pour plus de détails et d'expériences sur les images des documents, cet article est une traduction d'un article accepté en version longue (Martínez Tornés *et al.*, 2023).
5. http://l3i-share.univ-lr.fr/2023Finditagain/findit2.zip

constituer une ressource intéressante pour la communauté de détection des faux documents. Les expériences présentées peuvent être considérées comme un point de départ pour comparer avec d'autres méthodes, en particulier le développement d'approches TALN pour l'authentification de documents, ainsi que des approches multimodales. En effet, si l'analyse des documents et la détection de fraude sont majoritairement abordés comme des problématiques de vision par ordinateur, la prise en compte du contenu ainsi que sa cohérence et sa plausibilité est une perspective qui nous semble prometteuse, mais qui est longtemps restée limitée par le manque de données réalistes (ou pseudo-réalistes).

Remerciements

Ce travail a été soutenu par l'Agence de l'innovation de défense (AID), ainsi que le projet VERINDOC financé par la Région Nouvelle-Aquitaine. Nous tenons également à remercier les participants à la campagne de fraude pour leur contribution.

Références

ARTAUD C., SIDÈRE N., DOUCET A., OGIER J.-M. & POULAIN D'ANDECY V. (2018). Find it ! Fraud detection contest report. In *2018 24th International Conference on Pattern Recognition (ICPR)*, p. 13–18.

ASKELL A., BAI Y., CHEN A., DRAIN D., GANGULI D., HENIGHAN T., JONES A., JOSEPH N., MANN B., DASSARMA N. *et al.* (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv :2112.00861*.

BEHERA T. K. & PANIGRAHI S. (2015). Credit card fraud detection : a hybrid approach using fuzzy clustering & neural network. In *2015 Second International Conference on Advances in Computing and Communication Engineering*.

CRUZ F., SIDÈRE N., COUSTATY M., POULAIN D'ANDECY V. & OGIER J.-M. (2019). Categorization of document image tampering techniques and how to identify them. In *International Conference on Pattern Recognition*, p. 117–124 : Springer.

GUO B., ZHANG X., WANG Z., JIANG M., NIE J., DING Y., YUE J. & WU Y. (2023). How close is chatgpt to human experts ? comparison corpus, evaluation, and detection. *arXiv preprint arXiv :2301.07597*.

JAMES H., GUPTA O. & RAVIV D. (2020). Ocr graph features for manipulation detection in documents.

KOWSHALYA G. & NANDHINI M. (2018). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*.

MARTÍNEZ TORNÉS B., TABURET T., ROUIS K., BOROS E., GOMEZ-KRÄMER P., SIDERE N., DOUCET A. & POULAIN D'ANDECY V. (2023). Receipt Dataset for Document Forgery Detection. In *2023 International Conference on Document Analysis and Recognition (ICDAR)*.

MISHRA A. & GHORPADE C. (2018). Credit card fraud detection on the skewed data using various classification and ensemble techniques. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*.

RIZKI A. A., SURJANDARI I. & WAYASTI R. A. (2017). Data mining application to detect financial fraud in indonesia's public companies. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*.

SIDERE N., CRUZ F., COUSTATY M. & OGIER J.-M. (2017). A dataset for forgery detection and spotting in document images. In *2017 Seventh International Conference on Emerging Security Technologies (EST)*.

VIDROS S., KOLIAS C., KAMBOURAKIS G. & AKOGLU L. (2017). Automatic detection of online recruitment frauds : Characteristics, methods, and a public dataset. *Future Internet*, **9**(1).

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Portabilité linguistique des modèles de langage pré-appris appliqués à la tâche du dialogue humain-machine en français

Ahmed Njifenjou Virgile Sucal Bassam Jabaian Fabrice Lefèvre
LIA, Avignon Université, France
{ahmed-ndouop.njifenjou, virgile.sucal, bassam.jabaian,
fabrice.lefevre}@univ-avignon.fr

RÉSUMÉ

Dans cet article, nous proposons une étude de la portabilité linguistique des modèles de langage pré-appris (MLPs) appliqués à une tâche de dialogue à domaine ouvert. La langue cible (L_T) retenue dans cette étude est le français. Elle dispose de peu de ressources spécifiques pour la tâche considérée et nous permet de réaliser une évaluation humaine in situ. La langue source (L_S) est l'anglais qui concentre la majorité des travaux récents dans ce domaine. Construire des MLPs spécifiques pour chaque langue nécessite de collecter de nouveaux jeux de données et cela est coûteux. Ainsi, à partir des ressources disponibles en L_S et L_T, nous souhaitons évaluer les performances atteignables par un système de conversation en L_T. Trois approches sont proposées : TrainOnTarget où le corpus L_S est traduit vers L_T avant l'affinage du modèle, TestOnSource où un modèle L_S est couplé avec des modules de traduction au moment du décodage et TrainOnSourceAdaptOnTarget, qui utilise un MLP multilingue - ici BLOOM (Workshop *et al.*, 2023) - avec l'architecture MAD-X Adapter (Pfeiffer *et al.*, 2020) pour apprendre la tâche en L_S et l'adapter à L_T. Les modèles sont évalués dans des conditions de dialogue oral et les stratégies sont comparées en termes de qualité perçue lors l'interaction.

ABSTRACT

Linguistic portability strategies for open-domain dialogue with pre-trained language models from high to low resource languages

In this paper we propose a study of linguistic portability of pre-trained language models (PLMs) for open-domain dialogue systems in a high-resource language. The target language (L_T) is simulated with French as it lacks task-specific resources and allows an in-situ human evaluation. The source language (L_S) is English which concentrates the majority of recent works. Building specific PLMs for each possible language supposes collecting new datasets and is costly. Hence, leveraging resources from both L_S and L_T, we assess the performance achievable in L_T with three approaches : TrainOnTarget where a L_S dataset is translated in L_T before finetuning, TestOnSource where a L_S model is coupled with translation modules at inference and the TrainOnSourceAdaptOnTarget, using a multilingual PLM - here BLOOM (Workshop *et al.*, 2023) - with MAD-X Adapter architecture (Pfeiffer *et al.*, 2020) to learn the task in L_S and adapt it to L_T. Models are evaluated in spoken dialogue conditions with human and the strategies compared in terms of perceived interaction quality.

MOTS-CLÉS : Agent conversationnel, Transformers, Portabilité multilingue, Langue peu dotée.

KEYWORDS: Conversational agent, Transformers, Crosslingual portability, Low-resource language.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 148–158.

1 Introduction

Depuis l'apparition des modèles Transformers (Vaswani *et al.*, 2017), plusieurs variantes de MLPs ont été déployées dans le domaine du traitement automatique du langage. Les Transformers autorégressifs (utilisant le bloc décodeur) comme GPT (Radford & Narasimhan, 2018), BART (Lewis *et al.*, 2019) etc. se positionnent sur l'état de l'art pour de nombreuses tâches génératives, dont le dialogue à domaine ouvert. Mais pour cela les systèmes doivent développer certaines capacités humaines telles que l'empathie, et avoir un personnalité consistante durant l'interaction (Walker *et al.*, 2021). Dans cette optique, des corpus spécifiques ont été collectés en faisant interagir des humains. On peut citer par exemple : PersonaChat (Zhang *et al.*, 2018), Empathetic Dialogues (Rashkin *et al.*, 2019), Blended Skill Talk (Smith *et al.*, 2020) etc. sur lesquels les MLPs peuvent être affinés. La majorité des corpus disponibles sont en anglais, ou en chinois. L'absence de corpus d'apprentissage spécifiques du même type en français, empêche l'obtention directe de modèles similaires.

Dans ce travail, nous étudions les stratégies de portabilité des chatbots et des corpus d'une langue source (L_S, ici l'anglais) vers une langue cible (L_T, ici le français). En exploitant le maximum de ressources disponibles dans L_S et L_T (outils de traduction automatique neuronale (TAN), corpus et MLPs), nous avons mis en place et mené une évaluation humaine de différents systèmes obtenus par trois approches : TrainOnSource, TestOnTarget, TrainOnsourceAndAdaptOnTarget. Nous avons ensuite comparé ces modèles à un modèle de référence en L_S à savoir BlenderBot 1.0 (Roller *et al.*, 2020).

Pour cela, nous avons revisité les approches proposées pour la portabilité linguistique des modules de compréhension de la parole (Jabaian *et al.*, 2013; Lefèvre *et al.*, 2010) afin de les appliquer au cas des MLPs pour la conversation orale humain-machine. A notre connaissance, l'un des seuls travaux à avoir abordé la question du développement multilingue des ressources pour les chitchat bots basés sur les PLMs est (Lin *et al.*, 2020) qui propose des traductions en plusieurs langues du corpus PersonaChat et l'utilise pour apprendre des modèles en différentes langues. Pour la modélisation du dialogue, nous avons utilisé le même schéma d'apprentissage que celui proposé par (Wolf *et al.*, 2019). En plus de reprendre ce qu'ils ont fait avec un modèle de type GPT (Radford & Narasimhan, 2018), nous avons exploré l'application de la même approche en français et avec le modèle BLOOM (Workshop *et al.*, 2023), un modèle multilingue en libre accès [1].

2 Stratégies pour la portabilité des systèmes de dialogues de l'anglais vers le français

Dans cette étude préliminaire, plutôt que de nous concentrer sur l'amélioration de la performance intrinsèque du dialogue, nous évaluons comment les données et les modèles de L_S peuvent être exploités pour développer des modèles conversationnels simples basés sur des MLPs en L_T.

Les approches *TestOnSource* et *TrainOnTarget* s'appuient sur l'utilisation de modules de TAN à différentes étapes. Pour cela, nous avons utilisé l'API `Google Translate` comme dans (Lin *et al.*, 2020), excellentes performances et facilité d'utilisation expliquent ce choix parmi d'autres.

1. Le modèle est disponible sur https ://huggingface.co/bigscience/bloom

FIGURE 1 – **TestOnSource**

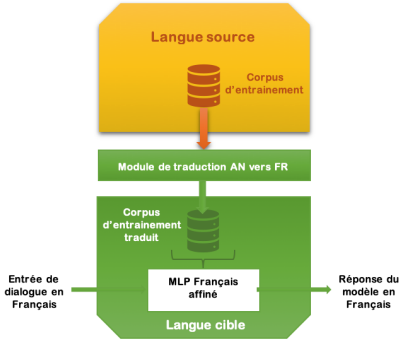


FIGURE 2 – **TrainOnTarget**

TestOnSource : L'approche consiste à utiliser les jeux de données, les modèles de dialogue et les MLPs disponibles en L_S et les combiner avec deux systèmes de TAN qui opèrent pendant les conversations bot-humain en L_T. Le premier (en orange dans la figure 1) traduit l'énoncé de l'utilisateur et le deuxième (en vert dans la figure 1) traduit la sortie du système. La disponibilité d'un grand nombre de ressources, notamment les modèles de dialogue à domaine ouvert en L_S, est un atout majeur pour cette approche. Par conséquent, il est intéressant d'évaluer les performances de ces systèmes sur des entrées traduites de L_T à L_S lors de l'inférence.

TrainOnTarget : Illustrée dans la figure 2, cette approche consiste à affiner les MLPs dans L_T (en vert) pour une tâche de chitchat sur un corpus traduit automatiquement depuis L_S (en jaune). Le français (L_T) dispose de quelques MLPs qui peuvent être utilisés comme base pour des modèles de dialogue. Cette approche suppose que les connaissances spécifiques à la langue, apprises par les MLPs en L_T, peuvent aider à gérer les échantillons bruités issus de la TAN.

TrainOnSourceAdaptOnTarget : Les approches précédentes s'appuient sur le fait qu'en dehors du chitchat, L_T est une langue dotée disposant de modèles de TAN et de MLPs ce qui n'est pas le cas de beaucoup de langues d'où l'idée d'utiliser des MLPs multilingues. Nous reproduisons l'architecture MAD-X (Pfeiffer *et al.*, 2020) pour le dialogue en utilisant **BLOOM** qui a des capacités de traduction et qui intègre une grande variété de langues peu dotées. Dans la figure 3, les flèches vides montrent la 1ère étape d'affinage des adaptateurs de tâche (sur les données de L_S) avec les adaptateurs de langue L_S gelés. Dans la 2ième étape (flèches pleines), les adaptateurs de langue (toujours gelés) passent de L_S à L_T et les mêmes adaptateurs de tâche sont affinés en utilisant peu de données L_T (ou des données traduites). En amont de ces étapes, les adaptateurs de langue L_S et L_T sont appris indépendamment en gelant les paramètres du Transformer.

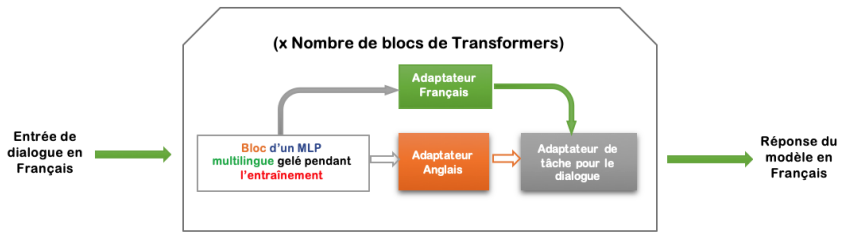


FIGURE 3 – Bloc du Transformer pour le **TrainOnSourceAdaptOnTarget**

3 Étude expérimentale

Afin de comparer et d'évaluer les trois approches présentées dans la section 2, nous avons affiné différent MLPs de L_S, L_T et une version multilingue.

Corpus d'entraînement : PersonaChat (Zhang *et al.*, 2018) est constitué d'un ensemble de dialogues entre deux humains en anglais, chacun se voyant attribué une personnalité initiale en quelques phrases.

Objectif d'entraînement : Nous avons utilisé un double objectif. D'une part la modélisation de langue avec comme entrée la concaténation de la personnalité, l'historique et la réponse. C'est uniquement sur cette dernière que la *fonction d'optimisation (loss function)* est calculée. D'autre part la classification multichoix, pour apprendre à sélectionner la bonne réponse dans un ensemble comprenant plusieurs distracteurs.

Description des modèles : Pour *TrainOnTarget*, nous avons affiné **GPT-fr** (Simoulin & Crabbé, 2021) (124M paramètres), comparable en termes d'architecture et de dimension au modèle Transfer-Transfo (Wolf *et al.*, 2019) basé sur **GPT-1** (117M) qui est utilisé pour *TestOnSource*. Enfin, pour *TrainOnSourceAdaptOnTarget*, le modèle utilisé est **BLOOM-560M**. Multilingue, il est le seul à permettre cette approche. Cette caractéristiques nous permet de l'utiliser également pour construire des modèles en L_S et en L_T pour les deux premières approches respectivement.

4 Évaluation

De part une structure *one-to-many* (Zhao *et al.*, 2017), l'évaluation automatique du dialogue n'est pas toujours en accord avec l'évaluation humaine, qui reste la plus fiable. Nous reportons ici les résultats de l'évaluation humaine accompagnés d'une analyse basée sur les conversations collectées. Les résultats de l'évaluation automatique sont reportés en annexe.

Évaluation humaine : Nous avons collecté 120 conversations via l'interface RASA-X (Bocklisch *et al.*, 2017) en 2 phases : dans la 1ère, nous avons déployé `GPT-fr`, `GPT` et `BlenderBot 1` et dans la 2ième, les 4 modèles basés sur `BLOOM`. Elles ont ensuite été notées de 1 à 5 selon trois critères sélectionnés sur la base de ceux figurant dans (Mehri & Eskénazi, 2020; Ji *et al.*, 2022; Roller *et al.*, 2020) : la cohérence, l'engagement et le naturel. Les résultats sont reporté dans le tableau 1.

TABLE 1 – Différence de notes moyennes avec un modèle de référence (BlenderBot1) par critère

Stratégies	**Modèles**	**Cohérence**	**Engagement**	**Naturel**
Référence (TestOS)	**BlenderBot 1**	**3.64**	**4.45**	**3.77**
TrainOnSource	madx-BLOOM (L_S)	-2,10	-2,37	-2,06
AdapatOnTarget	madx-BLOOM (L_T)	-2,02	-2,14	-2,30
TestOnSource	BLOOM (L_S)	-1,72	-2,07	-1,82
	GPT	-1,59	-2,02	-2,10
TrainOnTarget	GPT-fr	-2,09	-2,49	-2,30
	BLOOM (L_T)	**-1,32**	***-2,01***	**-1,55**

BlenderBot 1, la référence est un modèle plus large ($\sim$ 2.7Mds paramètres, distillés en 400M) appris avec des objectifs d'apprentissage complexes sur une grande variété de corpus. BLOOM_fr émerge comme le meilleur dans les trois catégories évaluées, en moyenne : **+0.26** pour la cohérence, **+0.01** pour l'engagement et **+0.56** pour le naturel par rapport à GPT_EN, son plus proche concurrent. Ce dernier a des notes proches de BLOOM_en avec un léger avantage sur la cohérence (+0.13) et l'engagement (+0.06) et un déficit sur le naturel (-0.29). Le dernier groupe est composé de GPT_FR, madxBLOOM_fr, madxBLOOM_en pour lesquels la note médiane pour tous les critères est proche de 1.5, i.e près de la moitié des conversations avec ces modèles ont reçu la note la plus basse possible.

Analyse des conversations : Le tableau 2 donne un autre aperçu des performances des modèles, en terme de nombre d'échanges. On observe la même tendance que précédemment : les modèles moins bien notés ont un plus faible nombre d'échanges par dialogue, puisque les utilisateurs avaient comme consignes de poursuivre au maximum la discussion. Cela peut expliquer aussi leurs notes d'engagement relativement faibles, ainsi que leurs scores de cohérence. En effet, suivant nos instructions, les conversations sont arrêtées dés que des comportements erratiques tels que des répétitions ou des hallucinations ont été observées par les utilisateurs.

TABLE 2 – Nombre de tours de parole moyen par dialogue

Modèles	BB1	GPT_FR	GPT	xBLOOM_fr	xBLOOM_en	BLOOM_fr	BLOOM_en
#échanges	**35,6**	15,4	**24,8**	12,9	20,6	**24,8**	**36,3**

BB1 = BlenderBot 1, xBLOOM = modèle avec architecture MAD-X et GPT = modèle anglais provenant de (Wolf *et al.*, 2019).

Nous avons utilisé le nombre de mots moyen par tour de parole pour évaluer le comportement relatif des testeurs vis-à-vis des modèles et inversement. Ces grandeurs ont été normalisées par testeur afin de s'affranchir des variabilités entre testeurs (personnalité, attente vis-a-vis d'un agent conversationnel (Walker *et al.*, 2021), habitudes,...) et ainsi observer des variations essentiellement liées aux modèles.

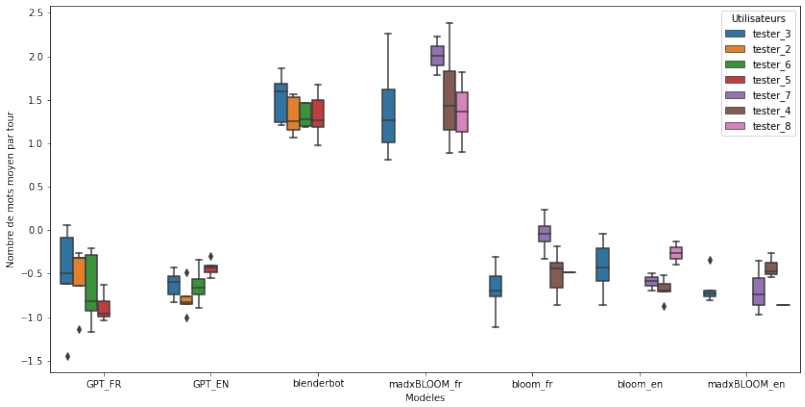


FIGURE 4 – **Nombre moyen de mots par tour de dialogue des differents modèles**

Dans la figure 4, on observe que `BlenderBot` se distingue de tous les autres à l'exception du modèle `madxBLOOM-fr` qui semble produire des réponses plus longue en moyenne. Ceci est dû au fait que, au moment du décodage, ce dernier a été contraint de générer au moins dix nouveaux tokens contrairement aux autres modèles car ayant tendance à générer des messages vides ou très courts avec des mots pas complets. Enfin, globalement les différents modèles ont un comportement qui varie peu d'un utilisateur à autre. Cela met en évidence la problématique **P3** mentionnée dans (Bowden & Walker, 2023) à savoir un manque de personnalisation qui pourrait avoir un impact sur l'expérience utilisateur.

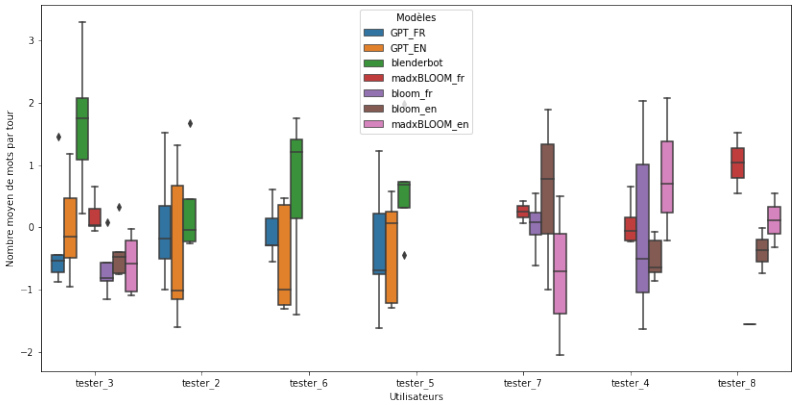


FIGURE 5 – **Nombre moyen de mots par tour de dialogue des testeurs**

Dans la figure 5 on observe *a contrario* que le comportement des utilisateurs varie beaucoup selon le modèle. Une analyse plus fine des dialogues concernés semble indiquer que pour `BlenderBot`, la référence, cela pourrait signifier un bon engagement de l'utilisateur tandis que pour `madxBLOOM-fr` (l'un des moins bien noté) cela pourrait illustrer les tentatives des utilisateurs de compenser la faible qualité des réponses en réorientant le modèle et en s'adaptant à lui. Aussi ce critère ne permet pas de distinguer clairement la qualité des modèles.

Prédiction des évaluations dans une configuration PARADISE (Walker *et al.*, 1997) : L'évaluation humaine implique un processus coûteux de collecte et annotations des conversations. Une solution est de prédire les notes d'évaluation directement à partir de mesures objectives sur les conversations. Pour vérifier cette possibilité, nous mettons en place une approche de type PARADISE. Le tableau 3 reporte les corrélations (r) entre les mesures objectives et les évaluations humaines.

On observe dans le tableau 3 que r est plus important avec les mesures directement dépendantes des modèles : **taille vocabulaire modèle** qui représente le nombre de lemmes différents utilisés par le modèle dans une conversation et le **nombre de mots moyens par tour de dialogue modèle**.

Nous avons entraîné différents modèles de régression des notations à partir des précédentes mesures normalisées (Régression linéaire, SVR, Arbre de Décision et MLP suivant ce qui a été fait dans (Walker *et al.*, 2021)). De notre jeu de 120 conversations, 10% sont aléatoirement assignées au jeu de test et nous entraînons les modèles sur celles restantes. Les notations sont elles aussi normalisées par évaluateur pour réduire le biais dans la notation. Dans le tableau 4 on observe des MSE relativement élevées sauf pour le critère engagement où on arrive avec le SVR à une MSE de **0.07**, un coefficient de détermination de **0.82** et surtout une corrélation de **92%** significative à $p \leq 0.001$. Avec les

TABLE 3 – Corrélations (r) entre les mesures objectives et les évaluations

Grandeurs mesurées	**Cohérence**	**Engagement**	**Naturel**
Nombre d'échanges	0.353	0.389	0.390
#Mots moyen par tour testeur	0.280	0.254	0.309
#Mots moyen par tour modèle	0.338	0.455	0.367
Taille vocabulaire testeur	0.429	0.457	0.468
Taille vocabulaire modèle	0.449	0.548	0.504
Taille vocabulaire conversation	0.466	0.526	0.510

Toutes les mesures significatives avec $p \leq 0,001$

mesures objectives utilisées, on ne pourrait donc prédire que les scores d'engagement avec une certaine qualité. Les évaluations humaines restent nécessaires, où d'autres mesures objectives, moins évidentes, doivent être mobilisées.

TABLE 4 – Résultats des modèles de régression sur les évaluations normalisées

	Cohérence			**Engagement**			**Naturel**		
Modèles	MSE	R^2	r	MSE	R^2	r	MSE	R^2	r
Regression linéaire	0.24	0.31	0.57	0.23	0.47	0.74*	0.27	0.41	0.65
SVR	0.34	0.05	0.48	**0.07**	**0.82**	**0.92****	0.35	0.22	0.57
Arbre de décision	0.42	-0.21	0.42	0.17	0.59	0.79*	0.41	0.096	0.43
MLP	0.28	0.20	0.55	0.21	0.53	0.74*	0.28	0.40	0.64

** significatif $p \leq 0.001$, * significatif $p \leq 0.01$

5 Conclusion

Le développement des modèles de dialogue à domaine ouvert en français est encore loin derrière les modèles anglais, ou même chinois, aujourd'hui. Il en va de même pour de nombreuses autres langues. La raison principale est le manque de corpus spécialisés. Cependant, la disponibilité de MLPs en français et d'outils TAN sont des atouts pouvant être mis à profit pour exploiter les ressources d'une langue plus dotée pour cette tâche. Dans cette optique, nous avons évalué trois approches différentes et comparé les modèles obtenus et un modèle de référence anglais utilisé avec un traducteur automatique. La stratégie TrainOnTarget avec un modèle multilingue a donné les meilleurs résultats (hors modèle de référence) lors de l'évaluation humaine. Ceci ouvre la voie à de futurs travaux sur l'utilisation de données traduites automatiquement avec des modèles multilingues tels BLOOM qui possèdent implicitement des capacités de traduction. L'amélioration des objectifs d'apprentissage pourrait alors permettre de rattraper les performances des modèles de référence des langues bien dotées pour la tâche, malgré l'obstacle que constitue la rareté des corpus spécifiques à chaque langue. Le fait qu'en dehors des dialogues à domaine ouvert, le français soit une langue bien dotée n'est pas totalement limitant pour ces approches. En effet, notre meilleur modèle était basé sur l'approche TrainOnTarget avec BLOOM, un modèle multilingue incluant plusieurs langues peu dotées et en accès libre.

Références

BOCKLISCH T., FAULKNER J., PAWLOWSKI N. & NICHOL A. (2017). Rasa : Open source language understanding and dialogue management. *CoRR*, **abs/1712.05181**.

BOWDEN K. K. & WALKER M. (2023). Let's get personal : Personal questions improve socialbot performance in the alexa prize.

JABAIAN B., BESACIER L. & LEFÈVRE F. (2013). Comparison and combination of lightly supervised approaches for language portability of a language understanding system. *IEEE Transactions on Audio, Speech and Language Processing*, **21**(3), 636–648.

JI T., GRAHAM Y., JONES G. J. F., LYU C. & LIU Q. (2022). Achieving reliable human assessment of open-domain dialogue systems. DOI : 10.48550/ARXIV.2203.05899.

LEFÈVRE F., MAIRESSE F. & YOUNG S. J. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. In *Interspeech 2010, 11th Annual Conference of the International Speech Communication Association*, p. 78–81, Chiba, Japan.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). BART : denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

LIN Z., LIU Z., WINATA G. I., CAHYAWIJAYA S., MADOTTO A., BANG Y., ISHII E. & FUNG P. (2020). Xpersona : Evaluating multilingual personalized chatbot.

MEHRI S. & ESKÉNAZI M. (2020). Unsupervised evaluation of interactive dialog with DialoGPT.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PFEIFFER J., VULIĆ I., GUREVYCH I. & RUDER S. (2020). MAD-X : An adapter-based framework for multi-task cross-lingual transfer.

RADFORD A. & NARASIMHAN K. (2018). Improving language understanding by generative pre-training.

RASHKIN H., SMITH E. M., LI M. & BOUREAU Y.-L. (2019). Towards empathetic open-domain conversation models : a new benchmark and dataset. In *ACL*.

ROLLER S., DINAN E., GOYAL N., JU D., WILLIAMSON M., LIU Y., XU J., OTT M., SHUSTER K., SMITH E. M., BOUREAU Y.-L. & WESTON J. (2020). Recipes for building an open-domain chatbot.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In P. DENIS, N. GRABAR, A. FRAISSE, R. CARDON, B. JACQUEMIN, E. KERGOSIEN & A. BALVET, Éds., *Traitement Automatique des Langues Naturelles*, p. 246–255, Lille, France : ATALA. HAL : hal-03265900.

SMITH E. M., WILLIAMSON M., SHUSTER K., WESTON J. & BOUREAU Y. (2020). Can you put it all together : Evaluating conversational agents' ability to blend skills.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention Is All You Need.

WALKER M. A., HARMON C., GRAUPERA J., HARRISON D. & WHITTAKER S. (2021). Modeling performance in open-domain dialogue with PARADISE.

WALKER M. A., LITMAN D. J., KAMM C. A. & ABELLA A. (1997). PARADISE : A framework for evaluating spoken dialogue agents. *CoRR*, **cmp-lg/9704004**.

WOLF T., SANH V., CHAUMOND J. & DELANGUE C. (2019). Transfertransfo : A transfer learning approach for neural network based conversational agents.

WORKSHOP B., :, SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M., TOW J., RUSH A. M., BIDERMAN S., WEBSON A., AMMANAMANCHI P. S., WANG T., SAGOT B., MUENNIGHOFF N., DEL MORAL A. V., RUWASE O., BAWDEN R., BEKMAN S., MCMILLAN-MAJOR A., BELTAGY I., NGUYEN H., SAULNIER L., TAN S., SUAREZ P. O., SANH V., LAURENÇON H., JERNITE Y., LAUNAY J., MITCHELL M., RAFFEL C., GOKASLAN A., SIMHI A., SOROA A., AJI A. F., ALFASSY A., ROGERS A., NITZAV A. K., XU C., MOU C., EMEZUE C., KLAMM C., LEONG C., VAN STRIEN D., ADELANI D. I., RADEV D., PONFERRADA E. G., LEVKOVIZH E., KIM E., NATAN E. B., TONI F. D., DUPONT G., KRUSZEWSKI G., PISTILLI G., ELSAHAR H., BENYAMINA H., TRAN H., YU I., ABDULMUMIN I., JOHNSON I., GONZALEZ-DIOS I., DE LA ROSA J., CHIM J., DODGE J., ZHU J., CHANG J., FROHBERG J., TOBING J., BHATTACHARJEE J., ALMUBARAK K., CHEN K., LO K., WERRA L. V., WEBER L., PHAN L., ALLAL L. B., TANGUY L., DEY M., MUÑOZ M. R., MASOUD M., GRANDURY M., ŠAŠKO M., HUANG M., COAVOUX M., SINGH M., JIANG M. T.-J., VU M. C., JAUHAR M. A., GHALEB M., SUBRAMANI N., KASSNER N., KHAMIS N., NGUYEN O., ESPEJEL O., DE GIBERT O., VILLEGAS P., HENDERSON P., COLOMBO P., AMUOK P., LHOEST Q., HARLIMAN R., BOMMASANI R., LÓPEZ R. L., RIBEIRO R., OSEI S., PYYSALO S., NAGEL S., BOSE S., MUHAMMAD S. H., SHARMA S., LONGPRE S., NIKPOOR S., SILBERBERG S., PAI S., ZINK S., TORRENT T. T., SCHICK T., THRUSH T., DANCHEV V., NIKOULINA V., LAIPPALA V., LEPERCQ V., PRABHU V., ALYAFEAI Z., TALAT Z., RAJA A., HEINZERLING B., SI C., TAŞAR D. E., SALESKY E., MIELKE S. J., LEE W. Y., SHARMA A., SANTILLI A., CHAFFIN A., STIEGLER A., DATTA D., SZCZECHLA E., CHHABLANI G., WANG H., PANDEY H., STROBELT H., FRIES J. A., ROZEN J., GAO L., SUTAWIKA L., BARI M. S., AL-SHAIBANI M. S., MANICA M., NAYAK N., TEEHAN R., ALBANIE S., SHEN S., BEN-DAVID S., BACH S. H., KIM T., BERS T., FEVRY T., NEERAJ T., THAKKER U., RAUNAK V., TANG X., YONG Z.-X., SUN Z., BRODY S., URI Y., TOJARIEH H., ROBERTS A., CHUNG H. W., TAE J., PHANG J., PRESS O., LI C., NARAYANAN D., BOURFOUNE H., CASPER J., RASLEY J., RYABININ M., MISHRA M., ZHANG M., SHOEYBI M., PEYROUNETTE M., PATRY N., TAZI N., SANSEVIERO O., VON PLATEN P., CORNETTE P., LAVALLÉE P. F., LACROIX R., RAJBHANDARI S., GANDHI S., SMITH S., REQUENA S., PATIL S., DETTMERS T., BARUWA A., SINGH A., CHEVELEVA A., LIGOZAT A.-L., SUBRAMONIAN A., NÉVÉOL A., LOVERING C., GARRETTE D., TUNUGUNTLA D., REITER E., TAKTASHEVA E., VOLOSHINA E., BOGDANOV E., WINATA G. I., SCHOELKOPF H., KALO J.-C., NOVIKOVA J., FORDE J. Z., CLIVE J., KASAI J., KAWAMURA K., HAZAN L., CARPUAT M., CLINCIU M., KIM N., CHENG N., SERIKOV O., ANTVERG O., VAN DER WAL O., ZHANG R., ZHANG R., GEHRMANN S., MIRKIN S., PAIS S., SHAVRINA T., SCIALOM T., YUN T., LIMISIEWICZ T., RIESER V., PROTASOV V., MIKHAILOV V., PRUKSACHATKUN Y., BELINKOV Y., BAMBERGER Z., KASNER Z., RUEDA A., PESTANA A., FEIZPOUR A., KHAN A., FARANAK A., SANTOS A., HEVIA A., UNLDREAJ A., AGHAGOL A., ABDOLLAHI A., TAMMOUR A., HAJIHOSSEINI A., BEHROOZI B., AJIBADE B., SAXENA B., FERRANDIS C. M., CONTRACTOR D., LANSKY D., DAVID D., KIELA D., NGUYEN D. A., TAN E., BAYLOR E., OZOANI E., MIRZA F., ONONIWU F., REZANEJAD H., JONES H., BHATTACHARYA I., SOLAIMAN I., SEDENKO I., NEJADGHOLI I., PASSMORE J., SELTZER J.,

SANZ J. B., DUTRA L., SAMAGAIO M., ELBADRI M., MIESKES M., GERCHICK M., AKINLOLU M., MCKENNA M., QIU M., GHAURI M., BURYNOK M., ABRAR N., RAJANI N., ELKOTT N., FAHMY N., SAMUEL O., AN R., KROMANN R., HAO R., ALIZADEH S., SHUBBER S., WANG S., ROY S., VIGUIER S., LE T., OYEBADE T., LE T., YANG Y., NGUYEN Z., KASHYAP A. R., PALASCIANO A., CALLAHAN A., SHUKLA A., MIRANDA-ESCALADA A., SINGH A., BEILHARZ B., WANG B., BRITO C., ZHOU C., JAIN C., XU C., FOURRIER C., PERIÑÁN D. L., MOLANO D., YU D., MANJAVACAS E., BARTH F., FUHRIMANN F., ALTAY G., BAYRAK G., BURNS G., VRABEC H. U., BELLO I., DASH I., KANG J., GIORGI J., GOLDE J., POSADA J. D., SIVARAMAN K. R., BULCHANDANI L., LIU L., SHINZATO L., DE BYKHOVETZ M. H., TAKEUCHI M., PÀMIES M., CASTILLO M. A., NEZHURINA M., SÄNGER M., SAMWALD M., CULLAN M., WEINBERG M., WOLF M. D., MIHALJCIC M., LIU M., FREIDANK M., KANG M., SEELAM N., DAHLBERG N., BROAD N. M., MUELLNER N., FUNG P., HALLER P., CHANDRASEKHAR R., EISENBERG R., MARTIN R., CANALLI R., SU R., SU R., CAHYAWIJAYA S., GARDA S., DESHMUKH S. S., MISHRA S., KIBLAWI S., OTT S., SANG-AROONSIRI S., KUMAR S., SCHWETER S., BHARATI S., LAUD T., GIGANT T., KAINUMA T., KUSA W., LABRAK Y., BAJAJ Y. S., VENKATRAMAN Y., XU Y., XU Y., XU Y., TAN Z., XIE Z., YE Z., BRAS M., BELKADA Y. & WOLF T. (2023). Bloom : A 176b-parameter open-access multilingual language model.

ZHANG S., DINAN E., URBANEK J., SZLAM A., KIELA D. & WESTON J. (2018). Personalizing Dialogue Agents : I have a dog, do you have pets too ? *arXiv.org*.

ZHAO T., ZHAO R. & ESKENAZI M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. DOI : 10.48550/ARXIV.1703.10960.

Annexe : Évaluation automatique

TABLE 5 – Évaluation automatique des différents modèles

Stratégies	**Modèles***	**Perplexité** ↓	**Hits@1** ↑	**BLEU** ↑
TrainOnTarget	GPT-fr	**10,82**	0,88	N/A
	BLOOM (L_T)	16,05	**0,95**	0,23
TestOnSource	*GPT*	18,49	0,84	N/A
	BLOOM (L_S)	**13,01**	**0,94**	0,22
Entraînement Cross-Lingue**	*XNLG*(L_S)	54,74	N/A	**2,25**
	madx-BLOOM (L_S)	**24,07**	0,82	0,13
	XNLG (L_T)	640.33	N/A	0,09
	madx-BLOOM (L_T)	**28,64**	0,81	**0,15**

* Les modèles en italique et les métriques associées proviennent de l'état de l'art : *GPT* (Wolf *et al.*, 2019), *XNLG* (Lin *et al.*, 2020).
** Le terme « *Entraînement cross-lingue* » est utilisé ici car l'approche des modèles de l'état de l'art n'est pas identique au *TrainOnSourceAdapatOnTarget*.

Le tableau 5 reporte la perplexité disponible pour les modèles issus de l'état de l'art, le Hits@1 qui mesure la précision avec laquelle la « vraie » réponse à une entrée de dialogue est classée première

parmi plusieurs distracteurs (second objectif d'entraînement) et le score BLEU (Papineni *et al.*, 2002) à des fins de comparaison avec d'autres modèles.

Dans le cadre cross-lingue, l'architecture MAD-X avec BLOOM apporte un gain important pour les modèles L_T : de 640 à 28 pour la perplexité. Ceci garantit des capacités de génération mais pas nécessairement de dialogue donnant des résultats médiocres dans l'évaluation humaine comme montré précédemment dans le tableau 1.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Détection d'événements à partir de peu d'exemples par seuillage dynamique

Aboubacar Tuo Romaric Besançon Olivier Ferret Julien Tourille
Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{aboubacar.tuo,romaric.besancon,olivier.ferret,julien.tourille}@cea.fr

RÉSUMÉ

Les études récentes abordent la détection d'événements à partir de peu de données comme une tâche d'annotation de séquences en utilisant des réseaux prototypiques. Dans ce contexte, elles classifient chaque mot d'une phrase donnée en fonction de leurs similarités avec des prototypes construits pour chaque type d'événement et pour la classe nulle « non-événement ». Cependant, le prototype de la classe nulle agrège par définition un ensemble de mots sémantiquement hétérogènes, ce qui nuit à la discrimination entre les mots déclencheurs et non déclencheurs. Dans cet article, nous abordons ce problème en traitant la détection des mots non-déclencheurs comme un problème de détection d'exemples « hors-domaine » et proposons une méthode pour fixer dynamiquement un seuil de similarité pour cette détection.

ABSTRACT

Few Shot Event Detection with Dynamic Thresholding

Recent studies in few-shot event trigger detection from text address the task as a word sequence annotation task using prototypical networks. In this context, the classification of a word is based on the similarity of its representation to the prototypes built for each event type and for the "non-event" class. However, the "non-event" prototype aggregates by definition a set of semantically heterogeneous words, which hurts the discrimination between trigger and non-trigger words. We address this issue by handling the detection of non-trigger words as an out-of-domain detection problem and propose a method for dynamically setting a similarity threshold for this detection.

MOTS-CLÉS : Détection d'événements à partir de peu d'exemples, Méta-apprentissage.

KEYWORDS: Few-shot Event Detection, Meta-learning.

1 Introduction

La détection d'événements est une tâche de l'extraction d'information qui vise à extraire des instances de types d'événements donnés à partir de textes (Nguyen & Grishman, 2015a,b). Cette extraction consiste à identifier des déclencheurs d'événements, qui sont des groupes de mots indiquant explicitement la présence d'un événement dans une phrase. Par exemple, dans la phrase « *John D. Idol will [**take over**] as Chief Executive.* », un événement « Start-Position » est déclenché par le déclencheur « *take over* ». Les approches d'apprentissage supervisé pour la détection d'événements ont été largement étudiées ces dernières années, notamment les méthodes fondées sur des traits lexico-syntaxiques (Li *et al.*, 2013; Liao & Grishman, 2011), les réseaux neuronaux convolutifs (Nguyen & Grishman, 2015a), les réseaux neuronaux récurrents (Nguyen *et al.*, 2016) et les modèles fondés sur les graphes (Liu *et al.*, 2018; Nguyen & Grishman, 2018; Yan *et al.*, 2019). Cependant, toutes

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 159–168.

ces approches reposent sur des ensembles de données annotées conséquents pour l'entraînement, généralement difficiles à obtenir.

La détection d'événements à partir de peu d'exemples (*Few-Shot Event Detection*, FSED) a donc suscité un grand intérêt ces dernières années avec l'émergence de méthodes d'apprentissage à partir de peu de données, notamment via le méta-apprentissage (Snell *et al.*, 2017; Vinyals *et al.*, 2016; Sung *et al.*, 2018; Geng *et al.*, 2019), et le développement de modèles de langue pré-entraînés capables de transférer leurs connaissances linguistiques à de nouvelles tâches. Elle a été mise en œuvre sous plusieurs formes : *identification d'événement*, qui détermine si un mot dans une phrase est un déclencheur selon un type d'événement (Bronstein *et al.*, 2015; Chen *et al.*, 2021), *classification d'événement*, dont l'objectif est de choisir le type d'événement associé à un déclencheur déjà identifié dans une phrase (Shen *et al.*, 2021; Deng *et al.*, 2020; Lai & Nguyen, 2019; Lai *et al.*, 2020, 2021), et *la détection*, qui réalise ces deux étapes conjointement (Cong *et al.*, 2021; Tuo *et al.*, 2022).

Ces efforts de recherche ont fait de la FSED une tâche d'annotation de séquences, qui se transforme en un problème de classification de mots traité à l'aide de réseaux prototypiques (Snell *et al.*, 2017), qui sont particulièrement adaptés à l'apprentissage à partir de peu d'exemples. Dans ce contexte, un prototype est construit pour chaque type d'événement et la classe « non-événement » (aussi appelée classe nulle) (Yang & Katiyar, 2020; Cong *et al.*, 2021; Tuo *et al.*, 2022). Cependant, l'hétérogénéité intrinsèque du prototype « non-événement » rend difficile la discrimination entre les mots déclencheurs et non déclencheurs fondée sur la similarité avec les prototypes.

Pour résoudre ce problème, nous formulons la FSED comme un problème de détection hors domaine (Schölkopf *et al.*, 2001), en considérant les mots de la classe nulle comme des exemples hors-domaine et en apprenant un seuil de similarité dynamique afin que ces exemples ne soient associés à aucune classe d'événement. En résumé, notre contribution est triple : (1) nous proposons une nouvelle façon de traiter la classe nulle dans la FSED ; (2) nous définissons un nouveau modèle pour la FSED en utilisant des réseaux prototypiques et une fonction de coût contrastive ; (3) nous calculons un seuil de décision dynamique en utilisant la fonction de répartition empirique (*ECDF*). Ces contributions se traduisent par des gains significatifs, évalués sur plusieurs jeux de données [1].

2 Approche

2.1 Formulation du problème

Nous formulons la FSED comme un apprentissage épisodique à N-ways et k-shots (Vinyals *et al.*, 2016) avec des réseaux prototypiques. La tâche de détection des déclencheurs est prise comme un problème d'annotation de séquence au niveau des mots, en s'appuyant sur le format BIO (*Beginning-Inside-Outside*) comme dans Cong *et al.* (2021); Tuo *et al.* (2022) [2]. À chaque épisode, nous considérons un sous-ensemble de phrases annotées appelé *support set*. Il contient N types d'événements et k exemples annotés par type (k étant petit, par exemple de 1 à 10). Un second ensemble, appelé *query set*, est utilisé pour faire des prédictions fondées sur les exemples annotés du *support set*. Chaque phrase peut contenir un ou plusieurs déclencheurs, associés chacun à un type d'événements. L'identification de ce type et de la position du déclencheur est effectuée en attribuant une étiquette à chaque mot, ce qui correspond à une classification multi-classe au niveau des mots, avec autant de classes

1. Cet article reprend par ailleurs le travail présenté dans (Tuo *et al.*, 2023).
2. Les déclencheurs événementiels peuvent être des multi-mots, en pratique en très faible nombre dans les corpus d'évaluation, mais n'admettent pas d'insertion, ce qui rend le format BIO suffisant.

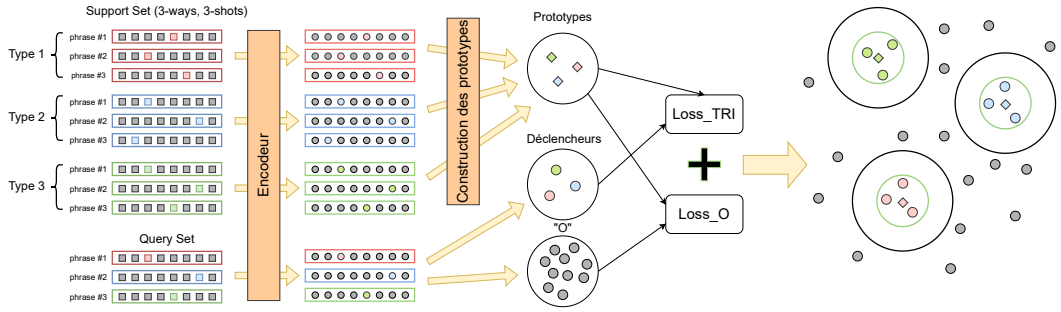


FIGURE 1 – Vue d'ensemble du modèle

que de types d'événements plus une classe nulle (étiquette « O ») pour les mots non-déclencheurs d'événements.

Nous construisons un prototype pour chaque classe à partir des exemples du *support set* en prenant la moyenne des représentations des k déclencheurs de cette classe. Ensuite, nous classons chaque mot du *query set* en fonction de sa similarité avec ces prototypes. Pendant l'apprentissage, ces similarités sont utilisées pour mettre à jour les poids du modèle via une fonction de coût. Cependant, cette formulation implique d'avoir un prototype pour la classe « O » qui est construit en pratique en rassemblant des mots qui ne sont pas sémantiquement homogènes. Nous proposons de traiter cette classe comme une classe « hors-domaine ». Inspirés par les efforts de recherche sur la classification d'exemple hors-domaine avec peu d'exemples (Tan *et al.*, 2019; Nimah *et al.*, 2021), nous évitons de construire le prototype « O » et proposons une approche fondée sur un seuillage dynamique adapté à chaque phrase en utilisant l'ECDF des similarités entre les mots et les prototypes.

2.2 Modèle

La figure 1 présente une vue d'ensemble de notre modèle dont les composants sont détaillés ci-après.

Encodeur Ce composant prend une phrase en entrée et produit une représentation contextuelle pour chaque mot. Pour une phrase $x = w_1, \ldots, w_L$, de longueur L, l'encodeur fournit $\bar{\mathbf{e}} = \mathbf{e_1}, \ldots, \mathbf{e_L}$, où $\mathbf{e_i}$ est la représentation du mot w_i.

Module prototypique Ce composant construit un prototype pour chaque type d'événements en faisant la moyenne des représentations des mots déclencheurs du *support set* et classe les mots du *query set* en fonction de leurs similarités avec ces prototypes. Contrairement à Tuo *et al.* (2022) et Cong *et al.* (2021), nous ne construisons pas de prototype pour la classe nulle. Nous nous fondons sur un seuil de similarité pour décider si un mot est déclencheur ou non.

Entraînement La fonction de coût habituellement utilisée dans les réseaux prototypiques est l'entropie croisée (*cross-entropy*). Nous proposons un apprentissage contrastif plus adapté à l'apprentissage de métrique. Contrairement à l'entropie croisée, dont l'objectif est d'apprendre à prédire une étiquette ou des valeurs à partir d'une entrée, les fonctions contrastives prédisent la similarité relative entre les entrées. Une telle fonction est plus appropriée dans notre cas puisque nous cherchons justement à rendre les déclencheurs plus proches de leurs prototypes que les mots « O » . Pour une classe y donnée, la fonction de coût comporte deux termes :

- **Loss-TRI** : rapproche le déclencheur $\mathbf{e_{tr}}$ de son prototype $\mathbf{c^y}$ et l'éloigne des autres prototypes $c^{j \neq y}$:

$$\mathcal{L}_{TRI}(\bar{\mathbf{e}}, y) = \sum_{j \neq y} \max(0, \mathcal{M}_0 + s(\mathbf{e_{tr}}, \mathbf{c^j}) - s(\mathbf{e_{tr}}, \mathbf{c^y})) \quad (1)$$

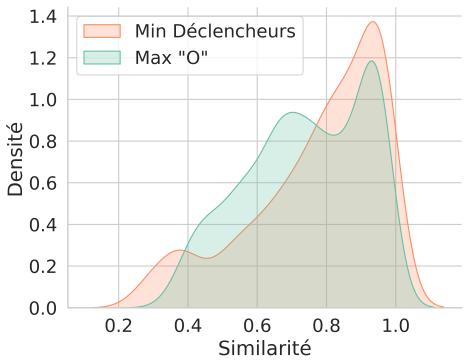
- **Loss-O** : éloigne les mots « O » $\mathbf{e_i}$ $(i \neq tr)$ de tous les prototypes c^j.

$$\mathcal{L}_O(\bar{\mathbf{e}}, y) = \max_{i \neq tr}(0, \max_j(s\left(\mathbf{e_i}, \mathbf{c^j}\right) - \mathcal{M}_1)) \quad (2)$$

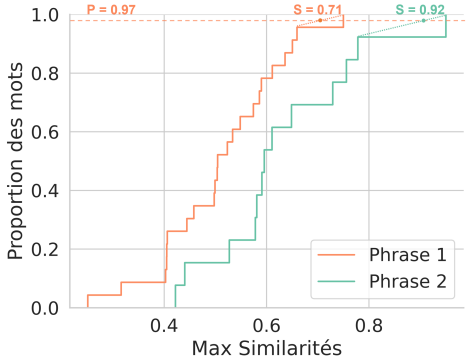
Dans ces fonctions de coût, la fonction $s(.)$ fait référence à la similarité entre la représentation d'un déclencheur et celle du prototype d'une classe tandis que $\mathcal{M}_0$ et $\mathcal{M}_1$ sont des hyperparamètres correspondant aux marges. Le modèle est entraîné en minimisant une fonction de coût correspondant à la somme de ces deux termes.

Classification et traitement de la classe « O » L'approche standard avec les réseaux prototypiques est de classer chaque mot en fonction de sa similarité avec les prototypes. Dans notre modèle, en l'absence de prototype pour la classe nulle, nous devons nous fier à un seuil en dessous duquel le mot est considéré comme un non-déclencheur. Typiquement, dans des travaux tels que Tan *et al.* (2019) et Nimah *et al.* (2021), un seuil global est défini en utilisant la distribution des valeurs de similarité sur un ensemble de validation. Cependant, dans notre cas, nous avons observé empiriquement que les distributions des valeurs de similarité entre un déclencheur et les prototypes varient trop d'une phrase à l'autre (cf. Figure 2a), ce qui rend impraticable l'utilisation d'un seuil global.

Pour résoudre ce problème, nous proposons de rechercher la probabilité correspondant au seuil optimal en utilisant la fonction de répartition sur les valeurs maximales de similarité. Ceci nous permet d'obtenir un seuil dynamique spécifique à la phrase considérée. Plus précisément, étant donné que les similarités des déclencheurs sont plus élevées que celles des mots « O » , nous supposons que, pour une phrase donnée, les similarités des déclencheurs ne seront présentes qu'au-dessus d'un certain quantile (assez élevé) dans la distribution des similarités. Nous supposons également que ce quantile est assez stable, même s'il ne correspond pas à la même valeur de similarité d'une phrase à l'autre. En pratique, pour une phrase donnée du *query set*, nous sélectionnons la phrase la plus similaire dans le *support set*. Puis, nous faisons varier le seuil entre les similarités minimum et maximum et adoptons celui maximisant la f1-mesure sur la phrase sélectionnée. Ensuite, nous déterminons la probabilité correspondant à ce seuil en utilisant la fonction de répartition sur les valeurs de similarité. Enfin, nous déterminons le seuil optimal pour la phrase du *query set* à partir de sa fonction de répartition et de la probabilité déterminée précédemment. Toutefois, comme les probabilités directement calculées à partir de la fonction de répartition dépendent du nombre de mots dans les phrases, nous interpolons linéairement la fonction de répartition sur un plus grand nombre de points avant d'estimer les probabilités, ce qui nous permet de donner artificiellement à toutes les phrases la même longueur (nous utilisons 512 points). Dans l'exemple de la Figure 2b, la « phrase 1 » (du *support set*) a son seuil optimal à $0{,}71$ correspondant à une probabilité de $0{,}97$. Nous reportons ensuite cette probabilité sur la fonction de répartition de la « phrase 2 » (du *query set*) pour obtenir son seuil optimal, égal à $0{,}92$.



(a) Similarités : min déclencheurs VS max « O »



(b) Fonctions de répartition de deux phrases

FIGURE 2 – La figure 2a montre que les mots déclencheurs et les mots « O » ne sont pas séparables avec un seuil global parce que la distribution des valeurs minimales des similarités des déclencheurs (Min déclencheurs) et la distribution des valeurs maximales des similarités des « O » (Max « O ») se chevauchent de façon importante. La Figure 2b montre, à partir de l'ECDF de deux exemples de phrases, que la similarité optimale varie d'une phrase à l'autre.

3 Expériences

3.1 Méthode d'évaluation et hyperparamètres

Nous expérimentons sur les ensembles de données ACE 2005 (Walker *et al.*, 2006), MAVEN (Wang *et al.*, 2020) et FewEvent (Deng *et al.*, 2020). Nous utilisons les découpages de Chen *et al.* (2021) pour ACE 2005 et MAVEN et celui de Cong *et al.* (2021) pour FewEvent. Dans tous les cas, les ensembles de test et d'apprentissage contiennent des classes distinctes, de sorte que lors de l'évaluation, le modèle doit faire face à de nouvelles classes qu'il n'a jamais vues auparavant.

Nous adoptons l'évaluation épisodique N ways, k shots, qui consiste à construire des épisodes avec N classes et k exemples annotés par classe. Dans l'évaluation épisodique standard (Vinyals *et al.*, 2016), les ensembles de test sont échantillonnés de façon à ce que toutes les classes soient distribuées uniformément, ce qui ne correspond pas à la distribution des mentions d'événements dans les données réelles. Ainsi, les scores de performance rapportés par cette méthode ne reflètent pas la distribution réelle des données. Nous adoptons la configuration plus réaliste de Yang & Katiyar (2020), qui construit le *support set* avec $N * K$ exemples et évalue le modèle sur tous les autres exemples.

Pour les expériences, nous avons utilisé le modèle pré-entraîné BERT-base comme encodeur et adopté la stratégie *Weighted* proposée par Tuo *et al.* (2022) pour obtenir des représentations contextuelles des mots. Nous adoptons une longueur maximale de séquence de 128 tokens, un taux d'apprentissage de 1e-5 et 30 000 épisodes N-ways, k-shots pour entraîner le modèle. Les hyper-paramètres $\mathcal{M}_0 = 1$ et $\mathcal{M}_1 = 0{,}4$ ont été obtenus sur l'ensemble de validation, pris entre $0{,}2$ et 1 (avec un pas de $0{,}2$).

3.2 Résultats et analyses

Nous comparons notre approche, **OUTFIT** (i.e. OUT oF trIgger deTection), à quatre autres modèles de l'état de l'art dans la configuration 5-ways 5-shots. **PA-CRF** (Cong *et al.*, 2021) et Tuo *et al.* (2022) sont des modèles de l'état de l'art qui calculent un prototype pour la classe « O » . **PA-CRF** estime les probabilités de transition entre les étiquettes BIO avec l'utilisation de CRF (*Conditional Random Fields)* (Lafferty *et al.*, 2001) tandis que Tuo *et al.* (2022) propose une meilleure exploitation

	Modèle	ACE 2005	MAVEN	FewEvent
5-ways, 5-shots	PROTO	49,2 ± 1,2	51,6 ± 1,4	53,6 ± 0,7
	PA-CRF (Cong *et al.*, 2021)	64,0 ± 0,6	65,2 ± 0,3	65,3 ± 2,0
	(Tuo *et al.*, 2022)	66,4 ± 1,8	67,1 ± 1,5	67,4 ± 1,1
	HCL-TAT† (Zhang *et al.*, 2022)	–	–	66,9 ± 0,7
	OUTFIT (ours)	**74,0*** ± **1,1**	76,9 ± 1,1	**79,6*** ± **4,2**
	− PoS tags	72,2 ± 2,2	**77,5*** ± **0,8**	77,9 ± 3,9
	− contrastive	66,5 ± 5,7	63,1 ± 12,6	75,9 ± 5,4
	− weighted	59,2 ± 3,6	50,0 ± 2,3	70,9 ± 2,7
	Seuil oracle	82,5 ± 1,9	87,2 ± 1,1	84,1 ± 0,5
1w,5s	FS-Causal† (Chen *et al.*, 2021)	76,9 ± 1,4	55,0 ± 0,4	–
	OUTFIT	80,9 ± 2,9	81,1 ± 1,1	79,1 ± 2,1

TABLE 1 – Performance de détection d'événement sur trois jeux de données, en moyenne et écart-type de la micro f1-mesure sur 5 essais. † indique les résultats issus de l'article original. * indique que la différence entre le meilleur modèle (**en gras**) et le deuxième (souligné) est statistiquement significative, en utilisant le test de significativité de (Dror *et al.*, 2019).

des couches du modèle BERT afin d'obtenir plus d'information pour l'apprentissage. **HCL-TAT** (Zhang *et al.*, 2022) est également un modèle sans prototype pour la classe nulle utilisant un seuil de décision égal à la moyenne des similarités pendant un épisode. Nous comparons ces méthodes à un modèle prototypique de base qui construit un prototype pour la classe nulle, utilise l'entropie croisée comme fonction de coût et un encodeur BERT-base (**PROTO**). **FS-Causal** (Chen *et al.*, 2021) est un modèle ajoutant une prise en compte explicite des relations de causalité entre les déclencheurs et leur contexte pour résoudre le problème dit de la malédiction des déclencheurs (*trigger curse*). En effet, un surapprentissage des déclencheurs peut nuire à la détection des déclencheurs rares dans l'ensemble de données. La prise en compte du contexte seul (sans les déclencheurs) permet de résoudre ce problème. Comme leurs résultats rapportés ne sont évalués que classe par classe, cela correspond à une configuration 1-way 5-shots. Nous avons également ajouté les résultats correspondant au seuil optimal trouvé directement sur les instances du *query set* (**Seuil oracle**), ce qui donne une indication du meilleur résultat pouvant être obtenu avec notre approche.

Dans les expériences préliminaires, nous avons remarqué que la précision ($\approx 65\%$) était relativement faible par rapport au rappel ($\approx 80\%$), ce qui indique que le modèle identifiait trop de mots comme déclencheurs. Pour augmenter la précision, nous avons filtré les prédictions en fonction de leurs catégories morphosyntaxiques (*PoS tags*), en ne conservant que les étiquettes les plus couramment associées aux déclencheurs d'événements dans l'ensemble d'apprentissage (verbe, adverbe et nom) [3].

Notre méthode établit une nouvelle performance de l'état de l'art avec une augmentation moyenne de 10 points de la f1-mesure pour les trois jeux de données considérés (cf. Tableau 1). Les analyses suggèrent que l'encodeur *Weighted* et l'apprentissage contrastif, combinés à notre nouvelle formulation, jouent un rôle important dans la performance globale du modèle. Plus spécifiquement, nous pouvons noter que la fonction contrastive contribue fortement à diminuer la variance des résultats. Nous pensons également que cette fonction, combinée à notre stratégie de recherche de seuil, contribue à la forte différence de performance avec HCL-TAT alors que nos problématiques sont initialement proches. Comme nos expériences préliminaires l'ont suggéré, le filtrage des déclencheurs candidats en fonction de leurs catégories morphosyntaxiques permet d'augmenter les performances de quelques points pour deux jeux de données. Toutefois, la condition sans ce filtrage, qui est la plus

3. Les déclencheurs événementiels sont essentiellement nominaux et verbaux mais ils peuvent inclure des mots qui ne sont ni des noms, ni des verbes, le cas le plus fréquent en anglais étant celui des verbes à particule.

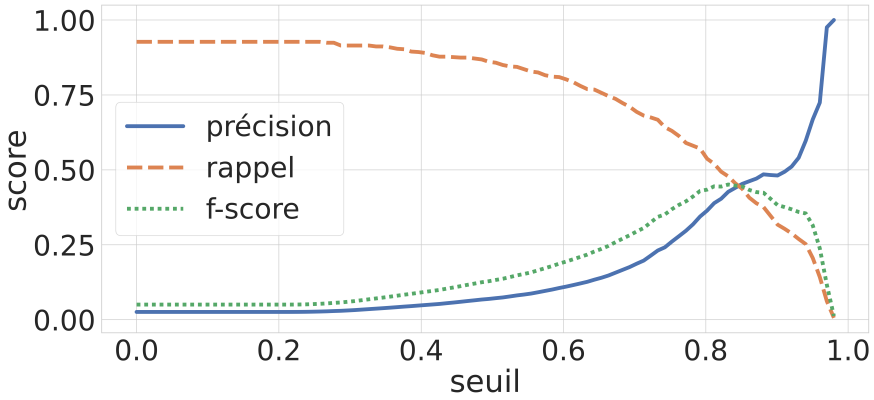


FIGURE 3 – Scores en utilisant un seuil global sur le jeu de données FewEvent

très directement comparable aux modèles de l'état de l'art, montre que celui-ci n'est pas le facteur principal des améliorations obtenues.

Dans la configuration 1-way 5-shots, notre modèle améliore également les performances par rapport à FS-Causal pour les deux jeux de données avec des résultats pour FS-Causal. Ce résultat montre d'abord que l'amélioration apportée par notre proposition n'est pas limitée à un unique cadre d'évaluation. Par ailleurs, considérer de nouveaux types d'événements un par un est la stratégie la plus générale pour l'adaptation à un nouveau domaine dans lequel le nombre de types d'événements n'est pas connu à l'avance. Cependant, l'écart important entre l'oracle et notre modèle suggère que notre approche pourrait être encore améliorée.

Enfin, la Figure 3 montre les scores pour un seuil global allant de 0 à 1 sur le jeu de données FewEvent, pour la condition 5-ways 5-shots. Nous remarquons que la meilleure f1-mesure pouvant être obtenue avec un seuil global est d'environ 0,45, ce qui justifie clairement l'intérêt d'adopter un seuil dynamique plutôt qu'un seuil global comme dans (Tan *et al.*, 2019) ou (Nimah *et al.*, 2021).

4 Conclusion et perspectives

Dans cet article, nous abordons la détection d'événement à partir de peu d'exemples comme une tâche de détection hors domaine en utilisant des réseaux prototypiques. Cette méthode évite de construire un prototype pour la classe nulle, qui est par nature hétérogène, et fournit un seuil dynamique pour décider si un mot est un déclencheur ou non. Les résultats expérimentaux suggèrent que cette nouvelle formulation offre une amélioration importante des performances par rapport aux autres méthodes de l'état de l'art. À notre connaissance, il s'agit du premier effort de recherche qui présente la FSED comme une tâche d'annotation de séquence tout en traitant la classe nulle comme un problème de détection hors domaine. Nous pensons que notre méthode pourrait être appliquée à d'autres tâches d'annotation de séquences et nous étudierons plus particulièrement son application à l'extraction d'arguments d'événements et à la reconnaissance d'entités nommées dans le cadre de travaux futurs.

Remerciements

Ces travaux ont été réalisés grâce au supercalculateur Factory-IA financé par le Conseil Régional d'Île-de-France.

Références

BRONSTEIN O., DAGAN I., LI Q., JI H. & FRANK A. (2015). Seed-Based Event Trigger Labeling : How far can event descriptions get us ? In *ACL-IJCNLP*, p. 372–376. DOI : 10.3115/v1/P15-2061.

CHEN J., LIN H., HAN X. & SUN L. (2021). Honey or Poison ? Solving the Trigger Curse in Few-shot Event Detection via Causal Intervention. In *Proceedings of EMNLP*, p. 8078–8088, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.637.

CONG X., CUI S., YU B., LIU T., YUBIN W. & WANG B. (2021). Few-Shot Event Detection with Prototypical Amortized Conditional Random Field. In *Findings of ACL-IJCNLP*, p. 28–40, Online. DOI : 10.18653/v1/2021.findings-acl.3.

DENG S., ZHANG N., KANG J., ZHANG Y., ZHANG W. & CHEN H. (2020). Meta-Learning with Dynamic-Memory-Based Prototypical Network for Few-Shot Event Detection. In *WSDM*, p. 151–159, Houston, TX, USA. DOI : 10.1145/3336191.3371796.

DROR R., SHLOMOV S. & REICHART R. (2019). Deep Dominance - How to Properly Compare Deep Neural Models. In *ACL*, p. 2773–2785, Florence, Italy. DOI : 10.18653/v1/P19-1266.

GENG R., LI B., LI Y., ZHU X., JIAN P. & SUN J. (2019). Induction Networks for Few-Shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3904–3913, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1403.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, p. 282–289, San Francisco, CA, USA.

LAI V., DERNONCOURT F. & NGUYEN T. H. (2021). Learning Prototype Representations Across Few-Shot Tasks for Event Detection. In *EMNLP*, p. 5270–5277.

LAI V. D. & NGUYEN T. (2019). Extending Event Detection to New Types with Learning from Keywords. In *W-NUT 2019*, p. 243–248, Hong Kong, China. DOI : 10.18653/v1/D19-5532.

LAI V. D., NGUYEN T. H. & DERNONCOURT F. (2020). Extensively Matching for Few-shot Learning Event Detection. In *Workshop NUSE*, p. 38–45, Online. DOI : 10.18653/v1/2020.nuse-1.5.

LI Q., JI H. & HUANG L. (2013). Joint Event Extraction via Structured Prediction with Global Features. In *ACL*, p. 73–82, Sofia, Bulgaria.

LIAO S. & GRISHMAN R. (2011). Acquiring topic features to improve event extraction : in pre-selected and balanced collections. In *RANLP* : Association for Computational Linguistics.

LIU X., LUO Z. & HUANG H. (2018). Jointly Multiple Events Extraction via Attention-based Graph Information Aggregation. In *EMNLP*, p. 1247–1256. DOI : 10.18653/v1/D18-1156.

NGUYEN T. H., CHO K. & GRISHMAN R. (2016). Joint Event Extraction via Recurrent Neural Networks. In *NAACL-HLT*, p. 300–309, San Diego, California. DOI : 10.18653/v1/N16-1034.

NGUYEN T. H. & GRISHMAN R. (2015a). Event Detection and Domain Adaptation with Convolutional Neural Networks. In *ACL-IJCNLP*, p. 365–371, Beijing, China. DOI : 10.3115/v1/P15-2060.

NGUYEN T. H. & GRISHMAN R. (2015b). Event Detection and Domain Adaptation with Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, p. 365–371, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-2060.

NGUYEN T. H. & GRISHMAN R. (2018). Graph Convolutional Networks with Argument-Aware Pooling for Event Detection. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

NIMAH I., FANG M., MENKOVSKI V. & PECHENIZKIY M. (2021). ProtoInfoMax : Prototypical Networks with Mutual Information Maximization for Out-of-Domain Detection. In *Findings of the Association for Computational Linguistics : EMNLP*, p. 1606–1617 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-emnlp.138.

SCHÖLKOPF B., PLATT J. C., SHAWE-TAYLOR J., SMOLA A. J. & WILLIAMSON R. C. (2001). Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, (7), 1443–1471. DOI : 10.1162/089976601750264965.

SHEN S., WU T., QI G., LI Y.-F., HAFFARI G. & BI S. (2021). Adaptive Knowledge-Enhanced Bayesian Meta-Learning for Few-shot Event Detection. In *Findings of ACL-IJCNLP*, p. 2417–2429, Online. DOI : 10.18653/v1/2021.findings-acl.214.

SNELL J., SWERSKY K. & ZEMEL R. (2017). Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30.

SUNG F., YANG Y., ZHANG L., XIANG T., TORR P. H. S. & HOSPEDALES T. M. (2018). Learning to compare : Relation network for few-shot learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 1199–1208.

TAN M., YU Y., WANG H., WANG D., POTDAR S., CHANG S. & YU M. (2019). Out-of-Domain Detection for Low-Resource Text Classification Tasks. In *EMNLP-IJCNLP*, p. 3566–3572 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1364.

TUO A., BESANÇON R., FERRET O. & TOURILLE J. (2022). Better Exploiting BERT for Few-Shot Event Detection. In *NLDB*, p. 291–298, Berlin, Heidelberg : Springer-Verlag. DOI : 10.1007/978-3-031-08473-7_26.

TUO A., BESANÇON R., FERRET O. & TOURILLE J. (2023). Trigger or not trigger : Dynamic thresholding for few shot event detection. In J. KAMPS, L. GOEURIOT, F. CRESTANI, M. MAISTRO, H. JOHO, B. DAVIS, C. GURRIN, U. KRUSCHWITZ & A. CAPUTO, Éds., *45th European Conference on Information Retrieval (ECIR 2023) : Advances in Information Retrieval, short article session*, volume 13981 de *Lecture Notes in Computer Science*, p. 637–645, Dublin, Ireland : Springer Nature Switzerland.

VINYALS O., BLUNDELL C., LILLICRAP T., KAVUKCUOGLU K. & WIERSTRA D. (2016). Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29.

WALKER C., STRASSEL S. & JULIE MEDERO K. M. (2006). ACE 2005 Multilingual Training Corpus. DOI : 10.35111/mwxc-vh88.

WANG X., WANG Z., HAN X., JIANG W., HAN R., LIU Z., LI J., LI P., LIN Y. & ZHOU J. (2020). MAVEN : A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1652–1671, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.129.

YAN H., JIN X., MENG X., GUO J. & CHENG X. (2019). Event detection with multi-order graph convolution and aggregated attention. In *EMNLP-IJCNLP*, p. 5766–5770.

YANG Y. & KATIYAR A. (2020). Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *EMNLP*, p. 6365–6375 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.516.

ZHANG R., WEI W., MAO X.-L., FANG R. & CHEN D. (2022). HCL-TAT : A Hybrid Contrastive Learning Method for Few-shot Event Detection with Task-Adaptive Threshold. In *Findings of the Association for Computational Linguistics : EMNLP*, p. 1808–1819 : Association for Computational Linguistics.

Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Sélection globale de segments pour la reconnaissance d'entités nommées

Urchade Zaratiana*†, Niama El Khbir†, Pierre Holat*†,
Nadi Tomeh†, Thierry Charnois†
* FI Group, † LIPN, CNRS UMR 7030, France
{zaratiana,elkhbir,holat,tomeh,charnois}@lipn.fr

RÉSUMÉ

La reconnaissance d'entités nommées est une tâche importante en traitement automatique du langage naturel avec des applications dans de nombreux domaines. Dans cet article, nous décrivons une nouvelle approche pour la reconnaissance d'entités nommées, dans laquelle nous produisons un ensemble de segmentations en maximisant un score global. Pendant l'entraînement, nous optimisons notre modèle en maximisant la probabilité de la segmentation correcte. Pendant l'inférence, nous utilisons la programmation dynamique pour sélectionner la meilleure segmentation avec une complexité linéaire. Nous prouvons que notre approche est supérieure aux modèles champs de Markov conditionnels et Semi-CRF pour la reconnaissance d'entités nommées.

ABSTRACT

Global Span Selection for Named Entity Recognition

Named Entity Recognition is an important task in Natural Language Processing with applications in many domains. In this paper, we describe a novel approach to named entity recognition, in which we output a set of spans (i.e., segmentations) by maximizing a global score. During training, we optimize our model by maximizing the probability of the gold segmentation. During inference, we use dynamic programming to select the best segmentation under a linear time complexity. We prove that our approach outperforms CRF and semi-CRF models for Named Entity Recognition.

MOTS-CLÉS : Reconaissance d'entités nommées, segmentation, Champ aléatoire conditionnel.

KEYWORDS: Named entity recognition, segmentation, Conditional Random Fields.

1 Introduction

La reconnaissance d'entités nommées (REN) est une tâche cruciale du traitement du langage naturel dont le but est d'identifier et de classer les entités pertinentes dans les textes telles que les personnes, les organisations et les lieux. La reconnaissance de telles entités est avantageuse pour des applications telles que l'extraction de relations (El Khbir *et al.*, 2022) et la construction de taxanomie (Dauxais *et al.*, 2022). Il existe deux paradigmes principaux pour la reconaissance d'entités : l'étiquetage de séquences (ES) (Huang *et al.*, 2015; Lample *et al.*, 2016; Akbik *et al.*, 2018) et les approches basées sur les segments (ABS) (Sohrab & Miwa, 2018; Yu *et al.*, 2020a; Li *et al.*, 2021). L'ES considère la reconaissance d'entités comme une prédiction au niveau du jeton, en utilisant par exemple les schémas BIO (Ramshaw & Marcus, 1995) ou BILOU (Ratinov & Roth, 2009), tandis que les ABS

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 169–177.

considèrent les segments (segments contigus de jetons) comme des unités de base au lieu des jetons et effectuent une classification au niveau du segment en attribuant une étiquette à chaque entité et une étiquette spéciale `null` aux segments sans entités (aussi segments non-entités).

L'ES est généralement réalisé en représentant les jetons à l'aide de modèles d'apprentissage profond, puis en utilisant un champ aléatoire conditionnel (Lafferty *et al.*, 2001) comme couche de sortie. La meilleure séquence d'étiquettes est calculée à l'aide de l'algorithme de Viterbi et l'apprentissage maximise typiquement la vraisemblance des séquences de référence. En revanche, les ABS énumèrent tous les segments candidats d'un texte d'entrée et calculent leur représentation avant de les fournir à une couche softmax pour la classification. L'un des avantages des ABS est qu'elles permettent une représentation plus riche des segments en comparaison à l'ES, puisque les caractéristiques au niveau des segments sont apprises de bout en bout.

Cependant, ces modèles basés sur les segments *unstructurés* prédisent l'étiquette de chaque segment indépendamment. Ils ont tendance à produire des entités qui se chevauchent, ce qui est interdit dans la reconaissance d'entités plate et imbriquée. Les travaux antérieurs utilisaient un algorithme de décodage (Johnson, 1973; Yu *et al.*, 2020b; Li *et al.*, 2021) pour obtenir un ensemble d'entités non chevauchantes. Les entités ayant obtenu le meilleur score sont sélectionnées de manière itérative tant qu'elles ne chevauchent pas avec celles sélectionnées précédemment. Le décodage glouton est efficace mais tend à souffrir d'un biais de myopie. Le fait de choisir des segments sans tenir compte des décisions futures peut conduire à des ensembles d'entités sous-optimaux.

Une formulation alternative de la REN sous forme de segmentation et d'étiquetage joints avec des semi-champs de Markov conditionnels (Semi-CRF) a été proposée dans la littérature : (Sarawagi & Cohen, 2005; Kong *et al.*, 2016; Ye & Ling, 2018). Cette approche présente deux avantages : (a) elle utilise un modèle globalement normalisé pour calculer la probabilité de chaque segmentation étiquetée, au lieu d'évaluer chaque segment indépendamment ; et (2) elle garantit le non-chevauchement des entités de sortie en utilisant une variante de l'algorithme de Viterbi pour le décodage.

Néanmoins, les Semi-CRF sont moins performants en pratique, comme nous le montrons dans nos expériences. Nous supposons que l'évaluation de segmentations composées d'entités et de non-entités à la fois est la principale faiblesse. Tout d'abord, les segments sans entités peuvent être segmentés de plusieurs façons, toutes aussi valides les unes que les autres, mais une seule d'entre elles est appliquée par les Semi-CRF, à la fois pendant l'apprentissage et l'inférence. De plus, la majorité des segments ne comportant pas d'entité, une masse considérable de probabilité est gaspillée sur des segmentations inintéressantes.

Dans cet article, nous proposons une nouvelle formulation pour la REN basée sur les segments qui combine des idées provenant d'approches en deux étapes (filtrage et décodage) et de modèles basés sur des champs de Markov conditionnels (CRF) globalement normalisés. Notre approche commence par le filtrage de tous les segments sans entités à l'aide d'un classificateur de segments et la construction d'un graphe (*chevauchant*) des segments restants. Un modèle globalement normalisé est ensuite utilisé pour calculer la probabilité de chaque *ensemble indépendant maximal (EIM)* dans le graphe. Chacun de ces ensembles correspond à une sélection d'entités non chevauchantes. L'apprentissage et l'inférence peuvent être réalisés efficacement en utilisant la programmation dynamique, comme nous l'expliquons dans la section 2.2. De plus, nous entraînons le classificateur de segments et le modèle global de sélection d'entités de manière jointe en utilisant un objectif multi-tâches. Nous montrons que notre approche surpasse à la fois l'ES et les Semi-CRF sur toutes les tâches et surpasse les modèles sur les deux étapes du filtrage et du décodage glouton dans la plupart des cas.

2 La REN basée sur les segments en deux étapes

Les approches de la littérature basées sur les segments utilisent un classificateur de segments non structuré et localement normalisé pour filtrer les segments non-entités, suivi d'un décodage glouton pour sélectionner un ensemble d'entités non chevauchantes (Li *et al.*, 2021; Fu *et al.*, 2021). Nous décrivons ces deux étapes dans cette section.

2.1 La classification de segments

Cette étape consiste à énumérer tous les segments de la séquence d'entrée et à calculer leur représentations en utilisant des transformeurs pré-entraînés tels que BERT. Conformément aux travaux antérieurs (Lee *et al.*, 2017; Luan *et al.*, 2019), la représentation s_{ij} d'un segment (i, j) de longueur k est calculée en concaténant la représentation de ses jetons d'extrémité gauche et droite (h_i et h_j respectivement) avec une caractéristique apprise de largeur du segment f_k. Un perceptron multicouche à deux couches avec activation *ReLU* est appliqué aux caractéristiques pour obtenir la représentation finale du segment

$$s_{ij} = \texttt{MLP}([h_i; h_j; f_k]) \tag{1}$$

Ensuite, la représentation du segment est introduite dans une couche linéaire (ou un perceptron multicouche) pour la classification du segment. Une tâche de REN avec L types d'entités aurait $L + 1$ étiquettes puisque nous allouons une étiquette `null` pour les segments non-entités. Le score de l'étiquette y pour un segment (i, j) est calculée comme suit :

$$\phi(i, j, y) = w_y^T s_{ij} \tag{2}$$

où w_y est un vecteur de poids apprenable (nous omettons le terme de biais pour des raisons de lisibilité). Ces scores sont ensuite normalisés à l'aide de la fonction softmax.

Le modèle est entraîné pour minimiser la log-vraisemblance négative des segments de référence de l'ensemble d'entraînement $\mathcal{T}$:

$$\mathcal{L}_{clf} = -\sum_{(i,j,y)\in\mathcal{T}} \log \frac{\exp \phi(i, j, y)}{\sum_{y'} \exp \phi(i, j, y')} \tag{3}$$

Pendant l'inférence, chaque segment (i, j) se voit attribuer l'étiquette $y(i, j) =_y \phi(i, j, y)$ avec le score $k(i, j) = \max_y \phi(i, j, y)$. Nous appelons $\mathcal{C}$ l'ensemble des entités candidates qui est l'ensemble de tous les segments auxquels on a attribué une étiquette différente de `null`. Cet ensemble peut contenir des segments qui se chevauchent, ce qui n'est pas autorisé dans les tâches de reconaissance plate, une étape de décodage est donc nécessaire.

2.2 Ensemble indépendant à poids maximal dans les graphes d'intervalles

Un *graphe de chevauchement* sur $\mathcal{C}$ est le graphe G dont les nœuds sont les éléments de $\mathcal{C}$ et qui contient une arête entre chaque paire d'entités chevauchantes. Ce graphe peut également être appelé un *graphe d'intervalles* puisque les segments peuvent être vues comme des intervalles de leurs positions de début et de fin. Un *ensemble indépendant (EI)* du graphe G est un ensemble de nœuds tel que deux nœuds de cet ensemble ne sont pas reliés par une arête. Un ensemble indépendant est dit *maximal* s'il n'est pas correctement contenu dans un autre ensemble indépendant. Chaque nœud (i, j)

du graphe se voit attribuer un nombre réel $r(i,j)$, le graphe G est dit être un graphe *pondéré*. Pour chaque sous-ensemble de nœuds $\mathcal{S} \subseteq \mathcal{C}$, $\sum_{(i,j)\in\mathcal{S}} r(i,j)$ est appelé le poids de $\mathcal{S}$. Un *Ensemble indépendant de poids maximal (EIPM)* est un ensemble indépendant tel que son poids est maximal sur tous les ensembles indépendants. Sous cette formulation, le problème du décodage revient à trouver un EIPM dans le graphe G :

$$\hat{\mathcal{S}} = argmax_{\mathcal{S}\in\Psi(\mathcal{C})} \sum_{(i,j)\in\mathcal{S}} r(i,j) \tag{4}$$

où $\Psi(\mathcal{C})$ est l'ensemble de tous les EIM de G.

Le décodage glouton Le décodage glouton construit une approximation de $\hat{\mathcal{S}}$ en ajoutant itérativement l'entité de C ayant le meilleur score et qui ne se chevauche avec aucune entité précédemment sélectionnée. Cet algorithme est d'une complexité de $O(n \log n)$ avec $n = |C|$.

Dans la section suivante, nous proposons une alternative exacte qui utilise un modèle globalement normalisé.

Le decodage exact La solution exacte de l'équation (4) peut être obtenue par la programmation dynamique en utilisant l'algorithme EIPM présenté par Gupta *et al.* (1982); Hsiao *et al.* (1992). Cet algorithme a une complexité temporelle linéaire en $O(n)$ avec n le nombre de noeuds dans le graphe, qui est supposé être trié par les extrémités d'intervalles, sinon, il peut être trié en un temps de $O(n \log n)$. En pratique, le nombre de nœuds n est bien inférieur à la longueur de la séquence d'entrée.

2.3 Un modèle EIPM globalement normalisé

Une façon d'estimer les poids $r(i,j)$ des nœuds du graphe est d'utiliser les scores produits par les classificateurs locaux : $r(i,j) = k(i,j)$. Dans cette section, nous proposons d'apprendre un modèle probabiliste dédié du EIM globalement normalisé et entraîné pour maximiser la probabilité du EIM de référence.

La probabilité d'un EIM est calculée par :

$$P(\mathcal{S}) = \mathcal{Z}^{-1} \exp \sum_{(i,j)\in\mathcal{S}} r(i,j) \tag{5}$$

Le score non normalisé d'un EIM est simplement égal à la somme des poids individuels des segments, chacune étant une projection linéaire de la représentation du segment :

$$r(i,j) = w^T s_{ij} \tag{6}$$

où w est un vecteur de paramètres à apprendre. La constante de normalisation est donnée par :

$$\mathcal{Z} = \sum_{\mathcal{S}\in\Psi(\mathcal{C})} \exp \sum_{(i,j)\in\mathcal{S}} r(i,j) \tag{7}$$

Alors que $\mathcal{Z}$, la fonction de partition, peut être ignorée pendant l'inférence, elle doit être calculée pendant l'apprentissage car nous utilisons la log probabilité négative du EIM or comme fonction de perte. La fonction de partition peut être calculée efficacement en utilisant une modification au programme dynamique de l'algorithme EIPM. Cependant, en pratique, nous énumérons simplement tous les EIM, ce qui est faisable puisque le nombre de segments restants est faible. L'énumération peut

Modèles	Conll-2003			OntoNotes 5.0			TDM			ACE Arabe		
	P	R	F	P	R	F	P	R	F	P	R	F
CRF	92.64	91.82	92.23	87.77	89.47	88.61	69.77	73.65	71.66	82.79	84.44	83.61
Semi-CRF	91.46	90.77	91.11	87.44	88.85	88.14	69.38	72.85	71.05	82.97	84.24	83.60
Standard	93.40	91.68	92.53	89.47	90.00	89.73	67.75	69.88	68.78	83.21	83.76	83.48
+ Glouton	93.82	91.40	92.60	90.43	89.04	89.73	75.12	67.82	71.26	83.73	83.56	83.64
+ Global	93.83	91.51	92.65	90.58	89.45	**90.01**	75.25	68.12	71.48	83.72	83.55	83.63
Global	94.84	90.72	92.73	89.05	89.77	89.41	63.30	72.75	67.53	83.54	83.65	83.60
+ Glouton	95.07	90.42	92.69	89.98	88.44	89.21	74.16	68.23	71.07	83.87	82.75	83.31
+ Global	95.11	90.52	**92.76**	90.18	88.85	89.51	75.55	70.34	**72.84**	84.14	83.35	**83.74**

TABLE 1 – **Résultats expérimentaux**. Nous rapportons la moyenne sur trois échantillons aléatoires.

être effectuée en un temps $O(n^2 + \beta)$ où n est le nombre de segments et β la somme des nombres de segments de tous les ensembles énumérés (Leung, 1984; Liang *et al.*, 1991).

Pendant l'apprentissage, nous modifions l'ensemble $\mathcal{C}$, c'est-à-dire la sortie du classificateur local, de sorte (1) qu'il contienne tous les segments de référence, et (2) qu'il ne contienne pas de segments ne se chevauchant pas avec les segments de référence . Ce faisant, nous nous assurons que les intervalles or forment un EIM dans le graphe de chevauchement sur $\mathcal{C}$. Enfin, nous utilisons une fonction de perte multitâche qui correspond à la somme de la perte du classificateur local (Eq. (3)) et de la perte du modèle global.

3 Expériences

3.1 Configuration

Baselines Nous comparons notre approche à un étiqueteur CRF, au modèle standard ABS et au modèle basé sur les segments avec Semi-CRF. Pour tous les modèles, nous utilisons des transformeurs pré-entraînés pour la représentation des jetons.

Jeux de données Nous évaluons notre modèle sur divers jeux de données de REN : TDM (Hou *et al.*, 2021), Conll-2003 (Tjong Kim Sang & De Meulder, 2003), et OntoNotes 5.0 (Weischedel *et al.*, 2013) pour les données anglaises, et ACE05 pour les données arabes (Walker *et al.*, 2006).

Mesures d'évaluation Nous évaluons les modèles en utilisant la correspondance exacte entre les entités prédites et les entités de référence. Nous rapportons la précision, le rappel et le F1.

Hyperparamètres Pour les jeux de données Conll-2003 et Ontonotes, nous utilisons `bert-base-cased`. (Devlin *et al.*, 2019) pour produire la représentation contextuelle, pour TDM nous utilisons SciBERT (Beltagy *et al.*, 2019) et pour ACE Arabe nous utilisons `bert-base-arabertv2` (Antoun *et al.*, 2020). Nous utilisons la taille de base, avec 12 couches de transformeurs, pour tous les modèles. Nous n'utilisons pas d'incorporations auxiliaires (par exemple l'*incorporation de caractères*) pour des raisons de simplicité. Tous les modèles sont entraînés avec l'optimiseur de Kingma & Ba (2017) avec un taux d'apprentissage de 2e-5, une taille de lot de 10 et une époque maximale de 25. Nous conservons le meilleur point de contrôle sur l'ensemble de validation pour les tests. Nous avons entraîné tous les modèles sur un serveur équipé de GPU V100.

3.2 Résultats

Nous rapportons les résultats de nos expériences dans la Table 1 pour les quatre ensembles de données en utilisant les modèles CRF, Semi-CRF, Standard et Global basés sur les segments. Pour les modèles Standard et Global, nous rapportons les résultats obtenus en utilisant (cf. + lignes Global) ou en n'utilisant pas le décodage (cf. + lignes Glouton).

Résultats principaux D'après la Table 1, nos modèles globaux avec décodage global obtiennent les meilleurs résultats sur la plupart des ensembles de données (tous sauf sur OntoNotes). De plus, le Semi-CRF a le score le plus bas sur toutes les données, ce qui peut expliquer sa faible adoption au fil des ans par rapport au CRF standard.

Décodage global vs. décodage glouton Pour les deux approches basées sur les segments, nous constatons que le décodage améliore généralement la performance du score F1 et la précision tout en diminuant le rappel. Nous pouvons expliquer ce comportement par le fait que quand on utilise le décodage, les segments non fiables sont supprimés, ce qui augmente la précision. Cependant, certains faux négatifs peuvent également être supprimés, d'où la légère diminution du rappel. De plus, pour les modèles standards, le décodage glouton et le décodage global ont des performances similaires, tandis que pour les modèles entraînés globalement, le décodage global a toujours les meilleures performances, ce qui montre l'efficacité de notre approche. De plus, nous pouvons observer sur les jeux de données Conll-2003, ACE Arabe et OntoNotes 5.0 que le décodage glouton peut même diminuer la performance du modèle, ce qui serait un effet du biais myopique.

4 Travaux connexes

Approches pour la REN Traditionnellement, les tâches de REN sont conçues comme un ES (Lample *et al.*, 2016; Akbik *et al.*, 2018), c'est-à-dire une classification au niveau du jeton. Récemment, de nombreuses approches ont été proposées qui vont au-delà de la prédiction au niveau du token. Par exemple, certains travaux ont abordé la REN nommées comme une tâche de réponse aux questions (Li *et al.*, 2020) et d'autres utilisent des modèles de séquence-à-séquence (Yan *et al.*, 2021; Yang & Tu, 2022). Dans ce travail, nous nous sommes concentrés sur les ABS (Liu *et al.*, 2016; Sohrab & Miwa, 2018; Fu *et al.*, 2021; Zaratiana *et al.*, 2022b,c,a).

Décodage pour la REN La REN est une tâche pour laquelle un algorithme de décodage doit être appliqué afin de garantir que les sorties du modèle sont bien entraînées. Par exemple, les CRF (Lafferty *et al.*, 2001) ont été proposés pour l'ES et le Semi-CRF pour les ABS. En raison de la faible performance du Semi-CRF (Sarawagi & Cohen, 2005), les chercheurs ont proposé d'entraîner une méthode locale basée sur les segments et d'utiliser un décodage glouton pour garantir des entités non chevauchantes pour le décodage. Dans ce travail, nous proposons un décodage exact/global pour produire un ensemble de segments non chevauchants qui maximisent le score global afin d'éviter le biais myope de l'approche gloutonne.

5 Conclusion

Dans ce travail, nous proposons une nouvelle approche de reconnaissance d'entités nommées basée sur les segments. Notre modèle atténue le biais myope de l'approche standard et obtient de meilleurs résultats que les modèles structurés tels que les CRF ou les Semi-CRF. Des travaux futurs pourraient explorer l'interaction entre les segments pour améliorer davantage les performances.

Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011013096R1 attribuée par GENCI.

Références

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, p. 1638–1649, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ANTOUN W., BALY F. & HAJJ H. M. (2020). Arabert : Transformer-based model for arabic language understanding. *ArXiv*, **abs/2003.00104**.

BELTAGY I., LO K. & COHAN A. (2019). Scibert : A pretrained language model for scientific text.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

DAUXAIS Y., ZARATIANA U., LANEUVILLE M., HERNANDEZ S. D., HOLAT P. & GROSMAN C. (2022). Towards automation of topic taxonomy construction. In T. BOUADI, E. FROMONT & E. HÜLLERMEIER, Éds., *Advances in Intelligent Data Analysis XX*, p. 26–38, Cham : Springer International Publishing.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

EL KHBIR N., TOMEH N. & CHARNOIS T. (2022). ArabIE : Joint entity, relation and event extraction for Arabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, p. 331–345, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

FU J., HUANG X. & LIU P. (2021). SpanNER : Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 7183–7195, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.558.

GUPTA U. I., LEE D. T. & LEUNG J. Y.-T. (1982). Efficient algorithms for interval graphs and circular-arc graphs. *Networks*, **12**, 459–467.

HOU Y., JOCHIM C., GLEIZE M., BONIN F. & GANGULY D. (2021). TDMSci : A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 707–714, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.59.

HSIAO J. Y., TANG C. Y. & CHANG R. S. (1992). An efficient algorithm for finding a maximum weight 2-independent set on interval graphs. *Information Processing Letters*, **43**(5), 229–235. DOI : https ://doi.org/10.1016/0020-0190(92)90216-I.

HUANG Z., XU W. & YU K. (2015). Bidirectional lstm-crf models for sequence tagging.

JOHNSON D. S. (1973). Approximation algorithms for combinatorial problems. *Proceedings of the fifth annual ACM symposium on Theory of computing*.

KINGMA D. P. & BA J. (2017). Adam : A method for stochastic optimization.

KONG L., DYER C. & SMITH N. A. (2016). Segmental recurrent neural networks. *CoRR*, **abs/1511.06018**.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/N16-1030.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LEE K., HE L., LEWIS M. & ZETTLEMOYER L. (2017). End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 188–197, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1018.

LEUNG J. Y.-T. (1984). Fast algorithms for generating all maximal independent sets of interval, circular-arc and chordal graphs. *J. Algorithms*, **5**, 22–35.

LI X., FENG J., MENG Y., HAN Q., WU F. & LI J. (2020). A unified mrc framework for named entity recognition.

LI Y., LEMAO LIU & SHI S. (2021). Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*.

LIANG Y., DHALL S. & LAKSHMIVARAHAN S. (1991). On the problem of finding all maximum weight independent sets in interval and circular-arc graphs. In *[Proceedings] 1991 Symposium on Applied Computing*, p. 465–470. DOI : 10.1109/SOAC.1991.143921.

LIU Y., CHE W., GUO J., QIN B. & LIU T. (2016). Exploring segment representations for neural segmentation models. In *IJCAI*.

LUAN Y., WADDEN D., HE L., SHAH A., OSTENDORF M. & HAJISHIRZI H. (2019). A general framework for information extraction using dynamic span graphs.

RAMSHAW L. A. & MARCUS M. P. (1995). Text chunking using transformation-based learning. *ArXiv*, **cmp-lg/9505040**.

RATINOV L. & ROTH D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, p. 147–155, Boulder, Colorado : Association for Computational Linguistics.

SARAWAGI S. & COHEN W. W. (2005). Semi-markov conditional random fields for information extraction. In L. SAUL, Y. WEISS & L. BOTTOU, Éds., *Advances in Neural Information Processing Systems*, volume 17 : MIT Press.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SOHRAB M. G. & MIWA M. (2018). Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2843–2849, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1309.

TJONG KIM SANG E. F. & DE MEULDER F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, p. 142–147.

WALKER C., STRASSEL S., MEDERO J. & MAEDA K. (2006). Ace 2005 multilingual training corpus. DOI : 10.35111/MWXC-VH88.

WEISCHEDEL R., PALMER M., MARCUS M., HOVY E., PRADHAN S., RAMSHAW L., XUE N., TAYLOR A., KAUFMAN J., FRANCHINI M., EL-BACHOUTI M., BELVIN R. & HOUSTON A. (2013). OntoNotes Release 5.0. DOI : 11272.1/AB2/MKJJ2R.

YAN H., GUI T., DAI J., GUO Q., ZHANG Z. & QIU X. (2021). A unified generative framework for various ner subtasks. In *ACL*.

YANG S. & TU K. (2022). Bottom-up constituency parsing and nested named entity recognition with pointer networks. In *ACL*.

YE Z. & LING Z.-H. (2018). Hybrid semi-Markov CRF for neural sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 235–240, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-2038.

YU J., BOHNET B. & POESIO M. (2020a). Named entity recognition as dependency parsing. In *ACL*.

YU J., BOHNET B. & POESIO M. (2020b). Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6470–6476, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.577.

ZARATIANA U., ELKHBIR N., HOLAT P., TOMEH N. & CHARNOIS T. (2022a). Global span selection for named entity recognition. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, p. 11–17, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2022b). GNNer : Reducing overlapping in span-based NER using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*, p. 97–103, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-srw.9.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2022c). Named entity recognition as structured span prediction. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, p. 1–10, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.