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 2 : travaux de recherche originaux – articles courts

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

Positionnement temporel indépendant des évènements : application à des textes cliniques en français

Nesrine Bannour[1] Xavier Tannier[2] Bastien Rance[3,4,5] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France
(2) Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, Paris, 75006, France
(3) Inserm, Centre de Recherche des Cordeliers, UMRS 1138, Université Paris Cité, Sorbonne Paris Cité, Paris 75006, France
(4) Assitance Publique - Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Paris 75015, France
(5) HeKA, Inria Paris, Paris 75006, France
nesine.bannour@lisn.upsaclay.fr, xavier.tannier@sorbonne-universite.fr, bastien.rance@aphp.fr, aurelie.neveol@lisn.upsaclay.fr

RÉSUMÉ

L'extraction de relations temporelles consiste à identifier et classifier la relation entre deux mentions. Néanmoins, la définition des mentions temporelles dépend largement du type du texte et du domaine d'application. En particulier, le texte clinique est complexe car il décrit des évènements se produisant à des moments différents et contient des informations redondantes et diverses expressions temporelles spécifiques au domaine. Dans cet article, nous proposons une nouvelle représentation des relations temporelles, qui est indépendante du domaine et de l'objectif de la tâche d'extraction. Nous nous intéressons à extraire la relation entre chaque portion du texte et la date de création du document. Nous formulons l'extraction de relations temporelles comme une tâche d'étiquetage de séquences. Une macro F-mesure de 0,8 est obtenue par un modèle neuronal entraîné sur des textes cliniques, écrits en français. Nous évaluons notre représentation temporelle par le positionnement temporel des évènements de toxicité des chimiothérapies.

ABSTRACT

Event-independent temporal positioning : application to French clinical texts.

The extraction of temporal relations entails identifying and classifying the relation between two mentions. However, the definition of temporal mentions is strongly dependent on the type of text and the application domain. Clinical text, in particular, is complex, describing events that occurred at different times, containing redundant information, and a variety of domain-specific temporal expressions. In this paper, we propose a new representation of temporal relations, which is independent of the domain and the goal of the extraction task. We are interested in extracting the relation between each text portion and the document creation time. Temporal relation extraction is cast as a sequence labeling task. An overall macro F-measure of 0.8 is obtained by a neural model trained on clinical texts written in French. We evaluate our temporal representation by the temporal positioning of chemotherapy toxicity events.

MOTS-CLÉS : Texte clinique, Extraction d'informations temporelles, étiquetage de séquences.

KEYWORDS: Clinical Text, Temporal Information Extraction, Sequence Labeling.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 1–14.

1 Introduction

L'extraction d'informations temporelles est une tâche cruciale pour la compréhension de textes en langage naturel. Cette tâche a été utilisée dans de nombreuses applications de traitement du langage naturel, dont la reconstitution de récits (Do *et al.*, 2012; Ning *et al.*, 2017; Han *et al.*, 2019) et le traitement de textes cliniques (Tourille *et al.*, 2017; Moharasan & Ho, 2019; Lin *et al.*, 2020). Une attention particulière a été portée à l'extraction d'informations temporelles dans les textes cliniques par le biais des campagnes d'évaluation i2b2-2012 (Sun *et al.*, 2013) et Clinical TempEval (Bethard *et al.*, 2015, 2016, 2017). Les informations temporelles contenues dans les textes cliniques des Dossiers Patients Informatisés (DPIs) permettent de mieux comprendre divers évènements cliniques tels que la progression de la maladie et les effets longitudinaux des médicaments (Lin *et al.*, 2016). Le corpus de la campagne i2b2-2012 a été annoté en se basant sur les schémas d'annotations TimeML (Pustejovsky *et al.*, 2003) et THYME-TimeML (Styler IV *et al.*, 2014). THYME-TimeML a été développé pour annoter l'évolution temporelle des évènements cliniques dans le corpus THYME. Ce schéma a été utilisé pour l'annotation du corpus de la campagne Clinical TempEval.

L'extraction d'informations temporelles a d'abord été représentée comme l'extraction des relations TLINKs, reliant les évènements (EVENT) à un ancrage temporel représenté par des expressions temporelles (TIMEX). Pour le domaine clinique, des relations DocTimeRel entre des évènements médicaux et la date du création du document (Document Creation Time, DCT) ont été introduites (Pustejovsky & Stubbs, 2011; Styler IV *et al.*, 2014). Plusieurs défis sont associés à la représentation des informations temporelles cliniques (Najafabadipour *et al.*, 2020). En effet, les expressions temporelles sont diverses et incluent des dates spécifiques au domaine, non standards et abrégées. Le texte clinique étant généralement non-standard, il est également difficile d'associer des évènements à des expressions temporelles. Dans certains cas, le moment associé à un évènement n'est même pas explicitement mentionné. De plus, le texte narratif va et vient dans le temps en décrivant des évènements survenus à des moments différents, et peut contenir des informations redondantes provenant de rapports cliniques antérieurs, ce qui rend difficile l'identification de l'ordre chronologique des évènements.

Plusieurs études se sont intéressées à extraire les relations DocTimeRel (Tourille *et al.*, 2017; Viani *et al.*, 2019; Lin *et al.*, 2020; Alfattni *et al.*, 2021) et les relations TLINKs (Alfattni *et al.*, 2020; Tourille *et al.*, 2016; Liu *et al.*, 2019; Lin *et al.*, 2020). Néanmoins, les performances de la plupart des systèmes proposés sont loin d'être suffisantes pour des applications pratiques (Gumiel *et al.*, 2021). D'une part, outre l'aspect long et coûteux du processus d'annotation, l'annotation des relations temporelles cliniques est beaucoup plus difficile car cela nécessite une expertise médicale. D'autre part, la définition du schéma d'annotation dépend principalement du type de texte clinique et de l'objectif de la tâche d'extraction. La plupart des travaux de recherche portent aussi sur des corpus écrits en anglais. Peu d'études utilisent les corpus en français (Tourille *et al.*, 2016, 2017).

Dans cet article, nous nous intéressons à extraire le positionnement temporel des mentions en fonction de la DCT. Contrairement à l'extraction des relations DocTimeRel, nous ne commençons pas par l'extraction des évènements cliniques pour ensuite les classifier selon la DCT. Au lieu de cela, nous souhaitons l'extraction de la relation temporelle de chaque portion du texte avec la DCT, indépendemment des évènements. Nous pouvons procéder par la suite à l'extraction des évènements et chaque évènement aura alors la même relation temporelle que la portion du texte qui le contient. Voici les principales contributions de cet article :

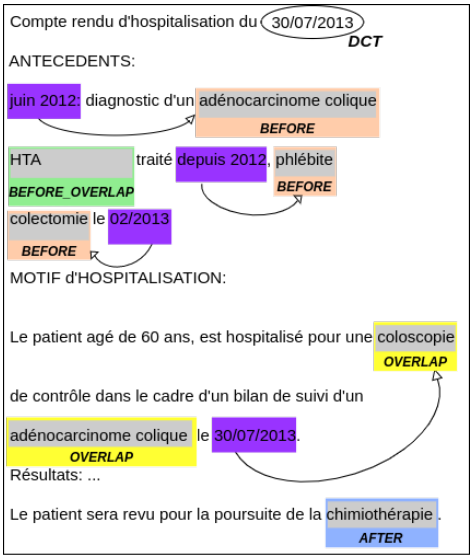
— Nous introduisons une nouvelle représentation des relations temporelles qui nous permet d'identifier des portions de textes homogènes du point de vue temporel et de caractériser ce

positionnement temporel, indépendamment du domaine et de la tâche de l'extraction.
— Pour évaluer notre représentation, nous annotons un corpus de textes cliniques, écrits en français en utilisant le schéma d'annotation THYME-TimeML et nous modélisons la tâche d'extraction de ces relations temporelles en tant qu'une tâche de classification de séquences. Le modèle de classification est comparé à un modèle *Baseline* à base de règles.
— Pour valider l'efficacité de notre représentation des relations temporelles, nous appliquons notre modèle de positionnement temporel sur un autre corpus clinique, nous identifions les évènements de toxicité des chimiothérapies dans ce corpus et nous inférons ensuite la relation temporelle de chaque évènement par rapport à la DCT.

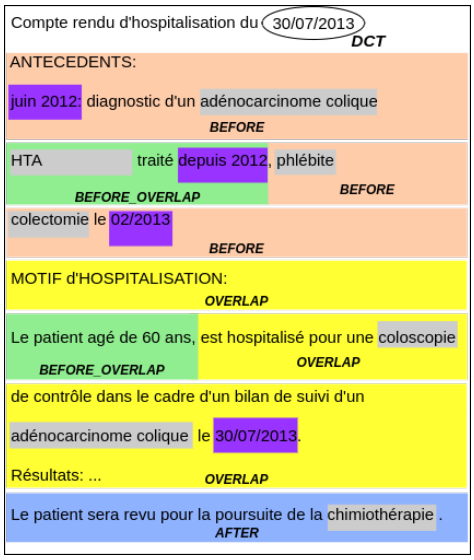
2 Méthodologie

Représentation des relations temporelles. Comme illustré dans la Figure 1a, les relations temporelles sont souvent représentées par les relations DocTimeRel et les relations TLINKs. L'extraction des DocTimeRel revient à l'identification des évènements et à la classification de leur relation temporelle avec la date de création du document. Chaque évènement sera associé à une catégorie selon le schéma THYME-TimeML (Figure 1) : *Before (orange)*, *Before_Overlap (vert)*, *Overlap (jaune)* et *After (bleu)*. Cependant, la définition des événements étant très spécifique au domaine, la tâche d'extraction des DocTimeRel diffère selon le domaine et aucune généralisation n'est possible. Des défis supplémentaires sont également rencontrés dans la définition des évènements cliniques dus à la complexité et à la variété des terminologies médicales présentes dans le texte clinique. L'extraction des relations TLINKs revient dans une première étape à extraire les paires possibles des évènements et des expressions temporelles. La stratégie la plus adoptée consiste à choisir les paires dans la même phrase et d'extraire les relations temporelles intra-phrastiques. Or, les spécificités des textes cliniques présentent des difficultés pour l'identification des frontières de phrase comme l'usage des termes contenant des marques de ponctuation et l'oubli des marqueurs de début et de fin de phrases. Par ailleurs, si l'évènement et l'expression temporelle sont dans des phrases différentes, d'autres stratégies doivent être adoptées pour résoudre les dépendances à longue distance. De ce fait, nous introduisons une nouvelle représentation des relations temporelles qui est indépendante des évènements. Comme illustré dans la Figure 1b, les portions de texte homogènes du point de vue temporel sont extraites et associées à une catégorie du schéma d'annotation THYME-TimeML qui reflète la relation avec la DCT. Les évènements auront par la suite la même catégorie que la portion qui les inclut. Ainsi, nous n'avons pas à gérer les problèmes de frontières de phrases ni la problématique de dépendance longue. Certes, cette représentation est moins fine que la représentation classique des informations temporelles, mais il s'agit d'une représentation qui est totalement indépendante du type des mentions à extraire et donc du domaine d'application.

Extraction et classification des relations temporelles. Nous modélisons la tâche d'extraction des relations temporelles comme une tâche de classification supervisée de séquences. Notre but est de classifier chaque portion du texte en une catégorie pré-définie. Pour ceci, nous entraînons un modèle de classification de tokens, en utilisant le modèle français CamemBERT (Martin *et al.*, 2020) de la librairie *transformers* de HuggingFace (Wolf *et al.*, 2020). Nous classifions chaque token en nous appuyant sur le format BIO (Beginning-Inside-Outside). De cette manière, le modèle sera capable de détecter les tokens qui marquent le changement temporel dans le texte. Les poids du modèle ont été optimisés avec Adam sans *weight decay* pendant 20 époques.



(a) Représentation classique des informations temporelles



(b) Positionnement temporel indépendant des évènements

FIGURE 1 – Représentation des informations temporelles. La DCT est entourée, les expressions temporelles sont représentées en violet, les évènements sont représentés en gris et encadrés par leurs relations DocTimeRel et les TLINKs sont représentées par les flèches. La Figure 1a correspond à la représentation classique des DocTimeRel entre la DCT et les évènements, et des TLINKs entre les évènements et les expressions temporelles. La Figure 1b illustre notre représentation du positionnement des portions du texte par rapport à la DCT, indépendemment des évènements.

Extraction des évènements de toxicité. Pour une première pré-annotation et identification des évènements de toxicité des chimiothérapies, nous avons utilisé l'algorithme QuickUMLS (Soldaini, 2016) en utilisant un dictionnaire de toxicité (Rogier *et al.*, 2021), contenant des termes de toxicité en français issus de différentes terminologies.

3 Expérimentations

Données d'évaluation [1]**.** Pour l'extraction des relations temporelles, nous avons sélectionné au hasard des comptes rendus hospitaliers, opératoires et de consultation de patients atteints de cancer du côlon issus d'un entrepôt de données de santé de l'Hôpital Européen Georges Pompidou (HEGP) (Jannot *et al.*, 2017). Nous avons annoté 180 documents pour l'entraînement et la validation de notre modèle et 42 documents pour l'évaluation. Nous utilisons les catégories du schéma THYME-TimeML ainsi que deux autres catégories *TemporalReference* et *End_Scope*. La catégorie *TemporalReference* sert à identifier le début d'un compte rendu associé à une nouvelle date de création de document, ce qui est nécessaire pour les cas où différents comptes rendus sont inscrits dans le même document. *End_Scope* marque la fin d'une portion de texte si la prochaine portion est une en-tête ou une signature. Cela nous permet uniquement d'exclure ces zones dans le pré-traitement. Trois annotateurs ont annoté un échantillon de 9 documents. Ces accords inter-annotateurs en termes de macro F-mesure sont obtenus : 0,62, 0,73 et 0,69. Pour la classification, nous aurons donc ces cinq catégories : *TemporalReference*,

1. L'accès aux données cliniques a été autorisé par le conseil scientifique et éthique de l'AP-HP (CSE21-15_TALONCO).

Before, *Before_Overlap*, *Overlap* et *After*. La catégorie temporelle par défaut de *TemporalReference* est *Overlap*. Le guide d'annotation détaillé est fourni en Annexe A. Pour la validation de l'efficacité de notre approche, nous utilisons un deuxième corpus qui contient des informations de toxicité de chimiothérapies de patients atteints du cancer du côlon et du poumon (Jannot *et al.*, 2017). Une validation manuelle des annotations des évènements de toxicité a été réalisée par un expert sur 25 documents cliniques, dont 5 documents qui appartiennent au corpus de test de notre modèle de positionnement temporel. L'annotation a été effectuée avec l'outil BRAT (Stenetorp *et al.*, 2012).

Métriques d'évaluation. Dans notre étude, nous sommes intéressés par la détection du changement temporel entre des grandes portions du texte. Dans ce cas, la segmentation en phrases et en tokens n'a donc plus de sens et nous évaluons la performance de notre système de classification temporelle au niveau des caractères en mesurant la macro précision, le macro rappel et la macro F-mesure. Nous calculons les intervalles de confiance à 95 % de nos résultats de classification en utilisant la technique d'*empirical bootstrap* (Dekking *et al.*, 2005, p.275). Pour cela, nous avons échantillonné notre corpus de test avec remise 1000 fois. Les métriques seront calculées pour chaque sous-échantillon. Pour mesurer l'empreinte carbone de l'entraînement et l'évaluation de notre modèle, nous utilisons l'outil Carbontracker (Anthony *et al.*, 2020).

Modèle *Baseline*. Pour évaluer la performance de notre modèle d'extraction de relations temporelles, nous avons développé un modèle à base de règles pour le positionnement temporel des portions de texte. Comme pour l'annotation, nous nous sommes appuyés sur des mots-clés qui sont souvent utilisés pour définir certaines sections médicales, en particulier dans les comptes-rendus opératoires et hospitaliers telles que *"Antécédents", "Indication", "Gestes réalisés", "Traitement de sortie"*, etc. Ces mots-clés sont généralement utiles pour l'annotation temporelle, bien qu'ils soient insuffisants pour couvrir tous les types de comptes rendus. Le modèle *Baseline* sera évalué sur notre corpus de test.

4 Résultats et discussion

Extraction des relations temporelles cliniques. La Table 1 présente une comparaison de notre modèle de positionnement temporel avec le modèle *Baseline* à base de règles. Les meilleurs résultats sont obtenus avec notre modèle avec une macro F-mesure de 0,82, ce qui est supérieur aux accords inter-annotateurs. Les résultats sont beaucoup plus bas avec le modèle *Baseline* avec une F-mesure de 0,39. Malgré l'utilité des mots-clés introduisant les sections médicales pour la catégorisation temporelle, nous ne disposons pas d'une liste exhaustive. De plus, les sections varient énormément selon l'hôpital et même selon le service hospitalier et aucune catégorisation en sections n'est présente dans les comptes rendus du type courrier. Nous pouvons avoir des changements temporels dans une même section médicale et même dans une phrase (Figure 1b). Une approche liée uniquement à la macro-structure du texte ne permet donc pas de faire une analyse temporelle sur le document. De ce fait, ce type de changement ne sera donc pas identifié par le modèle *Baseline*, ce qui peut expliquer les mauvais résultats de ce modèle. L'émission de CO_2 résultant de l'entraînement et du test de notre modèle est estimée à 167 g, qui est l'équivalent de 1,55 km parcourus en voiture. La Table 2 présente les résultats par catégorie de notre modèle. La Table 3 représente le nombre et le pourcentage de portions de texte pour chaque catégorie pour les corpus d'entraînement et de test. Les catégories les plus prévalentes sont les mieux prédites. Ainsi, une F-mesure de 0,88 est obtenue pour la catégorie *Overlap*, représentant 35,5 % du corpus d'entraînement et de 0,86 pour la catégorie *Before*, représentant 22,8 % du corpus d'entraînement. Pour les catégories moins représentées telles que

	Précision	**Rappel**	**F-Mesure**	**Équivalent CO_2 (g.)**
Modèle *Baseline*	0,43 [0,37-0,52]	0,59 [0,51-0,66]	0,39 [0,31-0,46]	-
Notre modèle	**0,82 [0,78-0,85]**	**0,79 [0,75-0,85]**	**0,80 [0,76-0,85]**	167

TABLE 1 – Performance globale des modèles étudiés sur le corpus du test

TemporalReference et *After*, les résultats sont moins bons (F-mesure de 0,75 et 0,79 respectivement). Les portions de texte ayant la catégorie *Before_Overlap* sont souvent des phrases incluses dans des portions de la catégorie *Before* avec une indication temporelle qui indique la continuité dans le temps (Figure 2, Annexe A). Ce changement temporel est rarement prédit correctement et malgré la couverture de la la catégorie *Before_Overlap* (18,5 % dans le corpus d'entrainement) la performance est moins élevée (0,74 de F-mesure).

	P	**R**	**F**
TemporalReference	0,78	0,71	0,75
Before	0,91	0,82	0,86
Before_Overlap	0,75	0,73	0,74
Overlap	0,86	0,90	0,88
After	0,78	0,81	0,79
Macro-moyenne	**0,82**	**0,79**	**0,80**

TABLE 2 – Performance du positionnement temporel des portions du texte sur le corpus de test

	# portions de texte (test)	**# portions de texte (train)**
TemporalReference	46 (8,4%)	207 (11,2%)
Before	121 (22,2%)	423 (22,8%)
Before_Overlap	118 (21,6%)	343 (18,5%)
Overlap	191 (35%)	658 (35,5%)
After	70 (12,8%)	223 (12%)
Total	546	1854

TABLE 3 – Nombre de portions de texte pour chaque catégorie dans le corpus de test et d'entraînement

Positionnement temporel des évènements de toxicité des chimiothérapies. Afin d'évaluer l'efficacité de notre représentation des relations temporelles, nous utilisons notre modèle de positionnement temporel sur 25 documents cliniques contenant des évènements de toxicité. Ces évènements ont été identifiés en se basant sur un dictionnaire de toxicité et validés par un expert. La Table 4 présente les résultats du positionnement temporel des portions du texte, indépendamment des évènements ainsi que les résultats du positionnement temporel des évènements de toxicité sur 5 documents appartenant à la fois au corpus de test de la temporalité et de la toxicité. Nous avons une bonne performance de notre modèle pour le positionnement temporel des portions du texte avec une F-mesure de 0,84, indépendamment des évènements. Pour le positionnement temporel des évènements de toxicité, une F-mesure de 0,7 est obtenue. En moyenne, un positionnement de 10 évènements de toxicité par document est effectué sur cet échantillon de 5 documents. Nous observons une distribution similaire pour les 20 autres documents annotés en toxicité et nous estimons alors que les résultats seront similaires pour ces autres documents.

5 Conclusion

Nous proposons une nouvelle représentation des relations temporelles dans les textes, qui est indépendante des évènements et donc du domaine d'application. La tâche de l'extraction est formulée

	Précision	Rappel	F-Mesure
Positionnement des portions du texte	0,83 [0,76-0,88]	0,85 [0,74-0,97]	0,84 [0,74-0,92]
Positionnement des évènements de toxicité	0,87 [0,65-1]	0,70 [0,38-1]	0,70 [0,44-1]

TABLE 4 – Performance du positionnement temporel des portions du texte et des évènements de toxicité sur 5 documents du corpus de test

comme une tâche de classification des portions de texte homogènes du point de vue temporel en des catégories temporelles. Nous obtenons de bonnes performances de positionnement temporel des portions de texte cliniques écrits en français. Notre représentation temporelle nous a permis d'inférer le positionnement temporel des évènements de toxicité des chimiothérapies et semble être une bonne méthode pour le positionnement temporel de tout type d'évènement indépendamment du domaine.

Remerciements

Nous remercions le conseil scientifique et éthique de l'entrepôt de données de santé de l'AP-HP et l'Hôpital Européen Georges Pompidou qui nous ont permis d'avoir accès aux corpus utilisés dans ce travail. Nesrine Bannour a bénéficié d'un financement de l'ITMO Cancer Aviesan. Bastien Rance est soutenu par le programme SIRIC CARPEM.

Références

ALFATTNI G., PEEK N. & NENADIC G. (2020). Extraction of temporal relations from clinical free text : A systematic review of current approaches. *Journal of Biomedical Informatics*, **108**, 103488.

ALFATTNI G., PEEK N. & NENADIC G. (2021). Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries. *Journal of Biomedical Informatics*, **123**, 103915.

ANTHONY L. F. W., KANDING B. & SELVAN R. (2020). Carbontracker : Tracking and predicting the carbon footprint of training deep learning models. In *ICML Workshop on "Challenges in Deploying and monitoring Machine Learning Systems"*.

BETHARD S., DERCZYNSKI L., SAVOVA G., PUSTEJOVSKY J. & VERHAGEN M. (2015). SemEval-2015 task 6 : Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 806–814, Denver, Colorado : Association for Computational Linguistics. DOI : 10.18653/v1/S15-2136.

BETHARD S., SAVOVA G., CHEN W.-T., DERCZYNSKI L., PUSTEJOVSKY J. & VERHAGEN M. (2016). SemEval-2016 task 12 : Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 1052–1062, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1165.

BETHARD S., SAVOVA G., PALMER M. & PUSTEJOVSKY J. (2017). SemEval-2017 task 12 : Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, p. 565–572, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/S17-2093.

DEKKING F. M., KRAAIKAMP C., LOPUHAÄ H. P. & MEESTER L. E. (2005). *A Modern Introduction to Probability and Statistics : Understanding why and how*, volume 488. Springer.

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

DO Q., LU W. & ROTH D. (2012). Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 677–687, Jeju Island, Korea : Association for Computational Linguistics.

GUMIEL Y. B., SILVA E OLIVEIRA L. E., CLAVEAU V., GRABAR N., PARAISO E. C., MORO C. & CARVALHO D. R. (2021). Temporal relation extraction in clinical texts : A systematic review. *ACM Computing Surveys (CSUR)*, **54**(7), 1–36.

HAN R., NING Q. & PENG N. (2019). Joint event and temporal relation extraction with shared representations and structured prediction. In *Conference on Empirical Methods in Natural Language Processing*.

JANNOT A.-S., ZAPLETAL E., AVILLACH P., MAMZER M.-F., BURGUN A. & DEGOULET P. (2017). The georges pompidou university hospital clinical data warehouse : A 8-years follow-up experience. *International Journal of Medical Informatics*, **102**, 21–28. DOI : https ://doi.org/10.1016/j.ijmedinf.2017.02.006.

LIN C., DLIGACH D., MILLER T. A., BETHARD S. & SAVOVA G. K. (2016). Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association*, **23**(2), 387–395.

LIN C., MILLER T., DLIGACH D., SADEQUE F., BETHARD S. & SAVOVA G. (2020). A BERT-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, p. 70–75, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.bionlp-1.7.

LIU S., WANG L., CHAUDHARY V. & LIU H. (2019). Attention neural model for temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, p. 134–139.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE E. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. *ArXiv*, **abs/1911.03894**.

MOHARASAN G. & HO T.-B. (2019). Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research*, **3**(2), 220–244.

NAJAFABADIPOUR M., ZANIN M., GONZÁLEZ A. R., TORRENTE M., GARCÍA B. N., BERMUDEZ J. L. C., PROVENCIO M. & RUIZ E. M. (2020). Reconstructing the patient's natural history from electronic health records. *Artificial intelligence in medicine*, **105**, 101860.

NING Q., FENG Z. & ROTH D. (2017). A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 1027–1037, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1108.

PUSTEJOVSKY J., CASTANO J. M., INGRIA R., SAURI R., GAIZAUSKAS R. J., SETZER A., KATZ G. & RADEV D. R. (2003). Timeml : Robust specification of event and temporal expressions in text. *New directions in question answering*, **3**, 28–34.

PUSTEJOVSKY J. & STUBBS A. (2011). Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, p. 152–160, Portland, Oregon, USA : Association for Computational Linguistics.

ROGIER A., COULET A. & RANCE B. (2021). Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from EHRs. In *Medinfo 2021 - 18th World Congress on Medical and Health Informatics*, Virtual conference, Australia. HAL : hal-03364585.

SOLDAINI L. (2016). QuickUMLS : a fast, unsupervised approach for medical concept extraction.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France : Association for Computational Linguistics.

STYLER IV W. F., BETHARD S., FINAN S., PALMER M., PRADHAN S., DE GROEN P. C., ERICKSON B., MILLER T., LIN C., SAVOVA G. & PUSTEJOVSKY J. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, **2**, 143–154. DOI : 10.1162/tacl_a_00172.

SUN W., RUMSHISKY A. & UZUNER Ö. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *Journal of the American Medical Informatics Association : JAMIA*, **20 5**, 806–13.

TOURILLE J., FERRET O., NÉVÉOL A. & TANNIER X. (2016). Limsi-cot at semeval-2016 task 12 : Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 1136–1142.

TOURILLE J., FERRET O., TANNIER X. & NÉVÉOL A. (2017). Temporal information extraction from clinical text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 739–745, Valencia, Spain : Association for Computational Linguistics.

VIANI N., MILLER T. A., NAPOLITANO C., PRIORI S. G., SAVOVA G. K., BELLAZZI R. & SACCHI L. (2019). Supervised methods to extract clinical events from cardiology reports in italian. *Journal of biomedical informatics*, **95**, 103219.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 38–45, Online : Association for Computational Linguistics.

A Temporal Annotation scheme for our clinical corpus

A.1 Definitions of temporal categories

To annotate the temporal information in a clinical report, we define a temporal annotation scheme based on the Document Creation Time (DCT) and the possible categories of the Document creation Time Relation (DocTimeRel). The DCT might be the current medical visit date, usually stated in the document heading. It might also be the length of time spent in the hospital. The DCT does not need to be annotated.

A.1.1 Document creation Time Relation

Document creation Time Relation is the relation between events and Document Creation Time. We consider these four possible categories for this time relation : Before, Before_Overlap, Overlap, and After. We annotate only the first word of each temporal portion. We consider that the start of a temporal portion denotes the end of the previous one.

A.1.2 Before

The Before category is used to annotate narrative portions referring to what occurred before the Document Creation Time.

Examples

- Antécédents, antécédents médicaux, antécédents chirurgicaux, Antécédents familiaux, Histoire de la maladie, Rappel clinique, Rappel sur la pathologie → All terms referring to the medical history section.
- **Except :** Maladie traitée depuis le → Before_Overlap since we have a temporal indication that the procedure/disease is still ongoing for the patient (cf. Figure 2).

A.1.3 Before_Overlap

The Before_Overlap category is used to annotate narrative portions that started before the document creation time and are still ongoing at that time.

Examples

- Comorbidités, Mode de vie, Autonomie, traitement habituel, traitement à l'entrée, Allergies, Traitements concomitants, Facteurs de risque, Indication, Indication opératoire, décision d'une intervention
- Patient de 70 ans
- HTA traitée depuis, dans le cadre d'un suivi d'un cancer → The patient is still suffering from the disease.
- METASTASES HEPATIQUES D'UN ADENOCARCINOME → The disease's name as a title in operative reports, which is generally capitalized (cf. Figure 3).

A.1.4 Overlap

The Overlap category is used to annotate narrative portions that happen at the same time as the document creation time.

Examples

- Examen pratique, Tolérance intercure, Au total, Conclusion, Gestes opératoires, Gestes réalisés, Motif d'hospitalisation, Biologie, Biologie de sortie, INTERVENTION, constantes à l'arrivée, Date d'hospitalisation, Date d'entrée, Date de l'intervention, Motif

— Examens complémentaires, Examens paracliniques → Sometimes, some complementary exams are conducted before the document creation time but because they are done for the purpose of the hospital stay, we annotate them as Overlap (cf. Figure 2).
— Je vois ce jour, Je revois en consultation

A.1.5 After

The After category is used to annotate narrative portions referring to what occurs after the document creation time.

Examples
— Traitement de sortie, Prochains rendez-vous, Rendez-vous à venir, Prescription de médicaments, Date de la prochaine cure, Ordonnance de sortie, Prochains examens
— Je reverrai ce patient, je prévois une coloscopie
— La pièce est envoyée pour un examen histologique

A.2 Other categories

A.2.1 TemporalReference

Because several medical reports might be written in the same document, the TemporalReference category specifies the beginning of a new clinical report. Because several medical reports might be written in the same document, the TemporalReference category specifies the beginning of a new clinical report. Each clinical report will then have its own Document Creation Time, and the annotations will be based on this DCT. The TemporalReference category's default Document Time Relation is assumed to be Overlap and does not need to be annotated.

Examples
— Compte-rendu opératoire, Compte-rendu d'hospitalisation, Paris, le 14 octobre 2018

A.2.2 End_Scope

We do not consider heading and signature information in our annotation. Therefore, we use the category End_Scope to mark the ending of a narrative portion if the next narrative portion is a heading or a signature. This way, we avoid annotating the contact information for the health care unit, which may be repeated in several clinical reports. Despite the fact that the clinical documents are de-identified, we avoid annotating specific patient information. In cases other than headings or signatures, the end of a temporal portion is implicitly considered the start of a new temporal portion.

A.3 Examples of annotations made in accordance with the above scheme and guidelines

Annotations of the first example (cf. Figure 2)

— From *Compte* to *d'hospitalisation* as TemporalReference
— From *Hospitalisé* to 30/07/2013 as Overlap
— From *Motif* to *d'HOSPITALISATION :* as Overlap, note that we don't annotate the temporal portion after the End_Scope category containing contact information of doctors
— From *HISTOIRE* to *ANTECEDENTS* as Before
— From *HTA* to *2012,* as Before_Overlap since we have a temporal indication that the disease is still ongoing for the patient
— From *phlébite* to *07/2012* as Before since it's part of the medical patient history
— From *ALLERGIES* to *Autonome* as Before_Overlap
— From *Examens* to *et* as Overlap despite the fact that the medical exams are conducted before the date of hospital admission
— From *sera* to *10/09/2013* as After. The signature of the document after the End_Scope category is not annotated

FIGURE 2 – A first example of hospital report annotations

Annotations of the second example (cf. Figure 3)

— From *COMPTE* to *OPERATOIRE* as Temporal Reference
— *ADENOCARCINOME* as Before_Overlap
— *COLECTOMIE* as Overlap
— From *Rappel* to *clinique :* as Before
— From *Indication* to *opératoire.* as Before_Overlap
— From *Gestes* to *réalisés :* as Overlap
— From *La* to *histologique.* as After

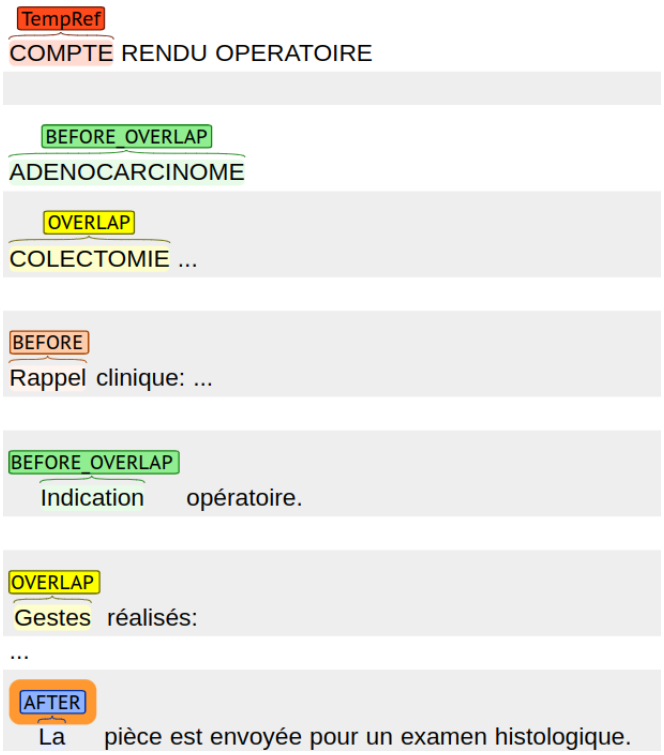


FIGURE 3 – A second example of annotating an operative report

Annotations of the third example (cf. Figure 4)

— From *Paris* to *2014,* as TemporalReference
— From *Je* to *jour* as Overlap
— From *Monsieur* to *comme* as Before_Overlap for the patient's age and since it is stated that the purpose of the medical visit is a disease follow-up
— From *antécédent* to *Rappel :* as Before
— From *Examen* to *pratique :* as Overlap
— From *A* to *mois* as After
— From *Dossier* to *staff* as TemporalReference, it's a new clinical report
— From *Dernières* to *2014 :* as Before, based on the document creation time of the second clinical report.
— From *Décisions* to *staff :* as Overlap
— From *Le* to *consultation.* as After

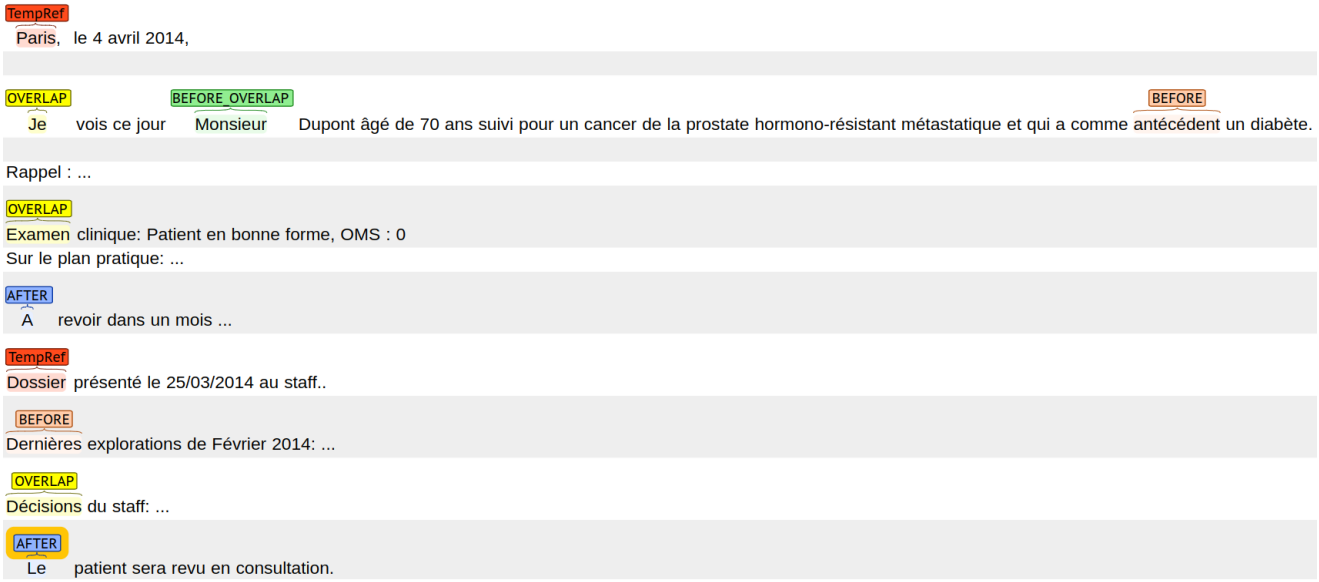


FIGURE 4 – A third example of annotating a clinical document containing two clinical reports

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

Une grammaire formelle pour les langues des signes basée sur AZee : une proposition établie sur une étude de corpus

Camille Challant, Michael Filhol
Université Paris-Saclay, CNRS, LISN, 91400 Orsay, France
{camille.challant, michael.filhol}@lisn.upsaclay.fr

RÉSUMÉ

Cet article propose de premières réflexions quant à l'élaboration d'une grammaire formelle pour les langues des signes, basée sur l'approche AZee. Nous avons mené une étude statistique sur un corpus d'expressions AZee, qui décrivent des discours en langue des signes française. Cela nous permet d'entrevoir des contraintes sur ces expressions, qui reflètent plus généralement les contraintes de la langue des signes française. Nous présentons quelques contraintes et positionnons théoriquement notre ébauche de grammaire au sein des différentes grammaires formelles existantes.

ABSTRACT

A formal grammar for sign languages based on AZee : a proposal established on a corpus study

This article provides some initial thoughts about the development of a formal grammar for sign languages, based on the AZee approach. We have conducted a statistical study on a corpus of AZee expressions, which describe French sign language discourses. This allows us to glimpse some constraints on these expressions, which reflect more generally the constraints of French sign language. We present some constraints and theoretically position this draft grammar among the various existing formal grammars.

MOTS-CLÉS : AZee, langue des signes française, modélisation, grammaire formelle.

KEYWORDS: AZee, French sign language, modeling, formal grammar.

1 Introduction

Les langues des signes (LS) sont des langues visuo-gestuelles, qui se distinguent des langues audio-vocales sur de nombreux points. Les mains et les doigts mais aussi le buste, le regard, les sourcils sont autant d'articulateurs qui entrent en jeu dans les LS, et s'animent dans l'espace de signation pour produire des énoncés. Ces différents articulateurs rendent possible la multilinéarité, c'est-à-dire le fait de pouvoir réaliser plusieurs choses simultanément : les LS ne sont donc pas nécessairement des séquences de signes placés les uns à la suite des autres. Ces caractéristiques (spatialisation, multilinéarité) deviennent de véritables défis lorsque l'on s'intéresse à la modélisation des LS, qui est nécessaire en traitement automatique des langues pour des tâches de synthèse, de reconnaissance ou encore de traduction automatique. Le modèle formel sur lequel nous travaillons, nommé AZee, permet de représenter des discours en LS en tenant compte des spécificités qui viennent d'être évoquées.

Nous proposons, dans cet article, de premières réflexions quant à l'élaboration d'une grammaire

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 15–22.

formelle pour la langue des signes française (LSF), établie à partir de l'étude d'un corpus décrit à l'aide d'AZee. Nous entendons par grammaire formelle un système de règles permettant de décrire la langue et de juger si un énoncé répond aux contraintes de celle-ci, système ne laissant aucune place à l'interprétation humaine et pouvant être utilisé par des programmes informatiques.

Nous commençons par présenter le modèle AZee et notre question de recherche, avant d'exposer notre méthode ainsi que nos premiers résultats. Nous comparons dans une dernière section notre potentielle grammaire AZee avec les différentes grammaires formelles existantes.

2 AZee

AZee (Filhol *et al.*, 2014) est une approche de description des LS fondée sur la notion essentielle de *règle de production*, qui associe à un sens identifié un ensemble de formes observables à produire. Cela permet de ne faire aucun présupposé concernant l'existence de niveaux linguistiques, de catégories grammaticales ou d'un ordre séquentiel.

Par exemple, en LSF, la règle de production `président` associe le sens 'président/présidence' à la forme illustrée en figure 1a. De même, la règle `info-about`, à deux arguments (*topic* et *info*), associe le sens '*info*, à propos de *topic*' à la synchronisation de formes présentée en figure 1b : les deux arguments sont placés en séquence et séparés par une durée contrôlée, chacun est maintenu (*hold*) sur une durée plus ou moins longue et un clignement des yeux (*el :cl*) se synchronise avec la fin du maintien d'*info*. Les différentes règles de production peuvent se combiner entre elles récursivement pour construire des *expressions AZee de discours*, qui reflètent le sens que l'on interprète à partir des formes qu'elles produisent. Un exemple d'une telle expression est donné en figure 1c : elle représente une production en LSF dont les formes correspondent à celles de la figure 1b, avec les règles `président` et `célèbre` respectivement en *topic* et *info*, et signifie « [le/un] président est célèbre ».

(a) forme pour "président/présidence" en LSF

(b) synchronisation de formes pour la règle `info-about`

```
:info-about
    'topic
    :président
    'info
    :célèbre
```

(c) Expression AZee de discours signifiant "le/un président est célèbre"

FIGURE 1 – Présentation d'AZee

Il est possible, de cette manière, de décrire de vraies productions en LSF avec AZee, ce que nous avons fait avec les 120 discours composant le corpus des "40 brèves" (40 entrées journalistiques en français écrit, chacune traduite en LSF par 3 traducteurs sourds) (Filhol & Tannier, 2014). Un corpus de 120 expressions AZee de discours représentant une heure de LSF au total est ainsi disponible, et

comporte 11 470 applications de règles de production (Challant & Filhol, 2022). Nous avons décidé de travailler sur ce corpus.

3 Question de recherche et méthode

Si une expression AZee modélise en effet une production langagière en LSF, alors toutes les contraintes qui pèsent sur la langue devraient être reflétées dans les expressions AZee de discours. Nous faisons donc l'hypothèse qu'il existe des contraintes qui régissent ces expressions. Si cette hypothèse est vérifiée, nous pourrons considérer ces contraintes comme des contraintes grammaticales de la LSF. Nous reviendrons sur cette considération plus en détails dans la section 5.

Plusieurs questions se posent alors pour identifier des contraintes sur les expressions AZee de discours : les différentes règles de production apparaissent-elles dans les mêmes contextes ? Leurs arguments sont-ils contraints dans leur complexité et si oui, comment ? Certaines règles de production sont-elles plus fréquentes que d'autres ? Afin d'être en mesure de répondre à ces questions, nous avons imaginé plusieurs tests, applicables par exemple, pour toute expression ou sous-expression E, à :

- $rootname(E)$: règle de production appliquée à la racine de E, identifiée par son nom ;
- $prodcount(E)$: nombre d'applications de règles de production dans E ;
- $contains\text{-}rule(E, R)$: vrai si et seulement si il existe une sous-expression E' dans E telle que $rootname(E') = R$ avec R le nom d'une règle de production ;
- $E.arg$: sous-expression utilisée comme argument arg de E (E est l'application d'une règle définissant arg comme nom d'un de ses arguments).

Ces tests sont combinables à l'aide d'opérateurs booléens afin de créer des requêtes plus complexes sur les expressions.

Nous avons ensuite appliqué ces tests sur le corpus d'expressions AZee présenté dans la section précédente. Nous présentons les premiers résultats que nous avons obtenus suite à cette étude dans la section suivante.

4 Premiers résultats

Dans un premier temps, nous pouvons constater que les règles de production à la racine des 120 expressions ne sont que des `context` (95 occurrences) et des `info-about` (25 occurrences). Il semblerait donc que la racine d'une expression de discours de genre journalistique soit plutôt contrainte. Nous pouvons formuler une hypothèse de contrainte sur toute expression de discours E :

$$rootname(E) \in \{\texttt{context}, \texttt{info-about}\}$$

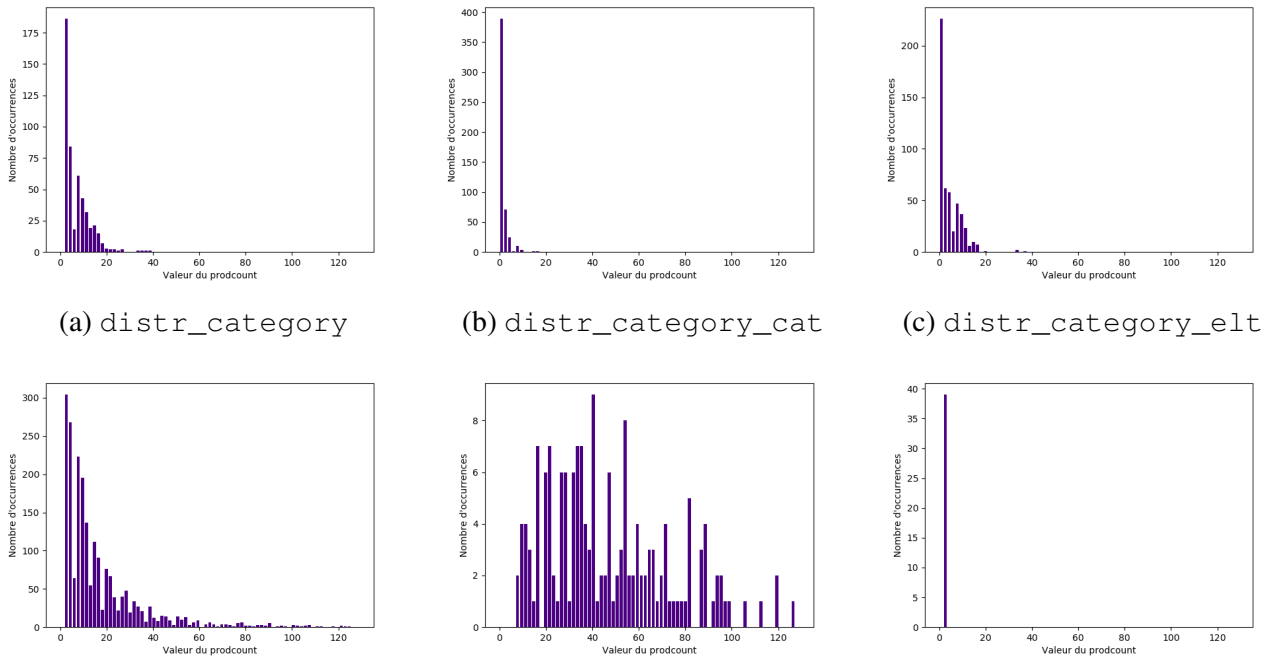
Dans un deuxième temps, nous nous sommes intéressés au poids des constituants, en mesurant leur $prodcount$, qui compte le nombre d'applications de règles de production contenues dans les expressions, ce qui reflète leur complexité.

Nous avons, pour chaque règle de production R connue, établi :

- la distribution $distr\_R$ des $prodcount(E)$ sur l'ensemble des expressions ou sous-expressions E du corpus telles que $rootname(E) = R$

— pour chaque argument arg défini par R, la distribution $distr\_R\_arg$ des $prodcount(E.arg)$ sur l'ensemble des expressions E telles que $rootname(E) = R$

Nous avons obtenu des distributions, dont nous montrons quelques exemples en figure 2. Celles-ci sont très différentes les unes des autres : certaines règles semblent présenter une limite maximale (comme ici `category` ou `info-about`), d'autres une limite minimale (`prise-de-parole`) et d'autres encore ont une valeur de $prodcount$ qui est fixe (`tens-unit`). Nous avons décidé de nous concentrer sur les règles de production que nous venons de citer, car elles présentent des contrastes intéressants. Nous les présentons succinctement dans le tableau 1.



(a) `distr_category` (b) `distr_category_cat` (c) `distr_category_elt`

(d) `distr_info-about` (e) `distr_prise-de-parole` (f) `distr_tens-units`

FIGURE 2 – Distribution du $prodcount$ de différentes règles de production

Nom de la règle	`category`	`prise-de-parole`	`tens-units`
Arguments	*cat, elt*	*sig*	*tens, units*
Nombre d'occurrences dans le corpus	500	165	44
Sens	*elt*, interprété comme une instance de *cat*	pause rhétorique avant *sig*	nombre formé par deux chiffres, *tens* et *units*

TABLE 1 – Présentation de trois règles de production

On remarque, sur la figure 2a, qu'il n'y a pas de prodcount supérieur à 20, ce qui montre une limite maximale pour les expressions dont `category` est à la racine. On remarque la même limite sur la figure 2c, c'est-à-dire pour les expressions utilisées comme argument *elt* de `category`. En revanche, cette limite est de 10 sur la figure 2b ce qui signifie que l'argument *cat* semble davantage limité en complexité que *elt*. Un contraste similaire est observé entre les arguments *topic* et *info* de la règle `info-about`, la limite se dégageant pour *info* (80) étant le double de celle pour *topic* (40).

Pour `prise-de-parole`, au contraire, aucune limite maximale ne se dégage (figure 2e), mais toutes les valeurs sont supérieures à 7. Ainsi, la limite semble être minimale : on observe un comportement inverse au précédent.

Enfin, un troisième comportement se dégage avec `tens-units`, comme l'illustre la figure 2f : sur les 44 occurrences du corpus, toutes les expressions ayant pour racine cette règle ont un $prodcount$ de 3. Plus précisément même, par argument, si $rootname(E) = \texttt{tens-units}$ alors :

$$prodcount(E.tens) = 1, prodcount(E.units) = 1$$

Dans un troisième temps, nous nous sommes penchés sur les règles pouvant être contenues dans la descendance d'une règle de production. Nous avons considéré l'ensemble $\mathcal{R}$ des règles ayant au moins un argument obligatoire et présentant plus de 20 occurrences dans le corpus, soit 19 règles de production.

Nous avons, pour chaque couple R_1, R_2 de règles de $\mathcal{R}$, compté le nombre de fois où R_2 apparaît dans R_1, ainsi que dans chacun de ses arguments pris séparément, c'est-à-dire :

— le nombre d'expressions E telles que $rootname(E) = R_1$ ET $contains\text{-}rule(E, R_2)$
— pour chaque argument arg défini par R_1, le nombre d'expressions E telles que $rootname(E) = R_1$ ET $contains\text{-}rule(E.arg, R_2)$

Dans le cas où $R_1 = \texttt{category}$, nous avons remarqué que 9 règles de production n'étaient jamais contenues dans l'argument *cat* alors qu'elles ne sont que 3 à n'être jamais contenues dans *elt*, et deux qui ne sont ni contenues dans *elt*, ni dans *cat* : `prise-de-parole` et `context`. En revanche, il n'y a pas de telle contrainte concernant `prise-de-parole` et `info-about` : toutes les règles de production de $\mathcal{R}$ apparaissent dans les expressions dont elles sont les racines. Pour finir, aucune règle de production de $\mathcal{R}$ n'est contenue dans les expressions ayant pour racine `tens-units`. Toutes les règles de production ne sont donc pas contraintes de la même façon sur leur descendance.

Enfin, dans un dernier temps, les règles `category`, `tens-units`, `info-about` peuvent être observées à la racine de n'importe quel argument d'une règle de production de l'ensemble $\mathcal{R}$. Au contraire, `prise-de-parole` ne se trouve à la racine que d'arguments de trois règles : `context`, `info-about` et `each-of`.

Pour conclure, à partir des observations réalisées sur notre corpus, nous avons pu identifier des contraintes sur les expressions AZee de discours. On remarque, par exemple, des règles de production qui peuvent accepter des arguments très complexes tandis que cela n'est jamais observé chez d'autres, certaines règles de production peuvent contenir n'importe quelle autre règle de production là où d'autres sont contraintes à ce niveau.

5 Discussion et positionnement théorique

Nous venons, dans la section précédente, d'identifier plusieurs contraintes auxquelles les différentes règles de production peuvent être soumises. Selon nous, et comme mentionné plus haut, une contrainte sur une expression AZee peut être perçue comme une contrainte grammaticale, qui gouverne la combinaison, la taille, la position ou encore la fréquence d'apparition des différentes unités de la langue. Ces contraintes formelles composent ensemble un système, qui peut être considéré comme une grammaire.

Ainsi construite, notre grammaire comporterait plusieurs caractéristiques qui la distingueraient des autres grammaires, à commencer par son absence de présupposés, sur plusieurs plans. Tout d'abord, la séquence n'est pas admise d'office : l'ordre linéaire dans lequel certains éléments apparaissent dans la forme produite par une expression est simplement le résultat de l'application de règles de production combinées. Les niveaux linguistiques ne sont pas non plus présupposés, pas plus que les catégories syntaxiques. Nous revenons ainsi à quelque chose de plus fondamental, sans pour autant nier l'existence de ces notions en LSF : il s'agit simplement là de remettre en question leur caractère fondamental. Si la séquence, les niveaux linguistiques ou les catégories syntaxiques sont des notions pertinentes pour décrire les LS, elles peuvent être définies à partir de critères formels plus fondamentaux, et sont donc en réalité émergentes. Par exemple, en nous appuyant sur les résultats présentés dans la section 4, nous pourrions définir des classes de règles qui répondent aux mêmes contraintes sur leur nombre d'arguments ou encore sur le poids de leurs arguments, Ensuite, notre grammaire a la caractéristique de porter de la sémantique à tous les niveaux de l'expression : chaque noeud est porteur de sens, quelle que soit sa position dans l'expression. Enfin, chaque expression AZee détermine des formes à produire, en prenant en compte les articulateurs manuels comme non manuels : aucune hiérarchie n'est présumée entre ces différents articulateurs.

Ces caractéristiques opposent notre approche aux grammaires génératives (Chomsky, 1965), centrées sur la syntaxe et basées sur un ordre séquentiel, qui ont sont encore très utilisées pour décrire les LS (Aristodemo & Hauser, 2021; Kimmelman & Pfau, 2021; Napoli & Sutton-Spence, 2014). Les grammaires génératives s'intéressent notamment à l'ordre des unités lexicales dans le discours, chacune étant pré-étiquetée avec une catégorie syntaxique. Les règles s'appliquant à ces catégories et aux nœuds dans les arbres syntaxiques (VP, NP, etc.) opèrent sans recours au niveau sémantique.

En revanche, plusieurs idées appartenant à d'autres approches formalistes retiennent notre attention, bien que la notion de catégorie syntaxique reste au cœur de toutes celles-ci. Par exemple, les grammaires cognitives (Martínez *et al.*, 2020; Langacker, 1987) accordent une grande place à la sémantique, ce qui est essentiel dans l'approche AZee. De plus, nous pouvons apercevoir quelques points communs entre notre grammaire et les grammaires de construction (Beuls & Van Eecke, 2023; van Trijp, 2015; Fillmore, 1988), qui prennent également pour base l'association forme-sens présente dans les langues. Ces grammaires mettent en avant un continuum lexique-syntaxe et des niveaux linguistiques qui ne sont pas clairement distingués, ce qui fait écho à notre approche. Enfin, les grammaires de propriétés (Blache, 2001) nous intéressent particulièrement : les propriétés sont des contraintes, et ce système de contraintes permet de caractériser les énoncés grâce à un gradient de grammaticalité (plutôt qu'un jugement binaire) dont la valeur peut être déterminée par la nombre de contraintes satisfaites, ce qui nous semble tout à fait approprié pour les langues orales que sont les LS.

6 Conclusion et perspectives

Pour conclure, nous avons présenté dans cet article les premières bases d'une grammaire formelle fondée sur AZee. Nous avons également positionné cette ébauche de grammaire au sein des grammaires formelles existantes. Une de nos perspectives à court terme est d'expérimenter l'extraction automatique de motifs réguliers de notre corpus, à l'instar de Herrera *et al.* (2022). Nous aimerions pouvoir adapter leurs méthodes et outils à nos données. De plus, nous souhaiterions augmenter notre corpus d'étude, tout en diversifiant le genre de discours décrits. Nous envisageons de décrire avec AZee le corpus Mocap1 (Benchiheub *et al.*, 2016), qui comporte un grand nombre de structures

dites iconiques, structures propres aux LS et reconnues comme très fréquentes. Nous pourrons ainsi l'explorer avec notre méthodologie et comparer les résultats obtenus avec ceux de notre corpus actuel. Enfin, une de nos perspectives est de générer avec un avatar (puisque cela est possible avec AZee !) des discours qui répondent – ou non – à nos contraintes, et de les présenter à des locuteurs natifs de la LSF afin de confirmer nos hypothèses.

Références

ARISTODEMO V. & HAUSER C. (2021). Similar but Different : Investigating Temporal Constructions in Sign Language. *Glossa : a Journal of General Linguistics 6(1) : 2*, **6**. DOI : 10.5334/gjgl.999.

BENCHIHEUB M.-E.-F., BERRET B. & BRAFFORT A. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language. In *7th International Conference on Language Resources and Evaluation - Workshop on the Representation and Processing of Sign Languages (LREC-WRPSL 2016)*, p. 7–12, May 23-28, Portoroz, Slovenia. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr, v1, https ://hdl.handle.net/11403/mocap1/v1.

BEULS K. & VAN EECKE P. (2023). Fluid Construction Grammar : State of the Art and Future Outlook. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, p. 41–50, Washington, D.C. : Association for Computational Linguistics.

BLACHE P. (2001). *Les grammaires de propriétés : des contraintes pour le traitement automatique des langues naturelles*. Collection Technologies et cultures. Hermès Science publications.

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.

CHOMSKY N. (1965). *Aspects of the Theory of Syntax*. Cambridge : MIT Press.

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

FILHOL M. & TANNIER X. (2014). Construction of a French–LSF Corpus. In *Building and Using Comparable Corpora, Language Resource and Evaluation Conference (LREC)*, p. 2–5, Reykjavik, Iceland.

FILLMORE C. J. (1988). The Mechanisms of "Construction Grammar". *Annual Meeting of the Berkeley Linguistics Society*, **14**(00), 35–55. DOI : 10.3765/bls.v14i0.1794.

HERRERA S., KAHANE S. & GUILLAUME B. (2022). Extraction de règles de grammaire à partir de treebanks : développement d'un outil et premiers résultats. *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, (93–98).

KIMMELMAN V. & PFAU R. (2021). Information Structure – Theoretical Perspectives. *The Routledge handbook of theoretical and experimental sign language research.*, p. 591–613.

LANGACKER R. W. (1987). *Foundations of Cognitive Grammar : Volume I : Theoretical Prerequisites*. Stanford University Press.

MARTÍNEZ R., SIYAVOSHI S. & WILCOX S. (2020). Advances in the Study of Signed Languages within a Cognitive Perspective. *Hesperia : Anuario de Filología Hispánica*, **23**, 29–56. DOI : 10.35869/hafh.v23i0.1654.

NAPOLI D. J. & SUTTON-SPENCE R. (2014). Order of the Major Constituents in Sign Languages : Implications for All Language. *Frontiers in Psychology*, **5**, 376. DOI : 10.3389/fpsyg.2014.00376.

VAN TRIJP R. (2015). Towards Bidirectional Processing Models of Sign Language : A Constructional Approach in Fluid Construction Grammar. In *Proceedings of the EuroAsianPacific joint conference on cognitive science*, p. 668–673, Turin : Univeristy of Torino.

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

Des ressources lexicales du français et de leur utilisation en TAL : étude des actes de TALN

Hee-Soo Choi[1, 2] Karën Fort[2, 3] Bruno Guillaume[2] Mathieu Constant[1]
(1) ATILF, CNRS, Université de Lorraine, 54000 Nancy, France
(2) LORIA, Université de Lorraine, 54506 Vandoeuvre-lès-Nancy, France
(3) Sorbonne Université, 75006 Paris, France
hee-soo.choi@loria.fr, karen.fort@loria.fr,
bruno.guillaume@loria.fr, mathieu.constant@atilf.fr

RÉSUMÉ

Au début du XXIe siècle, le français faisait encore partie des langues peu dotées. Grâce aux efforts de la communauté française du traitement automatique des langues (TAL), de nombreuses ressources librement disponibles ont été produites, dont des lexiques du français. À travers cet article, nous nous intéressons à leur devenir dans la communauté par le prisme des actes de la conférence TALN sur une période de 20 ans.

ABSTRACT

French lexicons and their usage in NLP : a study of TALN proceedings.

At the beginning of the 21st century, French was still considered a less-resourced language. Thanks to the efforts of the French automatic language processing (ALP) community, many freely available resources have been produced, including French lexicons. We are interested here in their future in the community through the prism of the proceedings of the TALN conference over a 20-year period.

MOTS-CLÉS : lexiques, ressources lexicales, français.

KEYWORDS: lexica, lexical resources, French.

1 Introduction

Depuis le début des années 2000, de nombreux efforts ont été réalisés pour créer des ressources langagières pour le français, notamment lexicales, qui soient librement disponibles. En 2019, Mariani *et al.* (2019b) démontrent l'importance de ces ressources en recueillant le nombre de mentions d'une large variété de ressources langagières sur une période de 50 ans, distinguant les lexiques, des corpus et des outils de Traitement Automatique des Langues (TAL). La ressource langagière la plus mentionnée est `WordNet` (Fellbaum, 1998) suivi de trois corpus `Timit`, `Wikipedia` et le `Penn Treebank` (Marcus *et al.*, 1993). Mais qu'en est-il pour les ressources lexicales du français ? Les efforts de la communauté ont-ils portés ? Inspirés des travaux de Mariani *et al.* (2019a,b), notre étude présente la dynamique des ressources lexicales du français librement disponibles dans les actes de la conférence TALN entre 2001 et 2022. À travers ce travail, nous nous intéressons à la manière dont ces ressources sont utilisées, par qui et sur quelles applications.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 23–36.

2 Les ressources lexicales publiées à TALN

2.1 Des lexiques aux plongements lexicaux statiques

Il existe aujourd'hui de nombreuses ressources lexicales du français utilisables en TAL, de différents types (lexiques, réseaux lexico-sémantiques, corpus...) et de différents niveaux de description linguistique (syntaxe, sémantique, morphologie...) (Gala, 2013). Si les premières ressources ont été créées manuellement par des linguistes, la communauté de TAL s'est efforcée de concevoir des ressources lexicales informatisées adaptées à ses besoins, notamment celui d'obtenir une ressource conséquente et la plus complète possible. Ainsi, une large variété de ressources a vu le jour grâce à des constructions semi-automatique, par myriadisation ou par fusion de ressources existantes (Choi, 2022).

Recenser une liste exhaustive de ressources constitue alors une tâche délicate dans la mesure où un certain nombre de ressources en englobent d'autres. De ce fait, afin de fixer un critère de sélection, nous avons considéré tous les lexiques du français présents dans la LRE map [1] et Ortolang [2] dont les liens étaient accessibles pour télécharger la ressource. Si nous faisons le choix de considérer les ressources multilingues, nous décidons de laisser de côté Wikipédia et Wiktionnaire qui n'ont pas été créés initialement pour la recherche ainsi que les lexiques spécialisés et de langues régionales. Nous obtenons ainsi une liste finale de 31 ressources présentées dans le tableau 1. Les personnes impliquées dans la conception des ressources correspondent aux auteurs des articles de référence et aux personnes présentées comme contributeurs dans les sites officiels des ressources.

Ressources	Création	Personnes impliquées dans la conception	Licence
Lexique-Grammaire (Gross, 1975)	Fin 1960	M. Gross et al. (LADL) [3]	LGPL-LR
DELA	1990	M. Gross et al. (LADL)	LGPL-LR
LVF (Dubois & Dubois-Charlier, 1997)	1997	J. Dubois, F. Dubois-Charlier	LGPL-LR
Dictionnaire Électronique des Synonymes (DES) (Ploux & Victorri, 1998)	1998	S. Ploux, B. Victorri, J-L. Manguin, M. Morel, L. Chardon	CC BY-NC-SA
Dicovalence (Van den Eynde & Mertens, 2006, 2010; Mertens, 2010)	2003	K. van den Eynde, P. Mertens	LGPL-LR
Lefff (Clément *et al.*, 2004; Sagot, 2010)	2003	B. Sagot, L. Clément	LGPL-LR
ProLexBase (Tran & Maurel, 2006)	2006	M. Tran, D. Maurel	LGPL-LR
Jibiki (Mangeot & Chalvin, 2006; Mangeot-Nagata, 2016)	2006	M. Mangeot-Nagata, A. Chalvin	Domaine Public
JeuxDeMots (Lafourcade & Joubert, 2008; Lafourcade & Le Brun, 2020)	2007	M. Lafourcade, A. Joubert, N. Le Brun	Domaine Public
VfrLPL (Rauzy & Blache, 2007)	2007	S. Rauzy, P. Blache	CRDO
LGLex (Constant & Tolone, 2010)	2008	E. Tolone, M. Constant	LGPL-LR
WOLF (Sagot & Fišer, 2008)	2008	B. Sagot, D. Fišer	CeCILL-C
RL-Fr (Lux-Pogodalla & Polguère, 2011)	2011	V. Lux-Pogodalla, A. Polguère, S. Ollinger	CC BY
DBnary (Sérasset, 2015)	2012	G. Sérasset	CC BY-SA
Dictionnaire morphosyntaxique du français (DM) (Trouilleux, 2012)	2012	F. Trouilleux	GNU GPL
DiLAF (Enguehard *et al.*, 2012)	2012	C. Enguehard, S. Kané, M. Mangeot, I. Modi, M. Sanogo	CC BY-SA
GLÀFF (Sajous *et al.*, 2013)	2013	F. Sajous, N. Hathout, B. Calderone	CC BY-SA
Marsalex (Blache & Rauzy, 2008)	2013	P. Blache, S. Rauzy	CC BY
Démonette (Hathout & Namer, 2014)	2014	N. Hathout, F. Namer	CC BY-NC
FLELex (François *et al.*, 2014)	2014	T. François, N. Gala, P. Watrin, C. Fairon, A. Pintard	CC BY-NC-SA
French FrameNet (Candito *et al.*, 2014)	2014	M. Candito, P. Amsili, L. Barque, F. Benamara, G. Chalendar, L. Vieu, M. Djemaa, P. Haas, R. Huyghe, Y. Mathieu, P. Muller, B. Sagot	LGPL-LR
VerbeNet (Danlos *et al.*, 2014)	2014	L. Danlos, T. Nakamura Q. Pradet	CC BY-SA
OpeNER-sentiment-lexicons (Maks *et al.*, 2014)	2014	I. Maks, R. Izquierdo, F. Frontini, R. Agerri, P. Vossen, A. Azpeitia	OpenSource
Morphalou (ATILF, 2019)	2015	S. Ollinger, C. Benzitoun, E. Jacquey, U. Fleury	LGPL-LR
TLFPhraseo (ATILF, 2016)	2016	E. Jacquey, J. Humbert	CC BY-NC-SA
Apertium RDF Graph (Villegas *et al.*, 2016)	2016	M. Villegas, M. Melero, N. Bel, J. Gracia	CC BY-SA
ReSyf (Gala *et al.*, 2013)	2018	N. Gala, M. Billami, T. François, C. Fairon, D. Bernhard	CC BY-NC
Nomage (Balvet *et al.*, 2011)	2019	A. Balvet, L. Barque, M. Condette, R. Huyghe, A. Jugnet, R. Marin, A. Merlo, P. Haas	LGPL-LR
VerNom (Missud *et al.*, 2020)	2020	A. Missud, P. Amsili, F. Villoing	CC BY-NC-SA
Holinet (Prost, 2022)	2022	J-P. Prost	CC BY
Lexique4linguists (Schalchli, 2022)	2022	G. Schalchli	CC BY

TABLE 1 – Liste des 31 lexiques sélectionnés du plus ancien au plus récent.

1. https://lremap.elra.info/?type=Lexicon&availability=Freely+Available&languages=french, consultée le 16 mai 2023.
2. https://www.ortolang.fr/market/lexicons, consultée le 16 mai 2023.

Outre les lexiques, nous observons également les occurrences de trois plongements lexicaux statiques disponibles pour le français : `Word2vec` (Mikolov *et al.*, 2013; Fauconnier, 2015), `FastText` (Bojanowski *et al.*, 2017; Grave *et al.*, 2018) et `GloVe` (Pennington *et al.*, 2014). Grâce à leurs représentations vectorielles codant des informations linguistiques, les plongements lexicaux constituent une autre forme de ressource lexicale, de plus en plus utilisés en TAL depuis leur apparition en 2013.

2.2 Corpus TALN 2001-2022

Notre corpus est composé d'articles de TALN de 2001 à 2022 extraits d'ACL Anthology au format PDF, convertis au format XML avec GROBID (Lopez, 2008 2023). Au total, 2 176 articles en PDF et 2 174 articles en XML ont été récupérés, deux articles de 2010 n'ayant pas été convertis. La conversion pouvant être défectueuse, une vérification semi-automatique a été faite en deux temps : nous avons déterminé si la balise *lang* correspond à celle du français ou de l'anglais et si des mots-outils de la langue concernée sont présents dans le corps de l'article. Nous avons ainsi recensé 80 articles mal convertis dont 39 pour l'année 2013 et 24 pour l'année 2001 [4]. Par ailleurs, nous avons fait le choix de retirer les articles JEP, invités, tutoriels, DEFT et des ateliers, ceux-ci pouvant biaiser les résultats. En effet, en 2014, deux ateliers sur les ressources ont eu lieu (FondamenTAL et RLTLN), qui ont largement augmenté les chiffres obtenus. Notre corpus final contient 1 511 articles exploitables au format XML (cf. Annexes).

Afin d'observer l'utilisation des lexiques sélectionnés dans les articles, les occurrences de chaque ressource dans le corps de l'article ont été extraites automatiquement. Nous avons décidé de considérer uniquement le corps de l'article avec la balise *body* pour éviter les biais dûs aux occurrences dans les résumés et la bibliographie. En amont, des pré-traitements simples tels qu'une suppression de certaines ponctuations, des majuscules et une tokenisation ont été appliqués. Nous avons également veillé à normaliser autant que faire se peut les noms des ressources. Par exemple, *glaff* est remplacé par *glàfff*, plus fréquent. Au total, seuls 231 articles présentent au moins un des lexiques considérés, ce qui représente environ de 15 % du corpus entier. De plus, deux lexiques n'apparaissent dans aucun article du corpus : `TLFPhraseo` et `Lexique4linguists`. Concernant les plongements lexicaux, 108 articles présentent au moins une des trois ressources entre 2014 et 2022.

En raison des difficultés à extraire des informations présentes à partir d'occurrences brutes, nous avons décidé de faire une annotation manuelle sur les articles présentant au moins un lexique ou des plongements lexicaux. Pour chaque article et pour chaque ressource présente dans celui-ci, on attribue une classe parmi les cinq suivantes :

— **Construction-Extension** : l'article présente comment la ressource a été élaborée ou traite de son extension/amélioration.
— **Utilisation** : l'article présente une expérience où la ressource est utilisée pour une tâche spécifique (construction d'une autre ressource, étiquetage morpho-syntaxique...).
— **Comparaison** : l'article utilise la ressource comme une référence pour faire une comparaison avec une autre ressource.
— **Mention** : l'article ne fait que mentionner la ressource (état de l'art).
— **Erreur** : l'article présente une fausse occurrence (à cause d'une erreur dans le fichier XML),

3. Pour le Lexique-Grammaire et DELA, les personnes considérées sont des personnes passées par le LADL ayant utilisé la ressource comme L. Danlos, E. Tolone, S. Voyatzi, T. Nakaruma, E. Laporte, S. Paumier. Cette liste n'est pas exhaustive.

4. Nous notons que le corpus ACL Anthology (Rohatgi, 2022) présente les mêmes articles défectueux. L'erreur de conversion provient vraisemblablement d'un problème dans les articles d'origine.

l'occurrence ne correspond pas au nom de la ressource (Ex : « wolf » peut désigner un nom propre), la ressource ne concerne pas le français (Ex : `FrameNet` de l'anglais, plongements lexicaux de `word2vec` ou lexiques multilingues utilisés pour une autre langue), l'article utilise l'outil et non la ressource (modèle `word2vec` et non les plongements lexicaux existants).

L'annotation a été effectuée en séparant les lexiques des plongements lexicaux. Tous les articles présentant un lexique ont été annotés par un des auteurs (annotateur 0). Un accord inter-annotateur a été calculé sur un échantillon de 20 articles avec deux paires d'annotateurs (annotateur 0 - annotateur 1, annotateur 0 - annotateur 2). Sur 20 articles, les annotateurs sont en accord sur 19 d'entre eux. Les articles contenant des plongements lexicaux ont été annotés par trois annotateurs. Un accord inter-annotateurs a été calculé de la même manière que pour les lexiques et montre que sur 20 articles, les annotateurs sont en accord sur 17 d'entre eux.

La figure 1 présente la distribution dans le temps des lexiques en fonction du type d'usage. Nous pouvons observer que les articles présentant les lexiques augmentent progressivement de 2004 à 2014, où un pic est atteint avec 27 articles. Cette année-là, plusieurs ressources voient le jour comme le `French FrameNet` (Candito *et al.*, 2014) et `Démonette` (Hathout & Namer, 2014). Cette dernière étant une fusion de ressources existantes, le nombre de ressources utilisées augmente également. D'autres ressources font également l'objet d'extension comme le `RL-Fr` (Lux-Pogodalla, 2014) ou `JeuxDeMots` (Lafourcade *et al.*, 2014). Après 2014, le nombre d'articles diminue mais se maintient autour d'une dizaine d'articles jusqu'en 2021 où l'on en compte seulement deux.

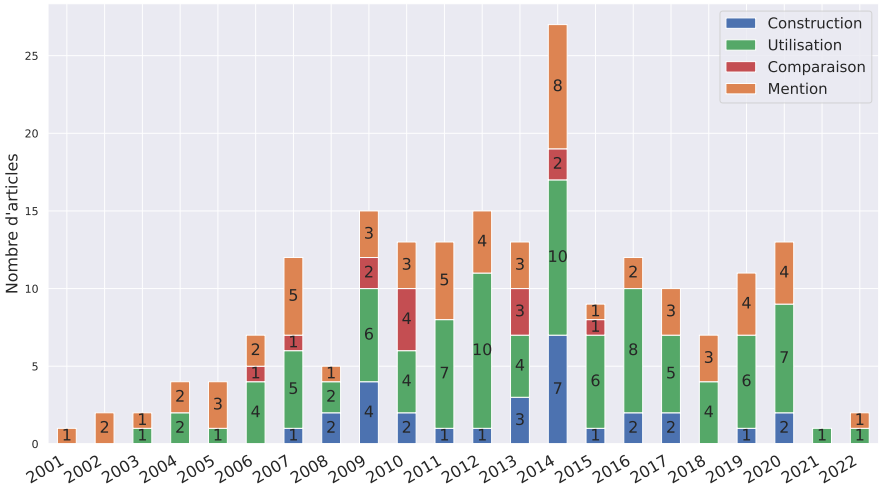


FIGURE 1 – Distribution dans le temps du nombre d'articles présentant au moins un lexique selon le type d'usage.

Concernant les plongements lexicaux, nous avons veillé à faire la distinction entre la ressource et l'outil. En effet, nous avons annoté en « Construction » le cas où les auteurs génèrent des plongements lexicaux et les mettent à disposition, contrairement au cas où le modèle est utilisé pour une application précise sans redistribution des plongements générés. Par ailleurs, comme pour les lexiques multilingues, nous ne prenons en compte que les articles traitant du français. Ils sont au nombre de 23 contre 50 pour l'anglais. Nous notons que cette vérification ne s'est pas faite de manière triviale en raison d'un certain nombre d'articles ne mentionnant pas explicitement la langue traitée (Ducel *et al.*, 2022).

3 Pour quoi et par qui sont utilisées les ressources lexicales ?

3.1 Des applications de différents types

La figure 2 présente la distribution en type d'usage pour 16 lexiques, telle qu'identifiée manuellement. Par souci de lisibilité, nous ne présentons pas les 15 lexiques restants, leurs occurrences étant strictement inférieures à 2, toutes classes confondues. Les ressources les plus présentes dans le corpus sont le Lefff (59 articles), JeuxDeMots (36 articles), Lexique-Grammaire (25 articles), Morphalou (21 articles) et WOLF (20 articles). En termes d'utilisation, le Lefff se distingue significativement des autres avec 36 articles. Morphalou et WOLF apparaissent tous deux dans une vingtaine d'articles cependant WOLF ne fait l'objet d'une utilisation que dans cinq d'entre eux tandis que Morphalou est utilisé dans dix articles. Les ressources les plus utilisées sont des lexiques présentant des contenus linguistiques différents. Si le Lefff et Morphalou sont des lexiques des formes fléchies du français, JeuxDeMots présente principalement des relations sémantiques entre les mots (voire des termes) sous la forme d'un réseau lexico-sémantique. De ce fait, ces ressources ne sont pas exploitées pour les mêmes applications.

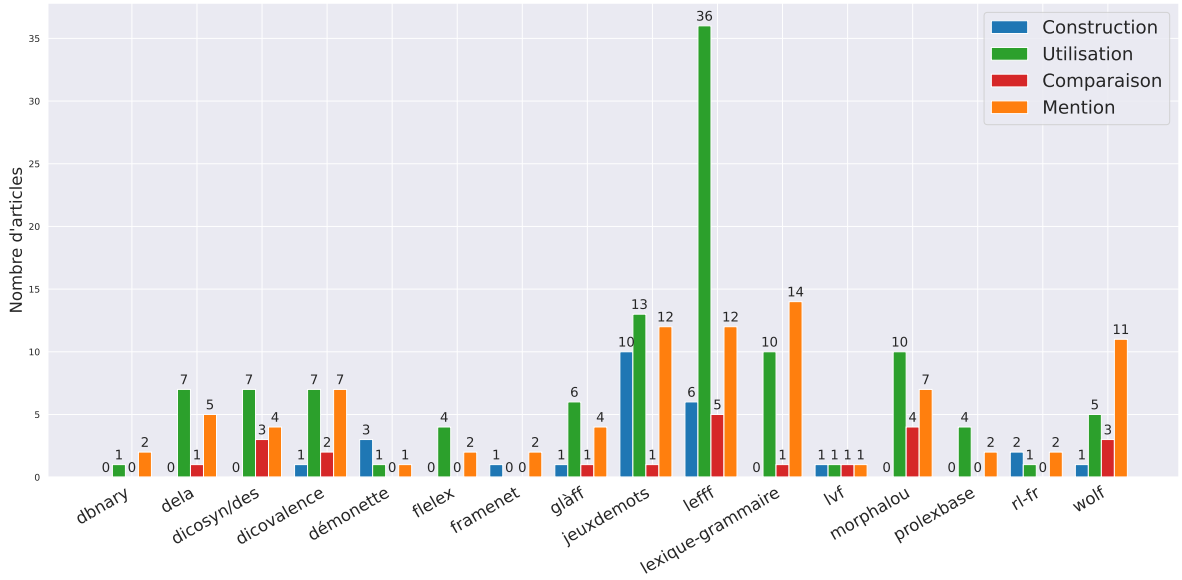


FIGURE 2 – Distribution des ressources les plus fréquentes selon le type d'usage.

Dans le domaine de la syntaxe et de la morphologie, le Lefff a été utilisé dans la construction de lexiques spécifiques (Strnadová & Sagot, 2011; Sagot, 2019), l'intégration à des étiqueteurs en parties du discours tels que MElt (Denis & Sagot, 2012), LGTagger (Constant & Sigogne, 2011) ou Macaon (Nasr *et al.*, 2011), la traduction automatique (Bawden, 2017; Burlot & Yvon, 2018), la fouille d'erreurs dans des sorties d'analyseurs syntaxique (Sagot & Villemonte De La Clergerie, 2006) ou la correction de textes bruités (Baranes, 2012).

En sémantique, JeuxDeMots est principalement exploité pour des travaux sur les relations sémantiques : l'hyperonymie et l'hyponymie (Gosset *et al.*, 2021), la synonymie (Francois *et al.*, 2016) ou la méronymie (Morlane-Hondère & Fabre, 2012). La ressource présente la particularité d'apparaître dans approximativement le même nombre d'articles qui décrivent sa construction et qui l'utilisent. Cela s'explique par les différentes extensions qu'a connues la ressource comme le jeu ColorIt permettant

de faire des associations de mots et de couleurs (Lafourcade *et al.*, 2014).

La figure 3 montre que les plongements lexicaux les plus utilisés sont `FastText` et `Word2vec`. Dans notre corpus, ils interviennent dans des tâches telles que la classification et l'extraction de relations (Khaldi *et al.*, 2020; Randriatsitohaina & Hamon, 2020), la reconnaissance d'entités nommées (Dupont, 2017) et la classification de questions (Eshkol-Taravella *et al.*, 2022). Nous pouvons remarquer que les plongements lexicaux en tant que ressource sont relativement peu réutilisés. En effet, le fait qu'ils codent des informations linguistiques leur donne un statut de ressource lexicale mais leur utilisation s'avère différente des lexiques symboliques dans la mesure où ce sont davantage les modèles qui sont utilisés pour générer des plongements lexicaux propres à une application.

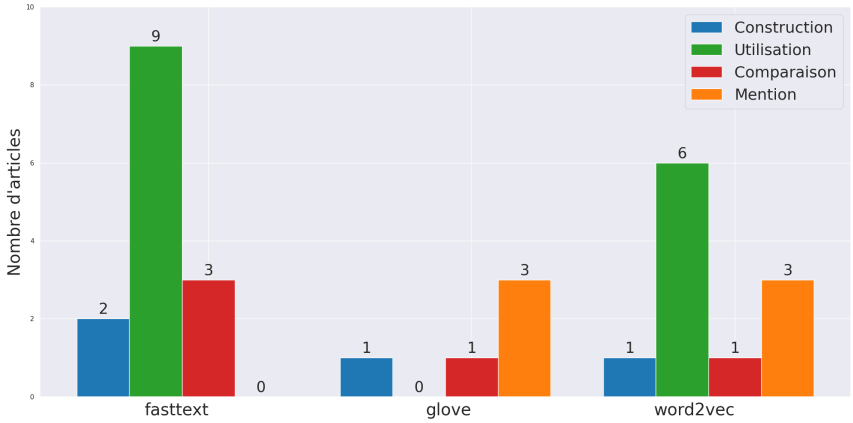


FIGURE 3 – Distribution des plongements lexicaux selon le type d'usage.

3.2 Réutilisation externe ou interne ?

Au-delà des applications, nous nous intéressons également aux personnes utilisant les ressources. Nous cherchons à déterminer si les utilisations d'une ressource sont menées par des personnes impliquées dans la construction de celle-ci. Pour ce faire, nous considérons une utilisation comme externe les cas où les auteurs de l'article ne font pas partie de la liste de créateurs donnée dans le tableau 1 [5].

Dans la figure 4, nous observons les proportions d'utilisations internes et externes pour les 15 ressources précédemment extraites. La figure montre que les ressources ont tendance à être réutilisées par des personnes extérieures à la ressource, à l'exception de `JeuxDeMots` où huit articles sur 13 comptent Mathieu Lafourcade dans leurs auteurs. Nous remarquons également que presque un tiers des utilisations du `Lefff` sont menées par un des créateurs de la ressource et que la moitié des utilisations du Lexique-Grammaire sont internes.

Observer si une ressource est utilisée par des personnes extérieures permet de questionner la réutilisabilité des ressources et ses facteurs déterminants. Inspirés par Cohen *et al.* (2005) décrivant les critères de réutilisabilité pour les corpus bio-médicaux, nous pouvons supposer que la taille de la ressource, sa couverture, son format, son âge ou la facilité de sa prise en main peuvent être des potentiels critères.

5. Nous précisons que les informations sur les concepteurs des ressources peuvent contenir des erreurs et que les articles RECITAL ne mentionnant pas les encadrants des étudiants peuvent avoir une influence sur les chiffres.

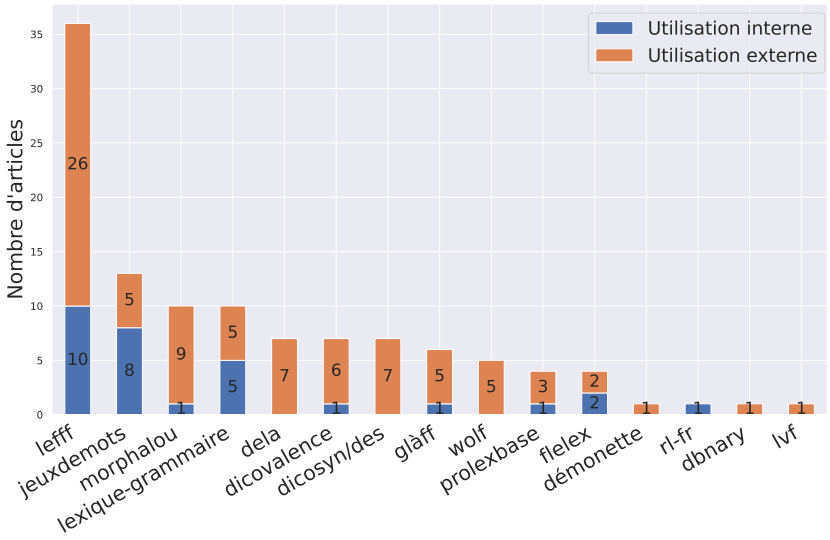


FIGURE 4 – Proportions de réutilisations internes ou externes pour les 15 ressources utilisées.

4 Discussion

À travers cette étude, nous observons que les ressources lexicales ont été utilisées de manière relativement stable depuis 20 ans. Un des exemples majeurs est le `Lefff`, utilisé près d'une fois par an depuis 2006 et considéré comme une ressource de référence pour le français. Elle doit ce statut notamment aux efforts de fusion de plusieurs ressources existantes (`Dicovalence`, `LVF`, `Lexique-Grammaire`) lui octroyant ainsi une meilleure couverture et complétude. Nous observons également que l'effet de l'âge a une influence sur nos chiffres, les ressources les plus récentes présentant une utilisation moindre. Par ailleurs, depuis plusieurs années, une nouvelle forme de ressource apparait avec les plongements lexicaux, de plus en plus présents du fait des récentes avancées des modèles de langue. Nous notons toutefois que les résultats ne montrent pas une massive utilisation des plongements lexicaux en tant que ressource, ces derniers étant davantage utilisés en tant qu'outil avec également la montée des modèles de langue tels que BERT (Devlin *et al.*, 2019). Bien que ces modèles de langue connaissent un fort succès depuis une dizaine d'années, des recherches visent de plus en plus à y intégrer des ressources symboliques comme des lexiques ou des bases de connaissances dans le but d'encoder des informations linguistiques externes et ainsi améliorer leur interprétabilité (Roy & Pan, 2020; Yang *et al.*, 2021).

Au-delà de leur place dans le domaine du TAL, nous pouvons souligner que certaines ressources lexicales ne sont pas seulement destinées au TAL mais également à la linguistique (`Lexique4linguists` (Schalchli, 2022)) ou à la didactique du FLE avec par exemple la création de `FLELex` (François *et al.*, 2014).

En observant la dynamique des ressources lexicales dans le temps dans les actes de TALN, nous avons cherché à mettre en évidence des critères permettant une meilleure réutilisabilité. Toutefois, limités probablement par un faible échantillon (moins de 300 articles) et des pertes durant la phase de conversion, nous ne pouvons noter aucune réelle tendance dans cette étude qui mériterait d'être étendu à une conférence plus conséquente telle que la conférence internationale de ressources langagières, LREC. Tous les scripts et annotations relatifs à l'article sont mis à disposition [6].

6. https://gitlab.inria.fr/papers2/taln2023

Références

ATILF (2016). Tlfphraseo. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

ATILF (2019). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BALVET A., BARQUE L., CONDETTE M.-H., HAAS P., HUYGHE R., MARIN R. & MERLO A. (2011). La ressource nomade. confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus [the nomage resource. compare theoretical expectations with observations of linguistic behavior of nominalizations in corpus]. *Traitement Automatique des Langues*, **52**(3), 129–152.

BARANES M. (2012). Vers la correction automatique de textes bruités : Architecture générale et détermination de la langue d'un mot inconnu (towards automatic spell-checking of noisy texts : General architecture and language identification for unknown words) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 3 : RECITAL*, p. 95–108, Grenoble, France : ATALA/AFCP.

BAWDEN R. (2017). Machine translation of speech-like texts : Strategies for the inclusion of context. In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, p. 1–14, Orléans, France : ATALA.

BLACHE P. & RAUZY S. (2008). Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 282–291, Avignon, France : ATALA.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (TACL)*, **5**, 135–146. DOI : 10.1162/tacl_a_00051.

BURLOT F. & YVON F. (2018). Évaluation morphologique pour la traduction automatique : adaptation au français (morphological evaluation for machine translation : Adaptation to French). In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, p. 61–74, Rennes, France : ATALA.

CANDITO M., AMSILI P., BARQUE L., BENAMARA F., DE CHALENDAR G., DJEMAA M., HAAS P., HUYGHE R., MATHIEU Y. Y., MULLER P., SAGOT B. & VIEU L. (2014). Developing a French FrameNet : Methodology and First results. In *LREC - The 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Islande. HAL : hal-01022385.

CHOI H.-S. (2022). État de l'art : Liage de ressources lexicales du français (state of the art : Linking French lexical resources). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, p. 55–68, Avignon, France : ATALA.

CLÉMENT L., SAGOT B. & LANG B. (2004). Morphology based automatic acquisition of large-coverage lexica. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbonne, Portugal : European Language Resources Association (ELRA).

COHEN K. B., FOX L., OGREN P. V. & HUNTER L. (2005). Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases : Mining Biological Semantics*, p. 38–45, Detroit : Association for Computational Linguistics.

CONSTANT M. & SIGOGNE A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the Workshop on Multiword Expressions : from Parsing and Generation to the Real World*, p. 49–56, Portland, Oregon, USA : Association for Computational Linguistics.

CONSTANT M. & TOLONE E. (2010). A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In M. D. GIOIA, Éd., *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie*, volume 1 de *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, p. 79–93. Aracne. ISBN 978-88-548-3166-7.

DANLOS L., NAKAMURA T. & PRADET Q. (2014). Vers la création d'un verbenet du français. In *Atelier FondamenTAL, TALN 2014*.

DENIS P. & SAGOT B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation*, **46**(4), 721–736. DOI : 10.1007/s10579-012-9193-0.

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.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les verbes français.* Larousse-Bordas, Paris, France.

DUCEL F., FORT K., LEJEUNE G. & LEPAGE Y. (2022). Langues par défaut ? analyse contrastive et diachronique des langues non citées dans les articles de TALN et d'ACL (contrastive and diachronic study of unmentioned (by default ?) languages in TALN and ACL we study the application of the #BenderRule in natural language processing articles, taking into account a contrastive and a diachronic dimensions, by examining the proceedings of two NLP conferences, TALN and ACL, over time). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 144–153, Avignon, France : ATALA.

DUPONT Y. (2017). Exploration de traits pour la reconnaissance d'entités nommées du français par apprentissage automatique (feature exploration for French named entity recognition with machine learning). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. 19es REncontres jeunes Chercheurs en Informatique pour le TAL (RECITAL 2017)*, p. 42–55, Orléans, France : ATALA.

ENGUEHARD C., KANÉ S., MANGEOT M., MODI I. & SANOGO M. L. (2012). Vers l'informatisation de quelques langues d'afrique de l'ouest (towards the computerization of some west-african languages) [in French]. In *JEP-TALN-RECITAL 2012, Workshop TALAf 2012 : Traitement Automatique des Langues Africaines (TALAf 2012 : African Language Processing)*, p. 27–40, Grenoble, France : ATALA/AFCP.

ESHKOL-TARAVELLA I., BARBEDETTE A., LIU X. & SOUMAH V.-G. (2022). Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral (automatic classification of spontaneous vs). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 305–314, Avignon, France : ATALA.

FAUCONNIER J.-P. (2015). French word embeddings.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. Bradford Books.

FRANCOIS T., BILLAMI M. B., GALA N. & BERNHARD D. (2016). Bleu, contusion, ecchymose : tri automatique de synonymes en fonction de leur difficulté de lecture et compréhension (automatic ranking of synonyms according to their reading and comprehension difficulty). In *Actes de la*

conférence conjointe JEP-TALN-RECITAL 2016. volume 2 : TALN (Articles longs), p. 15–28, Paris, France : AFCP - ATALA.

FRANÇOIS T., GALA N., WATRIN P. & FAIRON C. (2014). FLELex : a graded lexical resource for French foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande. HAL : hal-01758123.

GALA N. (2013). Ressources lexicales mono- et multilingues : une évolution historique au fil des pratiques et des usages. In *Ressources lexicales : contenu, évaluation, utilisation, évaluation.*, volume 30 de *Lingvisticæ Investigationes Supplementa*, p. 1–42. John Benjamins Publishing. HAL : hal-03203895.

GALA N., FRANÇOIS T. & FAIRON C. (2013). Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *eLex - Electronic Lexicography*, Tallinn, Estonie. HAL : hal-03194427.

GOSSET C., BOUMEDYEN BILLAMI M., LAFOURCADE M., BORTOLASO C. & DERRAS M. (2021). Extraction automatique de relations sémantiques d'hyperonymie et d'hyponymie dans un corpus métier (automatic extraction of hypernym and hyponym relations in a professional corpus). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 162–170, Lille, France : ATALA.

GRAVE E., 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)*, Miyazaki, Japon : European Language Resources Association (ELRA).

GROSS M. (1975). *Méthodes en syntaxe*. Hermann, Paris, France.

HATHOUT N. & NAMER F. (2014). Démonette, a french derivational morpho-semantic network. *Linguistic Issues in Language Technology*, **11**(5), 125–168.

KHALDI H., ABDAOUI A., BENAMARA F., SIGEL G. & AUSSENAC-GILLES N. (2020). Classification de relations pour l'intelligence économique et concurrentielle (relation classification for competitive and economic intelligence). In *Actes de la 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 2 : Traitement Automatique des Langues Naturelles*, p. 27–39, Nancy, France : ATALA et AFCP.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LAFOURCADE M. & LE BRUN N. (2020). Jeuxdemots : Un réseau lexico-sémantique pour le français, issu de jeux et d'inférences. *Revue Lexique*, **27**, 47–86.

LAFOURCADE M., LE BRUN N. & ZAMPA V. (2014). Les couleurs des gens. In *TALN : Traitement Automatique des Langues Naturelles*, Marseille, France. HAL : lirmm-01471671.

LOPEZ P. (2008–2023). Grobid. https://github.com/kermitt2/grobid.

LUX-POGODALLA V. (2014). Integrating lexicographic examples in a lexical network (intégration relationnelle des exemples lexicographiques dans un réseau lexical) [in French]. In *Proceedings of TALN 2014 (Volume 2 : Short Papers)*, p. 586–591, Marseille, France : Association pour le Traitement Automatique des Langues.

LUX-POGODALLA V. & POLGUÈRE A. (2011). Construction of a French Lexical Network : Methodological Issues. In *First InternationalWorkshop on Lexical Resources, WoLeR 2011*, p. 54–61, Ljubljana, Slovénie. HAL : hal-00686467.

MAKS I., IZQUIERDO R., FRONTINI F., AGERRI R., VOSSEN P. & ANDONI AZPEITIA (2014). Generating polarity lexicons with wordnet propagation in 5 languages. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Islande : European Language Resources Association (ELRA).

MANGEOT M. & CHALVIN A. (2006). Dictionary building with the jibiki platform : the GDEF case. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genève, Italie.

MANGEOT-NAGATA M. (2016). Collaborative Construction of a Good Quality, Broad Coverage and Copyright Free Japanese-French Dictionary. *International Journal of Lexicography*, **31**(1), 78–112. DOI : 10.1093/ijl/ecw035, HAL : hal-01712271.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

MARIANI J., FRANCOPOULO G. & PAROUBEK P. (2019a). The nlp4nlp corpus (i) : 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00036.

MARIANI J., FRANCOPOULO G., PAROUBEK P. & VERNIER F. (2019b). The nlp4nlp corpus (ii) : 50 years of research in speech and language processing. *Frontiers in Research Metrics and Analytics*, **3**. DOI : 10.3389/frma.2018.00037.

MERTENS P. (2010). Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE Conversion vers un format utilisable en TAL. In *Conference Traitement Automatique des Langues Naturelles (TALN)*, Montréal, Canada.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, p. 746–751.

MISSUD A., AMSILI P. & VILLOING F. (2020). VerNom : une base de paires morphologiques acquise sur très gros corpus (VerNom : a French derivational database acquired on a massive corpus). In *Actes de la 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 2 : Traitement Automatique des Langues Naturelles*, p. 305–313, Nancy, France : ATALA et AFCP.

MORLANE-HONDÈRE F. & FABRE C. (2012). Étude des manifestations de la relation de méronymie dans une ressource distributionnelle (study of meronymy in a distribution-based lexical resource) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2 : TALN*, p. 169–182, Grenoble, France : ATALA/AFCP.

NASR A., BÉCHET F., REY J.-F., FAVRE B. & LE ROUX J. (2011). MACAON an NLP tool suite for processing word lattices. In *Proceedings of the ACL-HLT 2011 System Demonstrations*, p. 86–91, Portland, Oregon : Association for Computational Linguistics.

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. DOI : 10.3115/v1/D14-1162.

PLOUX S. & VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Revue TAL*, **39**, 161–182. HAL : halshs-00009433.

PROST J.-P. (2022). Integrating a phrase structure corpus grammar and a lexical-semantic network : the HOLINET knowledge graph. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 613–622, Marseille, France : European Language Resources Association.

RANDRIATSITOHAINA T. & HAMON T. (2020). Identification des problèmes d'annotation pour l'extraction de relations (identification of annotation problem for the relation extraction). In *Actes de la 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 2 : Traitement Automatique des Langues Naturelles*, p. 323–331, Nancy, France : ATALA et AFCP.

RAUZY S. & BLACHE P. (2007). Un lexique syntaxique des verbes du français : VfrLPL. 7 pages.

ROHATGI S. (2022). Acl anthology corpus with full text. Github.

ROY A. & PAN S. (2020). Incorporating extra knowledge to enhance word embedding. In C. BESSIERE, Éd., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, p. 4929–4935 : International Joint Conferences on Artificial Intelligence Organization. Survey track, DOI : 10.24963/ijcai.2020/686.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte. HAL : inria-00521242.

SAGOT B. (2019). Développement d'un lexique morphologique et syntaxique de l'ancien français (development of a morphological and syntactic lexicon of Old French). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, p. 265–274, Toulouse, France : ATALA.

SAGOT B. & FIŠER D. (2008). Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Maroc.

SAGOT B. & VILLEMONTE DE LA CLERGERIE É. (2006). Trouver le coupable : Fouille d'erreurs sur des sorties d'analyseurs syntaxiques. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, p. 288–297, Leuven, Belgique : ATALA.

SAJOUS F., HATHOUT N. & CALDERONE B. (2013). GLÁFF, un Gros Lexique Á tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 285–298, Les Sables d'Olonne, France.

SCHALCHLI G. (2022). Lexique4linguists. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

SÉRASSET G. (2015). DBnary : Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, **6**(4), 355–361. DOI : 10.3233/SW-140147, HAL : hal-00953638.

STRNADOVÁ J. & SAGOT B. (2011). Construction d'un lexique des adjectifs dénominaux (construction of a lexicon of denominal adjectives). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, p. 67–72, Montpellier, France : ATALA.

TRAN M. & MAUREL D. (2006). Prolexbase - un dictionnaire relationnel multilingue de noms propres. *Trait. Autom. des Langues*, **47**(3), 115–139.

TROUILLEUX F. (2012). Le DM, a French Dictionary for NooJ. In B. B. KRISTINA VUČKOVIĆ & M. SILBERZTEIN, Éds., *Automatic Processing of Various Levels of Linguistic Phenomena : Selected Papers from the NooJ 2011 International Conference*, p. 16–28. Cambridge Scholars Publishing. HAL : hal-00702348.

VAN DEN EYNDE K. & MERTENS P. (2006). Le dictionnaire de valence dicovalence : manuel d'utilisation.

VAN DEN EYNDE K. & MERTENS P. (2010). Le dictionnaire de valence dicovalence : manuel d'utilisation version 2.0.

VILLEGAS M., MELERO M., BEL N. & GRACIA J. (2016). Leveraging RDF graphs for crossing multiple bilingual dictionaries. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 868–876, Portorož, Slovénie : European Language Resources Association (ELRA).

YANG J., XIAO G., SHEN Y., JIANG W., HU X., ZHANG Y. & PENG J. (2021). A survey of knowledge enhanced pre-trained models. *arXiv preprint arxiv :2110.00269*.

Annexes

Type	Ressources LRE map	Ressources Ortolang
Lexiques monolingues	Dicovalence, Lefff, JeuxDeMots, LGLex, French FrameNet, LVF, Lexique-Grammaire, VerbeNet, ReSyf, WOLF, GLAFF, DELA, FLELex, Démonette	Démonette, Morphalou, DES, Lexique4linguists, Holinet, DM, RL-Fr, VerNom, Nomage, Prolexbase, Dicovalence, TLFPhraseo, MarsaLex, VfrLPL
Lexiques multilingues	Apertium RDF Graph, OpeNER-sentiment-lexicons, DBnary, DiLAF	Prolexbase
Lexiques non sélectionnés	Wikitionary, FreeLang, Free dictionary download, Terminesp LD	Dictionnaire informatisé des Mots d'Affect, DSR, Termes de base du diagnostic orthophonique, LGeRM
Lexiques non disponibles (liens cassés)	JournalisticNL11, EUROSENTIMENT, CorpusDRF, DeQue, tl_dv2_ladl par-lvf, V2R, MotaMot, Multilingual glossary of technical and popular medical terms, Bilingual Dictionaries, Verbnet like classification of French verbs	
Doublons	Lefff (x2), LGLex 3.3 (x2), Dicovalence 2, LG, FLELLex, WOLF, DBnary (x2)	
Total	44	19

TABLE 2 – Lexiques du français libres d'utilisation présents dans la LRE map et Ortolang.

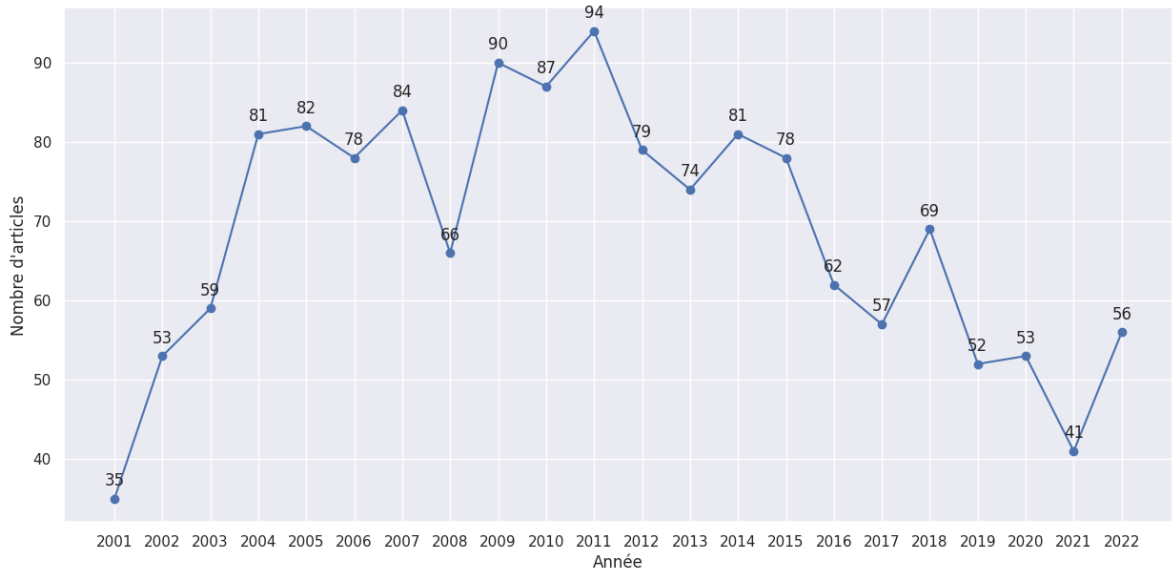


FIGURE 5 – Nombres d'article du corpus TALN 2001-2022 par année.

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

Attention sur les *spans* pour l'analyse syntaxique en constituants

Nicolas Floquet[1] Joseph Le Roux[1] Nadi Tomeh[1] Thierry Charnois[1]
(1) Laboratoire Informatique Paris Nord, 99 Av. Jean Baptiste Clément, 93430 Villetaneuse, France
floquet@lipn.univ-paris13.fr, leroux@lipn.fr, tomeh@lipn.fr,
thierry.charnois@lipn.univ-paris13.fr

RÉSUMÉ

Nous présentons une extension aux analyseurs syntaxiques en constituants neuronaux modernes qui consiste à doter les constituants potentiels d'une représentation vectorielle affinée en fonction du contexte par plusieurs applications successives d'un module de type transformer efficace (*pooling* par attention puis transformation non-linéaire). Nous appliquons cette extension à l'analyseur CRF de Zhang *et al.* (2020). Expérimentalement, nous testons cette extension sur deux corpus (PTB et FTB) avec ou sans vecteurs de mots dynamiques : cette extension permet d'avoir un gain constant dans toutes les configurations.

ABSTRACT

Attention over spans for syntactic parsing

We introduce an extension to recent neural constituency parsers that consists of providing potential constituents with a vector representation fine-tuned from the context by successive applications of an efficient transformer type module (pooling by attention, then a non-linear transformation). We implement this extension in the parser called CRF introduced by Zhang *et al.* (2020). We test this extension on two corpora, PTB and FTB, with or without dynamic word vectors : this extension achieves performance gains in all settings.

MOTS-CLÉS : Apprentissage profond, Attention, Analyse syntaxique en constituants.

KEYWORDS: Deep learning, Attention, Constituency Parsing.

1 Introduction

L'analyse syntaxique en constituants crée un arbre hiérarchique à partir d'une phrase d'entrée. Les feuilles de cet arbre sont les mots de la phrase et les nœuds internes sont les constituants. En tant que tâche fondamentale du TAL, l'analyse en constituants sert de base à de nombreuses applications en aval, notamment l'analyse sémantique ou l'extraction d'information, et il est donc crucial de produire correctement les structures syntaxiques nécessaires aux traitements ultérieurs. De nombreux modèles actuels (Stern *et al.*, 2017; Kitaev & Klein, 2018; Zhang *et al.*, 2020) fonctionnent en deux temps. Tout d'abord ils attribuent des scores aux différentes sous-chaînes (nous utiliserons le terme anglais *spans* dans la suite pour désigner les sous-chaînes) selon leur plausibilité d'être un constituant bien typé (groupe nominal, verbal, prépositionnel…). Ensuite un algorithme combinatoire, souvent CKY (Kasami, 1965; Baker, 1979), permet de construire l'arbre d'analyse de meilleure score. Ce score a en général une interprétation probabiliste, et l'analyse syntaxique revient donc dans ce cas à retourner l'arbre le plus probable (MAP, maximum a posteriori), ou qui combine les constituants

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 37–45.

les plus probables (MBR, *Minimum Bayes Risk*). La première étape, celle qui attribue les scores, est réalisée par une architecture neuronale construite à partir d'un module qui extrait des vecteurs caractéristiques des mots de la phrase (*extracteur*) puis un second module qui assemble les vecteurs de mots pour représenter les spans et qui leur attribue un score (*scoreur*). Un mécanisme de récurrence ou d'attention permet aux vecteurs de mots de prendre en compte le contexte, c'est-à-dire de s'affiner en fonction des autres mots dans la phrase.

Dans cet article nous étendons cette prise en compte du contexte aux spans. Pour cela, il faut d'abord *représenter* les spans, dans le cas des réseaux de neurones leur attribuer un vecteur réel caractéristique. Puis nous ajoutons des *transformers* dotés d'un mécanisme d'attention linéaire pour faire évoluer ces représentations en fonctions des autres spans de la forêt d'analyse (sans toutefois calculer cette forêt explicitement). Enfin à partir des représentations finales nous calculons les scores des spans. Expérimentalement, cette addition de transformers sur les représentations des spans dans un analyseur récent au plus haut de l'état de l'art actuel permet d'améliorer les performances sur plusieurs corpus de référence.

Dans la Section 2 nous présentons le modèle d'analyseur CRF que nous utilisons ainsi que l'extension que nous proposons pour représenter les spans. En Section 3 nous présentons l'évaluation sur deux corpus dans différents scénarios. Puis en Section 4 nous présentons quelques travaux connexes avant de conclure.

2 Modèle

L'analyse syntaxique en constituants consiste à décomposer une phrase en ses parties constituantes, qui sont organisées dans une structure hiérarchique basée sur leurs relations syntaxiques. Formellement, étant donné une phrase $\boldsymbol{x} = w_1 \dots w_n$, un arbre d'analyse $\boldsymbol{t}$ est un ensemble de syntagmes $(i, j, l) \in \boldsymbol{t}$ couvrant les mots $w_i \dots w_j$ et ayant une étiquette syntaxique $l \in \mathcal{L}$. Ils forment une segmentation hiérarchique qui couvre toute la phrase. Dans cet article, nous proposons une nouvelle représentation des syntagmes basée sur l'attention que nous intégrons dans l'analyseur CRF à deux étapes de Zhang *et al.* (2020).

2.1 Analyseur CRF à deux étapes

Modèle d'analyse et algorithmes. Notre analyseur de base construit $\boldsymbol{t}$ en deux étapes : la première génère un arbre non étiqueté $\mathbf{y} = \{(i, j)\}_{i<j, i,j \in [1,n]}$ contraint à l'ensemble de toutes les segmentations hiérarchiques valides de $\boldsymbol{x}$, appelé $\mathcal{Y}_{\boldsymbol{x}}$. La deuxième étape attribue des étiquettes à ses spans. L'analyseur fonctionne sur des arbres en Forme Normale de Chomsky où un segment *non étiqueté* (i, j) peut être formé en joignant deux spans adjacents (i, k) et $(k + 1, j)$.

Afin de comparer des arbres d'analyse alternatifs pour une entrée $\boldsymbol{x}$, le modèle utilise une fonction de notation d'ordre zéro qui se factorise en terme des scores individuels de spans : $s(\boldsymbol{y}, \boldsymbol{x}) = \sum_{(i,j)\in \boldsymbol{y}} s_c(i, j; \boldsymbol{x})$. Ces scores sont globalement normalisés pour obtenir une distribution sur les arbres valides : $\log p(\boldsymbol{y}|\boldsymbol{x}) = s_c(\boldsymbol{y}, \boldsymbol{x}) - A(\boldsymbol{x})$. La constante de normalisation $A(\boldsymbol{x}) = \log \sum_{\boldsymbol{y}\in\mathcal{Y}_{\boldsymbol{x}}} \exp(s(\boldsymbol{y}, \boldsymbol{x}))$ est calculée à l'aide de l'algorithme *inside* en temps $O(n^3)$. Pendant le décodage, l'algorithme CKY est utilisé pour trouver l'arbre optimal étant donné le modèle : $\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}\in\mathcal{Y}_{\boldsymbol{x}}} s(\boldsymbol{y}, \boldsymbol{x})$. CKY est presque identique à *inside* mais maximise au lieu de sommer sur les sous-arbres valides avec une complexité similaire. CKY peut également être utilisé avec le

décodage MBR (Minimum Bayes-Risk) en remplaçant $s_c(i, j; \boldsymbol{x})$ par la probabilité marginale du segment $p(i, j|\boldsymbol{x})$. Ces probabilités peuvent être calculées efficacement avec l'algorithme *outside*.

Dans la deuxième étape, une étiquette est sélectionnée pour chaque segment dans l'arbre décodé $\forall(i, j) \in \hat{\boldsymbol{y}} : \hat{l} = \arg\max_{l \in \mathcal{L}} s_l(i, j, l; \boldsymbol{x})$. La complexité de cette étape est de $O(n|\mathcal{L}|)$, car $\hat{\boldsymbol{y}}$ ne contient que $2n - 1$ spans qui doivent être scorés parmi les $O(n^2)$ spans de $\boldsymbol{x}$. Les probabilités des étiquettes $p(l|i, j; \boldsymbol{x})$ sont obtenues en normalisant localement les scores des étiquettes à l'aide de la fonction softmax.

Représentations et fonctions de score La fonction de score s_c opère sur les représentations des bornes des spans i et j. Chaque mot $w_i \in \boldsymbol{x}$ peut jouer le rôle d'une borne *gauche* ou *droite* avec une sémantique différente. Par conséquent, nous construisons une représentation spécialisée pour chaque rôle à l'aide d'un MLP (*Multilayer Perceptron*, Perceptron multicouche), similaire aux travaux de Stern *et al.* (2017) et Zhang *et al.* (2020). Autrement dit, pour chaque mot w_i, nous calculons $\boldsymbol{g}_i = \text{MLP}^g(\boldsymbol{h}_i)$ et $\boldsymbol{d}_i = \text{MLP}^d(\boldsymbol{h}_i)$ où $\boldsymbol{h}_i \in \mathbb{R}^k$ est une représentation contextualisée de w_i et $\boldsymbol{g}_i, \boldsymbol{d}_i \in \mathbb{R}^d$. La fonction de score $s_c(i, j) = \text{Biaffine}(\boldsymbol{g}_i, \boldsymbol{d}_j; \boldsymbol{W})$ est une fonction biaffine (Dozat & Manning, 2017) paramétrée par $\boldsymbol{W} \in \mathbb{R}^{d \times d}$. La représentation du mot $\boldsymbol{h}_i \in \mathbb{R}^k$ est calculée en deux étapes : 1) un encodage initial $\boldsymbol{e}_i \in \mathbb{R}^{k'}$ est obtenu soit en combinant des *embeddings* statiques tel que fournis par GloVe (Pennington *et al.*, 2014), ou fastText (Grave *et al.*, 2018) avec un encodage BiLSTM de la séquence de caractères d'entrée ; soit en utilisant un modèle de langue pré-entraîné tel que BERT (Devlin *et al.*, 2019) ; 2) la deuxième étape fait passer l'encodage initial par un BiLSTM (Dozat & Manning, 2017) à trois couches pour obtenir l'encodage final $\boldsymbol{h}_i$.

Des calculs similaires sont effectués pour scorer les étiquettes : $s_l(i, j, l; \boldsymbol{x}) = \text{Biaffine}(\bar{\boldsymbol{g}}_i, \bar{\boldsymbol{d}}_j; \boldsymbol{W}[l])$ où $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{L}| \times \bar{d} \times \bar{d}}$ regroupe tous les paramètres de s_l en un seul tenseur.

Apprentissage et implémentation Les paramètres des fonctions de scores sont appris pour minimiser la somme de deux fonctions de perte : la log-vraisemblance négative (NLL) des arbres de références dans une *treebank* $\mathcal{T}$ et la NLL des étiquettes individuelles des spans de ces arbres : $\mathcal{L}(x, y) = -\sum_{(\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{T}} \log p(\boldsymbol{y}|\boldsymbol{x}) - \sum_{\boldsymbol{y} \in \mathcal{T}; (i,j,l) \in \boldsymbol{y}} p(l|i, j; \boldsymbol{x})$.

L'apprentissage et le décodage sont particulièrement efficaces grâce à une implémentation *batchifiée* des algorithmes inside et CKY pour un calcul direct sur GPU. L'algorithme outside nécessaire au calcul des marginaux $p(i, j|\boldsymbol{x})$ est implémenté efficacement en utilisant la différenciation automatique car ces probabilités correspondent aux gradients du constant de normalisation $A(\boldsymbol{x})$ (Eisner, 2016).

2.2 Représentations des spans raffinées par attention

Notre objectif est d'affiner les représentations vectorielles biaffines des spans en permettant à leurs vecteurs initiaux d'interagir par le biais de l'attention. Les encodeurs basés sur le transformer (Devlin *et al.*, 2019) sont l'architecture de référence pour l'auto-attention mais ont une complexité quadratique puisque, pour chaque vecteur d'entrée, les poids d'attention doivent être calculés sur l'ensemble de l'entrée. Cette complexité quadratique est prohibitive dans notre cas puisque nous avons $O(n^2)$ spans pour n mots ce qui porte la complexité du transformer à $O(n^4)$. Pour résoudre ce problème, nous avons recours aux transformers linéaires (Katharopoulos *et al.*, 2020a; Choromanski *et al.*, 2021), une famille de transformers efficaces qui utilisent des noyaux pour calculer l'attention avec une

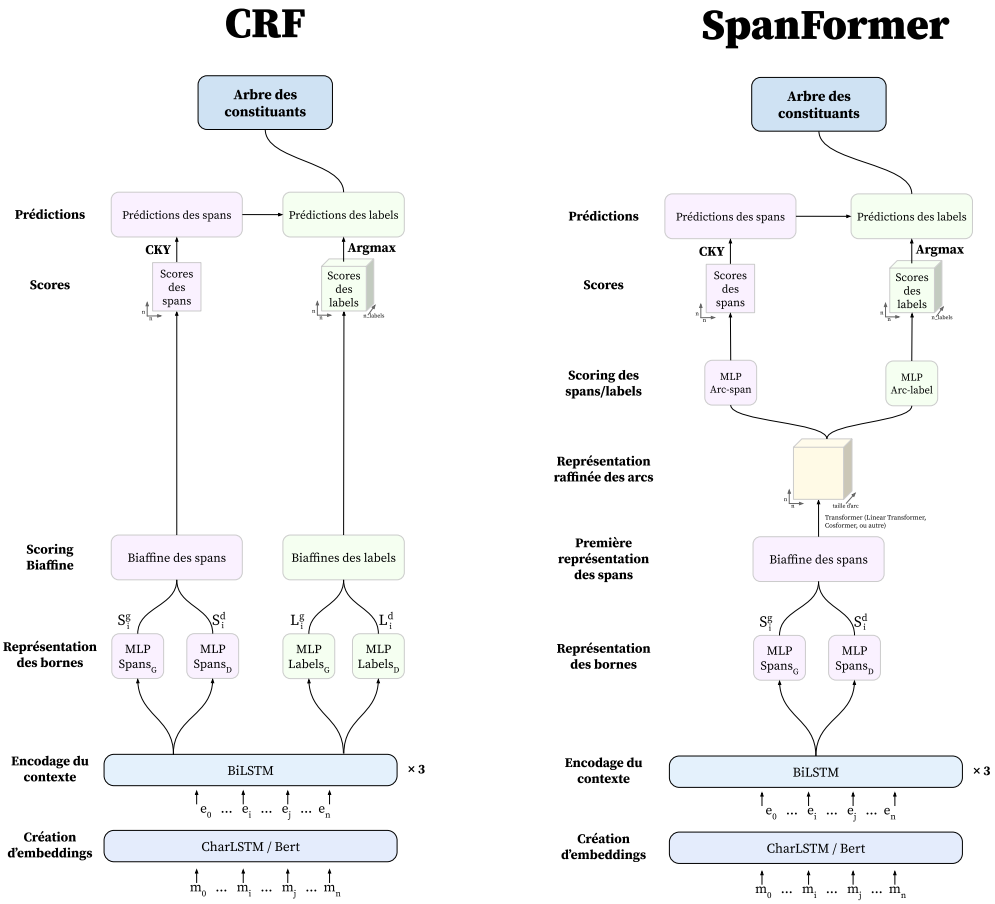


FIGURE 1 – Comparaison entre CRF et notre modèle SpanFormer, schéma adapté de la figure 2 de Zhang *et al.* (2020)

complexité spatio-temporelle linéaire. Plus précisément, nous empruntons la couche NormAttention à TransNormer (Qin *et al.*, 2022), conçue pour résoudre le problème des gradients non bornés responsables d'instabilités dans l'apprentissage des transforers linéaires à base de noyaux.

Étant donné les représentations initiales des spans $\mathbf{H}$, nous commençons par calculer les requêtes $\mathbf{Q} = \mathbf{H}\boldsymbol{W}_Q$, les clés $\mathbf{K} = \mathbf{H}\boldsymbol{W}_K$ et les valeurs $\mathbf{V} = \mathbf{H}\boldsymbol{W}_V$, tous des vecteurs dans $\mathbb{R}^d$. Contrairement aux transformers standards qui calculent l'attention comme $\text{Softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$ qui a des gradients non bornés et une complexité quadratique (Qin *et al.*, 2022), nous calculons la version linéaire de l'attention et utilisons LayerNorm (Ba *et al.*, 2016) pour le borner : $\text{LayerNorm}(\mathbf{Q}(\mathbf{K}^\top\mathbf{V}))$. Les autres opérations du bloc transformer sont standards, y compris les opérations de *skip* connexion et de perceptron multicouches. Nous utilisons plusieurs têtes d'attention et empilons plusieurs couches de transformer pour calculer les représentations finales des spans.

3 Expériences

Jeux de données Nos expériences sont effectuées sur le Penn TreeBank (PTB) en anglais et le corpus en français FTB (Abeillé *et al.*, 2000) de SPMRL (Seddah *et al.*, 2013) afin d'avoir des

résultats sur plusieurs langues et sur des *treebanks* de différentes tailles. Nous suivons la séparation usuelle train/dev/test et nous entraînons nos modèles sur des GPUs Nvidia A40. Nous avons utilisé les vecteurs de mots anglais Glove (Pennington *et al.*, 2014) ou BERT large cased (Devlin *et al.*, 2018) sur le PTB et les vecteurs de mots français fastText (Grave *et al.*, 2018) ou XLM-RoBERTa large (Conneau *et al.*, 2019) sur SPMRL.

Évaluation Nous reprenons les conditions d'évaluation de Zhang *et al.* (2020), comme eux nous utilisons donc la précision, le recall et le F-score (P/R/F) des arbres étiquetés avec l'outil d'évaluation du parenthésage EVALB, avec les configurations standards d'évaulation pour le PTB.

Paramètres Nous reprenons les paramètres de Zhang *et al.* (2020), nous utilisons aussi un charLSTM pour nos représentations initiales des mots, pour GloVe sur le PTB nous avons des *char embeddings* de dimension 50, les *word embeddings* et une sortie de charLSTM sont aussi de taille 100. (Pour fastText sur SPMRL, les *word embeddings* sont de taille 300). Les *dropout* valent 0.33 et nous utilisons aussi une taille de batch de 5000 tokens.

Pour les configurations avec GloVe/fastText, la patience est de 100, c'est à dire que notre entraînement s'arrête si le F-score du dev n'augmente pas pendant 100 époques. Avec BERT le nombre d'époques est fixé à 10. Pour BERT/XLM-RoBERTa, nous avons repris exactement les configurations données et nous avons aussi ajouté 3 arguments visiblement manquants au fichier de configuration, *weight_decay*, *decay* et *decay_steps* valant 0.01, 0.75 et 5000 respectivement. Le *learning rate* reste de 0.002 pour GloVe/fastText et 5×10^{-5} pour BERT/XLM-RoBERTa. Le *learning rate* du transformer est de 0.001.

3.1 Résultats

Comme on peut le voir dans la table 1, notre approche donne les meilleurs résultats pour chaque configuration sur une moyenne de 10 expériences, notamment avec une amélioration importante pour le FTB avec fastText.

Tailles des vecteurs de spans La taille des vecteurs représentant les spans doit être assez grande pour qu'ils soient porteurs d'information, mais assez petite pour limiter le coût en mémoire des $O(n^2)$ spans. Nos expériences ont montré qu'une taille entre 20 et 60 donne de bons scores, les résultats de la table 1 sont obtenus avec des tailles de 48, et 3 têtes d'attention pour le transformer. Nous avons obtenus de meilleurs résultats en produisant de grands spans de taille 480 avec la fonction biaffine, et en les projetant vers une taille plus petite de 48 avant de les raffiner avec le transformer.

Profondeur Nous avons testé nos modèles avec 1 à 4 couches de transformers et nous avons remarqué qu'en moyenne les scores n'augmentent pas significativement avec plus de couches, et qu'une suffit amplement.

Vitesse et mémoire Notre modèle SpanFormer est légèrement plus lent que CRF à l'entraînement, ce qui est dû au transformer, bien qu'efficace il a un coup conséquent qui fait qu'on traite 1080 phrases par seconde (p/s), contre 1319 pour CRF, soit 75% de la vitesse initiale. En revanche, lors de l'évaluation et du test, les vitesses sont plus similaires avec $607p/s$ pour SpanFormer et $624p/s$ pour CRF. La demande en mémoire est elle nettement accrue, typiquement un modèle avec une taille de span de 48 et 1 couche nécessitera environ 10Go de mémoire pour l'entraînement quand CRF n'avait besoin que de 5Go.

	Dev				Test			
	P	R	F	σ(F)	P	R	F	σ(F)
PTB (CharLSTM + GloVe)								
CRF (Zhang *et al.*, 2020)	94.03	94.25	94.14	0.07	94.14	93.90	94.02	0.06
Kitaev & Klein (2018)∗	–	–	–		93.90	93.20	93.55	
SpanFormer	94.16	94.17	**94.17**	0.09	94.34	93.92	**94.13**	0.09
PTB (BERT large cased)								
CRF (Zhang *et al.*, 2020)	95.78	95.89	95.83	0.06	96.01	95.55	95.78	0.08
Kitaev *et al.* (2019)∗	–	–	–		95.73	95.46	95.59	
Yang & Deng (2020)∗	–	–	–		96.04	95.55	95.79	
SpanFormer	95.80	95.90	**95.85**	0.04	96.02	95.58	**95.80**	0.07
FTB (CharLSTM + fastText)								
CRF (Zhang *et al.*, 2020)	84.48	85.23	84.86	0.09	83.94	84.59	84.26	0.15
SpanFormer	84.79	85.12	**84.96**	0.07	84.29	84.48	**84.39**	0.08
FTB (XLM-RoBERTa large cased)								
CRF (Zhang *et al.*, 2020)	88.50	88.82	88.66	0.08	88.87	89.15	89.01	0.25
SpanFormer	88.70	88.73	**88.76**	0.15	89.06	89.00	**89.04**	0.08

TABLE 1 – Comparaison de nos résultats sur CRF (Zhang *et al.*, 2020), notre SpanFormer, et les résultats publiés d'autres modèles la métrique déterminant le meilleur modèle est le F-Score (moyennes et écart-types de 10 expériences avec différentes valeurs initiales des paramètres pour nos expériences).
∗ Résultats reportés par Zhang *et al.* (2020)

4 Travaux connexes

Notre méthode de calcul de représentation des spans via une attention globale est similaire aux *Edge Transformers* de Bergen *et al.* (2021) qui avaient proposé une attention globale sur les arcs pour les analyses en dépendances. Outre les formalismes d'analyse (constituants vs. dépendances), nous soulignons deux différences majeures : premièrement nous n'utilisons pas de masque d'attention particulier [1] ce qui, deuxièmement, nous oblige à utiliser des transformers efficaces vu que le nombre d'items en relation d'attention est quadratique dans la taille d'une phrase [2], ce qui crée un problème d'utilisation mémoire si l'on veut utiliser des co-processeurs spécialisés, de type GPU. Nous utilisons des Transformers linéaires (Katharopoulos *et al.*, 2020b) qui, en évitant l'opération de normalisation des compatibilité requête-clé par softmax, garantissent une complexité en espace linéaire dans le nombre d'items, donc quadratique dans la taille de la phrase d'entrée.

D'autres travaux tels que ceux de Zaratiana *et al.* (2022) montrent l'intérêt d'enrichir les représentations des spans afin d'obtenir de meilleures prédictions dans les tâches en TAL, ce que nous faisons aussi à l'aide d'un transformer.

L'attention est devenue une étape essentielle dans les analyseurs [3] depuis l'article de Dozat & Manning

1. (Bergen *et al.*, 2021) utilisent l'attention *triangulaire* censée encourager la découverte de relations de transitivité logiques.
2. Il y a $2n^2 - n$ arcs en considérant une phrase de n mots complémentée d'une racine fictive.
3. Nous parlons de l'attention dans la partie des analyseurs propre à l'analyse syntaxique ou sémantique. La plupart des analyseurs utilisent comme la majorité des systèmes actuels en TAL des transformers pour plonger les mots dans des espaces

(2017). Nous notons que cette attention était déjà non-normalisée comme c'est le cas dans les transformers efficaces actuels. La plupart des analyseurs à l'état de l'art actuel, utilisent une attention particulière adaptée à l'analyse. Par exemple, Le Roux *et al.* (2019) utilisent une attention croisée spécialisée pour représenter les étapes d'une analyse en transitions, tandis que Kitaev & Klein (2018) travaillent sur la relation entre l'attention entre les contenus lexicaux et les contenus positionnels.

5 Conclusion

Nous proposons un modèle d'analyse syntaxique en constituants faisant usage de représentations vectorielles des spans raffinées à l'aide d'un transformer. Notre méthode mène à de meilleures performances sur toutes les configurations testées, ces améliorations sont conséquentes sur le FTB avec fastText. De plus, nous arrivons à ces performances avec un coût additionnel en temps raisonnable grâce à notre utilisation de transformers linéaires. Dans de futurs travaux nous essayerons de cibler l'attention pour mieux prendre en compte les incompatibilités entre certains spans et nous appliquerons notre approche dans différentes tâches du TAL.

Références

ABEILLÉ A., CLÉMENT L. & KINYON A. (2000). Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece : European Language Resources Association (ELRA).

BA L. J., KIROS J. R. & HINTON G. E. (2016). Layer normalization. *CoRR*, **abs/1607.06450**.

BAKER J. K. (1979). Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, **65**(S1), S132–S132. DOI : 10.1121/1.2017061.

BERGEN L., O'DONNELL T. J. & BAHDANAU D. (2021). Systematic generalization with edge transformers. In A. BEYGELZIMER, Y. DAUPHIN, P. LIANG & J. W. VAUGHAN, Éds., *Advances in Neural Information Processing Systems*.

CHOROMANSKI K. M., LIKHOSHERSTOV V., DOHAN D., SONG X., GANE A., SARLOS T., HAWKINS P., DAVIS J. Q., MOHIUDDIN A., KAISER L., BELANGER D. B., COLWELL L. J. & WELLER A. (2021). Rethinking attention with performers. In *International Conference on Learning Representations*.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, **abs/1911.02116**.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

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.

vectoriels.

DOZAT T. & MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

EISNER J. (2016). Inside-outside and forward-backward algorithms are just backprop (tutorial paper). In *Proceedings of the Workshop on Structured Prediction for NLP*, p. 1–17, Austin, TX : Association for Computational Linguistics. DOI : 10.18653/v1/W16-5901.

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

KASAMI T. (1965). *An efficient recognition and syntax analysis algorithm for context-free languages*. Rapport interne, Air Force Cambridge Research Laboratory.

KATHAROPOULOS A., VYAS A., PAPPAS N. & FLEURET F. (2020a). Transformers are rnns : Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20 : JMLR.org.

KATHAROPOULOS A., VYAS A., PAPPAS N. & FLEURET F. (2020b). Transformers are rnns : Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20 : JMLR.org.

KITAEV N., CAO S. & KLEIN D. (2019). Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 3499–3505, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1340.

KITAEV N. & KLEIN D. (2018). Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2676–2686, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1249.

LE ROUX J., ROZENKNOP A. & LACROIX M. (2019). Representation learning and dynamic programming for arc-hybrid parsing. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, p. 238–248, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/K19-1023.

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

QIN Z., HAN X., SUN W., LI D., KONG L., BARNES N. & ZHONG Y. (2022). The devil in linear transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 7025–7041, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & VILLEMONTE DE LA CLERGERIE E. (2013). Overview of the SPMRL 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

STERN M., ANDREAS J. & KLEIN D. (2017). A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume*

1 : Long Papers), p. 818–827, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1076.

YANG K. & DENG J. (2020). Strongly incremental constituency parsing with graph neural networks. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

ZARATIANA U., TOMEH N., HOLAT P. & CHARNOIS T. (2022). 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.

ZHANG Y., ZHOU H. & LI Z. (2020). Fast and accurate neural crf constituency parsing. In C. BESSIERE, Éd., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, p. 4046–4053 : International Joint Conferences on Artificial Intelligence Organization. Main track, DOI : 10.24963/ijcai.2020/560.

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

Les textes cliniques français générés sont-ils dangereusement similaires à leur source ? Analyse par plongements de phrases

Nicolas Hiebel[1] Olivier Ferret[2] Karën Fort[3] Aurélie Névéol[1]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) Sorbonne Université / Université de Lorraine, CNRS, Inria, LORIA, France
[1]prenom.nom@lisn.upsaclay.fr, [2]olivier.ferret@cea.fr,
[3]karen.fort@loria.fr

RÉSUMÉ

Les ressources textuelles disponibles dans le domaine biomédical sont rares pour des raisons de confidentialité. Des données existent mais ne sont pas partageables, c'est pourquoi il est intéressant de s'inspirer de ces données pour en générer de nouvelles sans contrainte de partage. Une difficulté majeure de la génération de données médicales est que les données générées doivent ressembler aux données originales sans compromettre leur confidentialité. L'évaluation de cette tâche est donc difficile. Dans cette étude, nous étendons l'évaluation de corpus cliniques générés en français en y ajoutant une dimension sémantique à l'aide de plongements de phrases. Nous recherchons des phrases proches à l'aide de similarité cosinus entre plongements, et analysons les scores de similarité. Nous observons que les phrases synthétiques sont thématiquement proches du corpus original, mais suffisamment éloignées pour ne pas être de simples reformulations qui compromettraient la confidentialité.

ABSTRACT

Are Synthesized Clinical Texts in French Dangerously Similar to Their Source ? An Analysis Using Sentence Embeddings.

Textual resources available in the biomedical field are scarce due to confidentiality issues. This problem can be addressed by automatically generating shareable data from existing restricted data. However, in the medical field, generated data should not contain sensitive information from restricted training data while remaining as close as possible to it. This paradox makes the evaluation of synthetic text challenging. In this study, we extend the evaluation of generated clinical corpora in French with semantic representations of sentences using sentence embeddings. We use cosine similarity between sentence embeddings to find similar sentences and analyze the similarity scores. We observe that the generated sentences are thematically close to those of the original corpus while being distant enough to avoid compromising confidentiality.

MOTS-CLÉS : Génération, Évaluation, Similarité, Texte clinique, Texte synthétique, Français.

KEYWORDS: Generation, Evaluation, Similarity, Clinical Text, Synthetic Text, French.

1 Introduction

Le manque de ressources est l'un des problèmes les plus communément rencontrés dans le Traitement Automatique des Langues (TAL), que ce soit en termes de domaine, de tâche, ou les deux. C'est le

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 46–54.

cas dans le domaine biomédical, où les ressources disponibles dans des langues autres que l'anglais sont rares (Névéol *et al.*, 2018). Ainsi, les données stockées dans les hôpitaux ne sont accessibles que par un nombre très restreint de personnes. Les données ne pouvant pas être diffusées, le partage des connaissances au sein de la communauté scientifique est difficile. Les possibilités de reproduction d'expériences et de comparaisons méthodologiques sont limitées.

Une piste pour résoudre ce problème est de générer de nouvelles données similaires aux données privées tout en préservant la confidentialité. La mise à disposition des données générées pourraient alors devenir un terrain de test, de comparaison, de discussion et d'entraide dans la recherche en TAL biomédical. Les modèles de type *transformer* (Vaswani *et al.*, 2017) ont montré leur efficacité dans différentes tâches de génération de textes (traduction, résumé, etc.). Les modèles auto-régressifs pré-entraînés comme GPT-2 (Radford *et al.*, 2019), GPT-3 (Brown *et al.*, 2020) et plus récemment InstructGPT (Ouyang *et al.*, 2022) et son successeur ChatGPT ont la capacité de générer des textes bien écrits. Les connaissances acquises lors du pré-entraînement de ces modèles pourraient être utiles dans des domaines peu dotés où il n'est pas possible d'entraîner un modèle génératif en partant de zéro.

L'évaluation automatique de la génération repose essentiellement sur des mesures de similarité avec une référence (Frisoni *et al.*, 2022) que l'on considère comme la réponse idéale attendue dans un contexte donné. La référence est comparée avec la ou les hypothèses du système. Les plus connues sont les mesures BLEU (Papineni *et al.*, 2002) et ROUGE (Lin, 2004) qui se basent sur des recouvrements de n-grammes. Cependant, dans le cas d'une génération ouverte, il n'y a pas de référence.

Nous proposons ici des alternatives à ces mesures pour évaluer une génération ouverte dans le domaine médical. Nous utilisons la méthode de génération mise en oeuvre par Hiebel *et al.* (2023). Dans ce travail, la génération est faite à l'aide d'un ajustement (*fine-tuning*) de modèles auto-régressifs pré-entraînés sur un corpus clinique dans plusieurs configurations.

Nous utilisons ici les corpus générés dans ce travail et nous proposons une évaluation sémantique des corpus en utilisant des similarités phrastiques obtenues à l'aide du modèle de plongements de phrases SENTENCE-BERT (SBERT) (Reimers & Gurevych, 2019) ajusté selon plusieurs configurations. Cette évaluation automatique supplémentaire permet d'estimer la proximité sémantique entre le contenu des corpus générés et le corpus réel. Nos contributions pour l'évaluation de la génération dans le domaine clinique sont les suivantes :

— nous présentons deux utilisations des plongements contextuels de phrases pour déterminer la distance des phrases des corpus générés avec les phrases du corpus original ;
— nous confirmons la qualité de la génération d'un point de vue sémantique ;
— nous identifions une piste d'amélioration dans la génération pour faciliter l'évaluation.

2 Méthode

2.1 Corpus utilisés

E3C (Magnini *et al.*, 2020) est un corpus multilingue librement disponible composé de documents médicaux provenant de différentes sources. Comme le travail de Hiebel *et al.* (2023), nous sélectionnons ici les cas cliniques en français du corpus.

CAS (Grabar *et al.*, 2018) est un corpus médical français contenant des cas cliniques français dé-identifiés dont la publication a été consentie par les patients concernés.

DEFT STS (Cardon & Grabar, 2020) est un corpus français de paires de phrases annotées en scores de similarité. Les phrases proviennent du corpus CLEAR (Grabar & Cardon, 2018), un corpus de phrases parallèles ayant pour but d'associer des phrases complexes avec leur version simplifiée. Sous ensemble de ce corpus, le corpus DEFT STS contient 1 010 paires de phrases qui ont été annotées en se reposant sur l'intuition des annotateurs. Certaines de ces paires portent sur le domaine biomédical.

CLISTER (Hiebel *et al.*, 2022) est un corpus clinique français de 1 000 paires de phrases annotées en scores de similarité. Ce corpus contient des phrases du corpus CAS. Les annotations ont été faites spécialement pour s'adapter au domaine clinique, avec une notion de compatibilité clinique entre les phrases, qui consiste à observer si les phrases peuvent ou non correspondre au même patient.

Est Républicain (ATILF & CLLE, 2020) est un corpus journalistique composé d'articles parus dans le quotidien éponyme. Nous avons ici sélectionné une sous-partie du corpus de manière à obtenir un corpus hors domaine de même taille que le corpus clinique E3C.

2.2 Génération des textes cliniques synthétiques

La génération des textes cliniques synthétiques est faite en ajustant des modèles auto-régressifs sur le corpus E3C.

Deux modèles différents ont été testés, le modèle multilingue BLOOM (Scao *et al.*, 2022) et un modèle français que nous appelons ici LLF (Simoulin & Crabbé, 2021). Pour une comparaison équitable, nous avons choisi deux modèles d'environ un milliard de paramètres.

Chaque modèle a été entraîné avec deux configurations différentes, dont l'une où des annotations en entités cliniques sont ajoutées au texte sous la forme de balises XML pour que le modèle génère directement des annotations. Les documents ont été générés avec comme unique contrainte une amorce (*prompt*) sous forme d'un token marquant le début d'un document. La génération a été faite en utilisant des paramètres de décodage favorisant la diversité des documents générés avec une température de 1,5 pour la génération annotée, 1,2 pour la génération non annotée, ainsi qu'une pénalité de répétition de 10.

La table 1 présente les statistiques des corpus réels et générés que nous souhaitons comparer, ne comprenant donc pas les corpus de similarité phrastique. Les corpus générés en incluant des annotations sont notés avec le suffixe « $_{+T}$ ».

	Toks	Docs	Toks/doc	Phrases/doc	Long. phrases	Self-Bleu	Perplexité
Est Républicain	306 866	8 226	37,3	2,0	18,4	0,52	77,8
CAS	231 662	717	323,1	15,8	20,4	0,66	17,0
E3C	328 645	1 009	325,7	15,2	21,4	0,68	22,0
Bloom$_{E3C}$	329 328	943	349,2	1,8	194	0,70	9,97
Bloom$_{E3C+T}$	346 413	1 997	173,5	3,1	56,0	0,73	9,27
LLF$_{E3C}$	328 498	1 028	319,6	6,9	46,3	0,68	10,03
LLF$_{E3C+T}$	336 154	978	343,7	7,2	47,7	0,67	11,9

TABLE 1 – Statistiques des corpus réels et des corpus générés, hors corpus de similarité phrastique.

2.3 Utilisation de plongements de phrases

Modèles de plongement de phrases Les plongements de phrases ont été obtenus en utilisant un modèle pré-entraîné multilingue de l'outil SBERT [1]. Afin d'observer différentes formes de similarité, nous avons calculé les plongements de phrases de tous les corpus selon trois configurations : une version avec le modèle pré-entraîné SBERT sans ajustement, une version avec le modèle ajusté sur le corpus DEFT STS et une version du modèle ajusté sur CLISTER.

Calcul des scores de similarité Pour obtenir une représentation de la proximité des phrases des différents corpus générés avec le corpus original E3C, nous recherchons pour chaque phrase du corpus généré les 100 phrases les plus proches dans le corpus E3C ainsi que les scores de similarité avec une similarité cosinus. Nous utilisons pour cela la bibliothèque FAISS (Johnson *et al.*, 2021). Nous obtenons pour chaque phrase générée 100 scores de similarité. Nous répétons le processus avec le corpus de cas cliniques CAS et le corpus hors domaine de l'Est Républicain.

Scores BLEU Nous souhaitons observer la proximité des phrases générées avec le corpus original à l'aide de la mesure BLEU. Cependant, contrairement au calcul des plongements de phrases en amont qui permet de de rechercher rapidement les éléments similaires dans la matrice de plongements, la mesure BLEU nécessite de calculer les similarités des phrases directement deux à deux, ce qui pose un problème de complexité dans notre cas où il faut comparer toutes les phrases du corpus source à toutes les phrases du corpus cible. C'est pourquoi nous observons ici uniquement les scores BLEU des phrases similaires déjà trouvées à l'aide d'un modèle de plongements de phrases.

Recherche des phrases du corpus E3C Comme deuxième approche pour observer la distance entre les corpus et E3C, nous essayons à partir des phrases d'E3C de retrouver dans la combinaison des phrases d'E3C et du corpus comparé les autres phrases du corpus E3C. Ainsi, plus il est simple de retrouver les autres phrases d'E3C, plus la distance entre les corpus est grande. Nous utilisons pour cela des mesures de recherche d'information.

3 Résultats

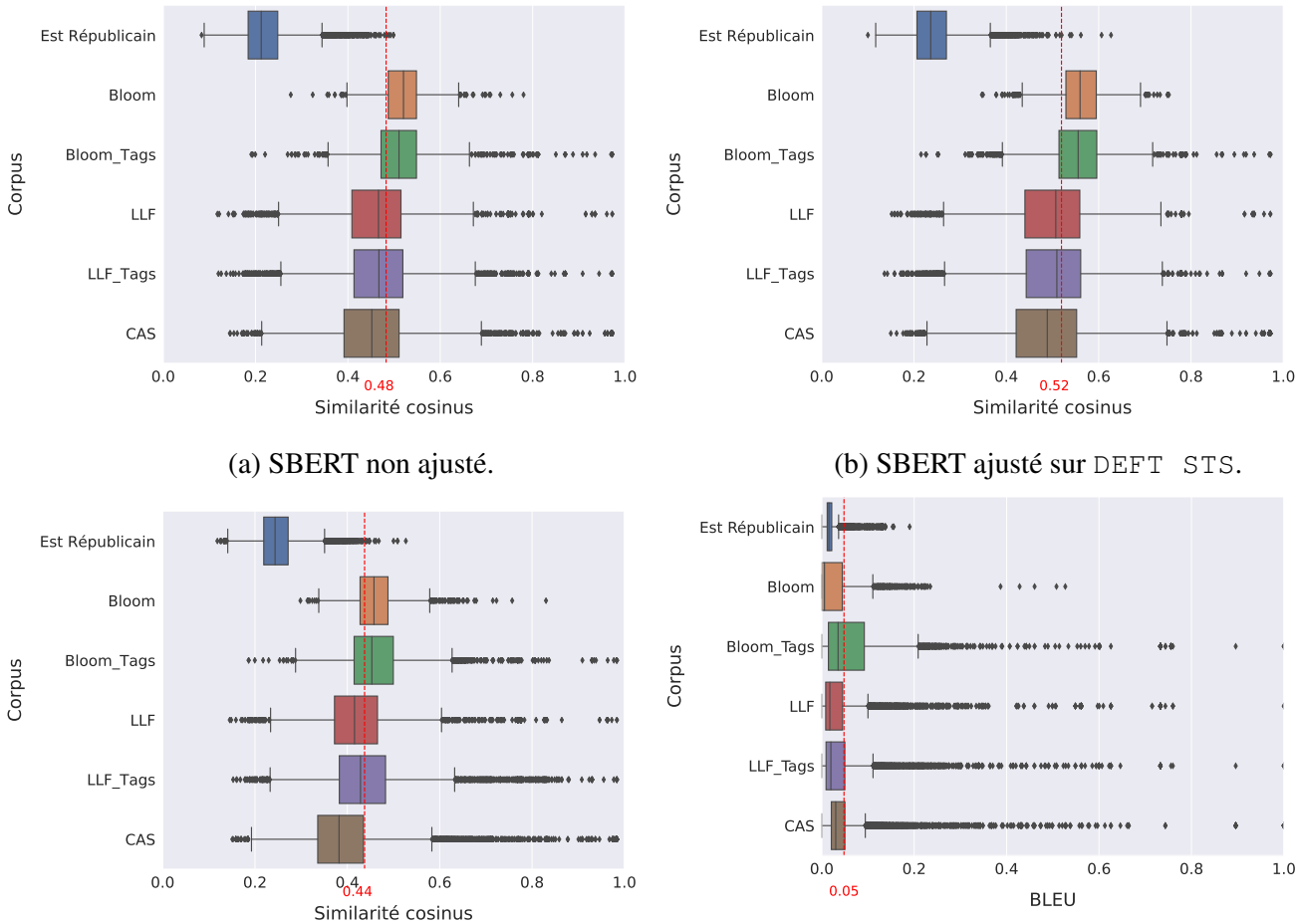
3.1 Distribution des similarités sémantiques des phrases

La figure 1 présente sous forme de diagrammes en boîtes les distributions des scores de similarité des phrases des corpus comparées aux phrases du corpus E3C. On y remarque facilement la mise à l'écart du corpus hors domaine de l'Est Républicain (en haut), surtout sur les figures 1a, 1b et 1c.

En ce qui concerne les modèles pour les corpus cliniques, les similarités les plus hautes sont obtenues avec le modèles SBERT ajusté sur le corpus DEFT STS. Cela peut s'expliquer à la fois par le fait qu'une portion des paires de phrases de ce corpus provient du domaine médical, et par la définition de similarité dans ce corpus qui se base sur des principes de simplification. Nous cherchons à repérer une thématique commune entre les phrases, et cela explique une similarité plus forte entre les corpus que celle obtenue avec le modèle SBERT non ajusté. Par ailleurs, les similarités obtenues avec le modèle SBERT ajusté sur CLISTER sont les plus faibles. Le modèle a été spécialisé sur le domaine clinique avec un critère de compatibilité clinique. Les similarités plus faibles des corpus cliniques avec ce modèle montrent que les phrases ont une thématique commune mais ne parlent pas forcément des

1. *distiluse-base-multilingual-cased-v1*

mêmes patients, ce qui est encourageant pour la qualité de la génération. Avec ce modèle, le corpus réel CAS se démarque un peu plus des corpus générés qu'avec les autres modèles, ce qui souligne une ressemblance entre E3C et CAS davantage surfacique que sémantique.



(a) SBERT non ajusté.

(b) SBERT ajusté sur DEFT STS.

(c) SBERT ajusté sur CLISTER.

(d) Scores BLEU.

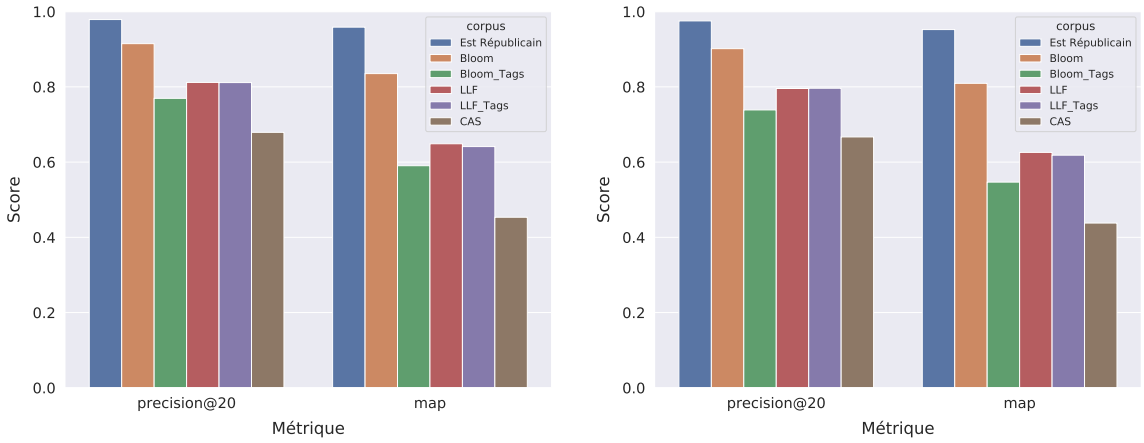
FIGURE 1 – Moyennes des scores de chaque phrase du corpus E3C avec les phrases les plus proches dans les corpus générés, CAS et un corpus hors domaine (Est Républicain). Pour les figures 1a, 1b et 1c, les similarités des 100 phrases les plus proches sont obtenues par similarité cosinus des plongements de phrases, issus de modèles SBERT. Pour la figure 1d, nous récupérons les 20 scores BLEU les plus élevés parmi les 100 phrases les plus proches trouvées avec les plongements de SBERT ajustés sur CLISTER. La ligne pointillée rouge correspond à la moyenne des scores des corpus cliniques.

La figure permet également d'observer la proximité des corpus générés ayant pour base le même modèle pré-entraîné. On constate que le corpus généré $\texttt{Bloom}_{E3C}$ se démarque particulièrement des autres corpus générés, avec une distribution plus resserrée et des similarités hors distributions moins extrêmes que dans les autres corpus. Cela s'explique sûrement par les très longues phrases du corpus $\texttt{Bloom}_{E3C}$. De par leur longueur, celles-ci contiennent mécaniquement plus d'informations, et chaque élément va influer sur l'encodage de la phrase dans un espace vectoriel réduit. Cela va donc

gommer les très fortes similarités de certaines sections de la phrase avec les phrases plus courtes de E3C, et inversement, il sera plus facile de trouver des phrases de E3C avec des points de similarité. Enfin, on observe sur la figure 1d que les scores BLEU des phrases sont en proportion beaucoup plus faibles que les scores obtenus avec les modèles de plongements de phrases, malgré l'étude des 20 phrases les plus similaires au lieu de 100. Le corpus $\texttt{Bloom}_{E3C+T}$ reste le plus similaire à E3C avec cette mesure, mais les autres corpus cliniques sont difficiles à différencier, illustrant le fait que cette mesure n'est pas la plus adaptée dans ce contexte.

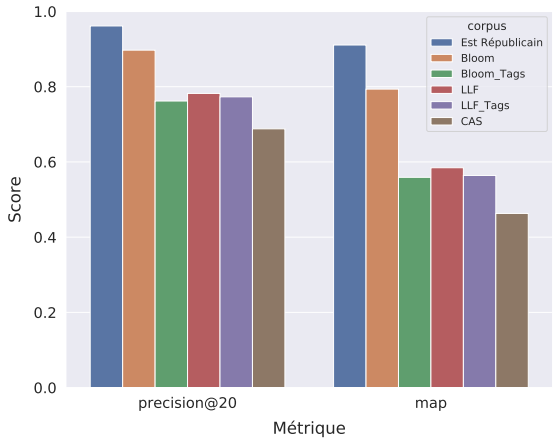
3.2 Recherche des phrases d'E3C

On observe sur la figure 2 les résultats de la recherche des phrases du corpus E3C dans les différentes combinaisons des phrases d'E3C et des corpus comparés.



(a) SBERT non ajusté.

(b) SBERT ajusté sur DEFT STS.



(c) SBERT ajusté sur CLISTER.

FIGURE 2 – Scores MAP et Précision@20 de la recherche des phrases de E3C parmi la combinaison des phrases de E3C et des phrases du corpus comparé. Les phrases sont représentées par des plongements de mots issus de Sentence-BERT. Nous recherchons parmi les 100 phrases les plus similaires.

On distingue logiquement le corpus de l'Est Républicain, dans lequel les phrases de E3C sont presque systématiquement retrouvées avec les trois modèles de plongements de phrases. Pour les corpus cliniques, on observe que Bloom_{E3C} présente des scores nettement plus élevés, CAS présente les scores les plus faibles et les scores des corpus générés Bloom_{E3C+T}, LLF_{E3C} et LLF_{E3C+T} se situent entre les deux.

Avec les trois modèles, les phrases de E3C sont retrouvées beaucoup plus facilement lorsqu'elles sont mélangées avec celles du corpus Bloom_{E3C}, au point que l'on se rapproche presque des scores obtenus avec le corpus hors domaine, particulièrement dans les premières positions (précision@20). Cela peut probablement s'expliquer par la différence de longueur des phrases, qui signifie aussi un nombre de phrases beaucoup plus faible. Proportionnellement, même en sélectionnant des phrases au hasard, il y a beaucoup plus de chance de tomber sur une autre phrase d'E3C qu'une phrase de Bloom_{E3C}. Concernant les modèles, on constate que les écarts entre les corpus cliniques sont moins importants lorsqu'on utilise les plongements obtenus par SBERT ajusté sur CLISTER.

Il est intéressant de constater qu'il est plus difficile de retrouver les autres phrases d'E3C en les mélangeant aux phrases du corpus CAS avec les trois modèles. Malgré les scores de similarité plus faibles pour CAS dans la figure 1, on constate ici que ce sont les phrases du corpus CAS qui sont les plus proches de celles d'E3C. Une première explication pourrait être les longueurs des phrases, très proches entre les corpus cliniques réels et beaucoup plus longues dans les corpus générés, spécialement pour Bloom_{E3C}, avec lesquelles les phrases d'E3C sont le plus facilement différenciées, se rapprochant presque du corpus hors domaine.

4 Conclusion

Nous avons présenté une évaluation automatique de corpus générés dans le domaine clinique à l'aide de mesures de similarité entre plongements de phrases issus de modèles implémentant plusieurs définitions de similarité sémantique. Les analyses montrent que les phrases synthétiques sont thématiquement proches des phrases du corpus dont elles sont inspirées, sans pour autant concerner les mêmes patients. Ces résultats suggèrent que les corpus générés sont proches du corpus original tout en apportant suffisamment d'innovation pour ne pas facilement pouvoir retrouver des patients. Cependant, une grande différence de longueur de phrases entre les corpus générés et le corpus original peut fausser ces comparaisons. Une perspective intéressante serait donc de contraindre la taille des phrases. Cela permettrait d'augmenter la proximité avec les corpus réels et d'améliorer l'évaluation de la génération.

Remerciements

Ce travail a été réalisé dans le cadre d'un projet de l'Agence Nationale de la Recherche, CODEINE (artificial text COrpus DEsIgNed Ethically), ANR-20-CE23-0026-01. Nous remercions par ailleurs Natalia Grabar (Université de Lille, CNRS, STL) qui nous a permis d'utiliser les corpus CAS et DEFT STS pour cette étude.

Références

ATILF & CLLE (2020). Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CARDON R. & GRABAR N. (2020). A French corpus for semantic similarity. In *Proceedings of the 12th Language Resources and Evaluation Conference*, p. 6889–6894, Marseille, France : European Language Resources Association.

FRISONI G., CARBONARO A., MORO G., ZAMMARCHI A. & AVAGNANO M. (2022). NLG-metricverse : An end-to-end library for evaluating natural language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, p. 3465–3479, Gyeongju, Republic of Korea : International Committee on Computational Linguistics.

GRABAR N. & CARDON R. (2018). CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, p. 3–9, Tilburg, the Netherlands : Association for Computational Linguistics. DOI : 10.18653/v1/W18-7002.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). CAS : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, p. 122–128, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/W18-5614.

HIEBEL N., FERRET O., FORT K. & NÉVÉOL A. (2023). Can synthetic text help clinical named entity recognition ? a study of electronic health records in French. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 2320–2338, Dubrovnik, Croatia : Association for Computational Linguistics.

HIEBEL N., FORT K., NÉVÉOL A. & FERRET O. (2022). CLISTER : Un corpus pour la similarité sémantique textuelle dans des cas cliniques en français (CLISTER : A corpus for semantic textual similarity in French clinical narratives). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 287–296, Avignon, France : ATALA.

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

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.

MAGNINI B., ALTUNA B., LAVELLI A., SPERANZA M. & ZANOLI R. (2020). The E3C Project : Collection and Annotation of a Multilingual Corpus of Clinical Cases. In J. MONTI, F. DELL'ORLETTA & F. TAMBURINI, Éds., *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 de *CEUR Workshop Proceedings* : CEUR-WS.org.

NÉVÉOL A., DALIANIS H., VELUPILLAI S., SAVOVA G. & ZWEIGENBAUM P. (2018). Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, **9**(1), 12. DOI : 10.1186/s13326-018-0179-8.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). Training language models to follow instructions with human feedback. DOI : 10.48550/ARXIV.2203.02155.

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.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). *Language Models are Unsupervised Multitask Learners*. Rapport interne, OpenAI.

REIMERS N. & GUREVYCH I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3980–3990 : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1410.

SCAO T. L. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. DOI : 10.48550/ARXIV.2211.05100.

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.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & 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.

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

Analyse sémantique AMR pour le français par transfert translingue

Jeongwoo Kang[1,2] Maximin Coavoux[1] Cédric Lopez[2] Didier Schwab[1]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
(2) Emvista, Immeuble Le 610 Bâtiment D 10, rue Louis Breguet, 34830 Jacou, France
[1]{prénom}.{nom}@univ-grenoble-alpes.fr
[2]{prénom}.{nom}@emvista.com

RÉSUMÉ

Abstract Meaning Representation (AMR) est un formalisme permettant de représenter la sémantique d'une phrase sous la forme d'un graphe, dont les nœuds sont des concepts sémantiques et les arcs des relations typées. Dans ce travail, nous construisons un analyseur AMR pour le français en étendant une méthode translingue zéro-ressource proposée par Procopio *et al.* (2021). Nous comparons l'utilisation d'un transfert bilingue à un transfert multi-cibles pour l'analyse sémantique AMR translingue. Nous construisons également des données d'évaluation pour l'AMR français. Nous présentons enfin les premiers résultats d'analyse AMR automatique pour le français. Selon le jeu de test utilisé, notre parseur AMR entraîné de manière zéro-ressource, c'est-à-dire sans donnée d'entraînement, obtient des scores Smatch qui se situent entre 54,2 et 66,0.

ABSTRACT

French AMR parsing with cross-lingual transfer learning

Abstract Meaning Representation (AMR) is a formalism for representing the semantics of a sentence with a graph of concepts connected with typed edges. In this work, we build a French AMR parser by extending a *zero-shot* cross-lingual method proposed by Procopio *et al.* (2021). We investigate the benefits of using bilingual data over multi-lingual data in cross-lingual AMR parsing. We also create evaluation datasets for French AMR parsing. With two evaluation datasets, we obtained respectively 54.2 and 66.0 Smatch score for the French AMR. Our French AMR parser is trained in a zero-shot setting, i.e. without French AMR training data.

MOTS-CLÉS : Analyse sémantique automatique, transfert translingue, seq2seq, AMR.

KEYWORDS: Semantic parsing, cross-lingual transfert, seq2seq, AMR.

1 Introduction

Abstract Meaning Representation (Banarescu *et al.*, 2013, AMR) est un formalisme permettant de représenter la sémantique d'une phrase sous la forme d'un graphe, dont les nœuds sont des concepts et les arcs des relations typées (figure 1). AMR a été initialement conçu pour analyser des phrases en anglais. Pourtant, Damonte & Cohen (2018) ont montré qu'AMR peut être utilisé également pour l'analyse sémantique multilingue, notamment pour l'espagnol, l'italien, l'allemand et le chinois. De plus, Damonte & Cohen (2018) ont construit des données d'évaluation AMR pour ces quatre langues par traduction du jeu d'évaluation anglais, et les ont diffusées. Ces données d'évaluation ont mené au développement d'analyseurs AMR translingues (Blloshmi *et al.*, 2020; Uhrig *et al.*, 2021; Procopio *et al.*, 2021) exploitant

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 55–62.

les données d'entraînement anglaises (seules données d'entraînement existantes). Les méthodes translingues consistent à utiliser la projection d'annotations sur des phrases traduites automatiquement en langue cible (Damonte & Cohen, 2018; Blloshmi *et al.*, 2020), l'apprentissage par transfert multitâche (Procopio *et al.*, 2021; Xu *et al.*, 2021) ou encore la distillation (Cai *et al.*, 2021).

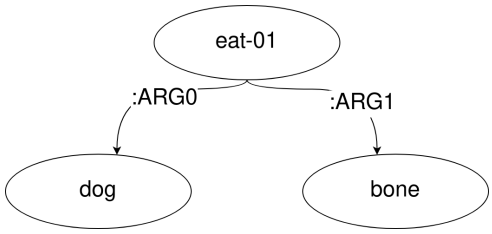


FIGURE 1 – Graphe AMR pour la phrase "The dog eats a bone."

Notre objectif est de construire un parseur AMR pour le français, langue pour laquelle il n'existe pas de données AMR ni de travaux antérieurs, à l'exception de ceux de Vanderwende *et al.* (2015) [1]. Nous reprenons l'approche de Procopio *et al.* (2021), basée sur un modèle séquence-à-séquence (seq2seq) multilingue pré-entraîné et affiné sur deux tâches : la prédiction de graphes sémantiques pour l'anglais, et la traduction automatique multilingue. Nous construisons par ailleurs notre propre jeu de test par projection d'annotations et traduction automatique suivie d'une vérification manuelle. En résumé nos contributions sont les suivantes :

— Nous reproduisons le travail de Procopio *et al.* (2021) et publions notre implémentation [2], qui est *de facto* la première implémentation disponible, Procopio *et al.* (2021) n'ayant pas publié leur code à ce jour.

— Nous construisons et publions le premier jeu d'évaluation AMR en français, obtenu par traduction automatique partiellement post-édité et projection d'annotations.

— Nous décrivons un système d'analyse AMR pour le français, par extension de Procopio *et al.* (2021) vers cette langue, et présentons les résultats de nos expériences d'apprentissage par transfert translingue. Nous obtenons des résultats intéressants (Smatch 54 et 66) sans donnée d'entraînement en français. À notre connaissance, notre travail est le premier à évaluer un modèle AMR sur le français.

2 Analyse AMR translingue

Speaking the Graph Languages (SGL) Nous utilisons l'approche SGL (Procopio *et al.*, 2021) fondée sur un modèle de séquence à séquence dont l'entrée est une phrase et la sortie est la forme linéarisée d'un graphe AMR (tableau 1). Procopio *et al.* (2021) adoptent un modèle de langue multilingue pré-entraîné : *mBART* (Liu *et al.*, 2020). Ils reformulent l'analyse AMR multilingue comme une tâche de traduction multilingue. Pendant l'entraînement, le modèle de traduction de *mBART* apprend à *traduire* des textes anglais en des graphes sémantiques AMR et UCCA (Abend & Rappoport, 2013, *Universal Conceptual Cognitive Annotation*). [3] En plus de ces deux formalismes de graphes sémantiques, le modèle apprend la

1. Vanderwende *et al.* (2015) utilisent un analyseur sémantique pour extraire la forme logique d'une phrase et appliquent un ensemble de règles de conversion vers AMR. Ils ne proposent pas d'évaluation empirique de leur méthode pour le français. À l'inverse, nous générons un graphe AMR directement à partir d'une phrase d'entrée et évaluons notre système.

2. Le code est accessible sur `https://github.com/Emvista/French-Amr-Parser.git` et également archivé sur `https://doi.org/10.5281/zenodo.7944999`.

3. UCCA est un formalisme sémantique inspiré par la Théorie Linguistique de Base (Dixon, 2009, 2010, *Basic Linguistic Theory*). Contrairement à l'AMR pour lequel il n'y a pas d'alignement entre les concepts et les tokens, UCCA donne des représentations ancrées, c'est-à-dire que chaque token d'une phrase est une feuille du graphe sémantique.

traduction multilingue avec quatre corpus parallèles : EN↔ES, EN↔DE, EN↔ IT, EN↔ZH (tableau 1). Notons que, contrairement aux données de traduction multilingue, pour l'analyse sémantique, la direction de la traduction est unidirectionnelle, depuis texte source anglais vers les graphes sémantiques. Puisque SGL fonctionne avec deux formalismes d'analyse sémantique, AMR et UCCA, leur approche n'est pas seulement multilingue mais aussi multi-formalismes.

entrée : <en>I have a dog.	**sortie :** <es>Tengo un perro.
entrée : <it>sono andato a letto.	**sortie :** <en>I went to bed.
entrée : <en>The boy wants to go.	**sortie :** <amr>(w / want-01 :ARG0 (b/boy) :ARG1 (g/go-02 :ARG0 b))
entrée : <en>John kicked his ball.	**sortie :** <ucca> [<root_0> H [<H_0> A [<A_0> T [John]] P [<P_0> T [kicked]] A [<A_0> E [<E_0> T [his]] C [<C_0> T [ball]]]]]

TABLEAU 1 – Exemples de données d'entraînement.

Adaptation de l'approche SGL au français Le choix de cette méthode a deux motivations. Tout d'abord, il s'agit d'une approche zéro-ressource qui ne nécessite pas de données annotées en AMR pour la langue cible et permet donc de pallier leur absence pour le français. À l'entraînement, le modèle apprend l'analyse en AMR, uniquement à partir de textes anglais. Au moment de l'inférence, il parvient à générer des graphes AMR à partir de textes dans des langues pour lesquelles le modèle n'a pas vu de données AMR correspondantes. Ceci est rendu possible par la capacité multilingue de *mBART* ainsi que par la tâche de traduction multidirectionnelle, pour laquelle le modèle est affiné. Deuxièmement, SGL est facile à étendre à une nouvelle langue. Pour permettre au modèle de générer des AMR pour des textes en français, il suffit d'ajouter un corpus parallèle FR ↔ EN aux données d'entraînement. Nous utilisons notre propre implémentation de SGL pour l'adapter au français.

Multilingue vs. Bilingue Des travaux (Zhang *et al.*, 2020; Conneau *et al.*, 2020) montrent l'avantage d'utiliser des données multilingues pour entraîner un modèle de traduction. Tang *et al.* (2021) montrent que le modèle de traduction multilingue est plus performant que le modèle de traduction bilingue quand il s'agit d'une traduction *many-to-one* où le modèle apprend à encoder en plusieurs langues et à décoder en une seule langue, notamment l'anglais. D'autre part, le transfert négatif (Wang *et al.*, 2019) est également un problème connu dans les modèles multilingues (Wang *et al.*, 2020; Conneau *et al.*, 2020) : les tâches multilingues entravent les performances individuelles pour chaque tâche. La méthode SGL est potentiellement sujette à ce problème, étant formalisée comme une tâche de traduction multilingue.

Nous souhaitons comparer l'usage d'un seul modèle multilingue à celui de multiples modèles bilingues (un pour chaque langue cible). Pour répondre à cette question, nous entraînons un parseur AMR français dans deux contextes : **Bilingue_FR** où le modèle est entraîné avec des données cibles uniquement (corpus AMR/UCCA en anglais et corpus parallèle EN↔ FR) et **Multilingue** où le modèle est entraîné avec des données cibles mais aussi avec des données multilingues (corpus AMR/UCCA en anglais et corpus parallèles EN↔ ES, DE, IT, ZH, FR). Concernant le modèle bilingue, nous entraînons également un modèle espagnol (**Bilingue_ES**) et un modèle allemand (**Bilingue_DE**) de la même manière avec un corpus parallèle pour chaque langue cible. Cela nous permettra de comparer plusieurs modèles bilingues de manière plus complète (section 4). Une comparaison similaire entre modèles bilingues et multilingues est décrite par Blloshmi *et al.* (2020). En revanche, notre perspective est différente : Blloshmi *et al.* (2020) utilisent des données cibles synthétiques, alors que nous sommes en contexte zéro-ressource.

3 Données pour l'évaluation de l'analyse AMR en français

Pour évaluer notre système, nous construisons des données d'évaluation pour le français par deux méthodes : la projection d'annotations et la traduction automatique.

Le petit prince Le corpus du *Petit Prince* d'Antoine de Saint-Exupéry est annoté pour l'analyse AMR en anglais et est disponible sur le site d'AMR [4]. Le livre étant initialement écrit en français, nous pouvons obtenir des paires [phrase française, graphe AMR] en alignant chaque phrase anglaise du corpus AMR sur son équivalent français. Notons que les textes français et anglais doivent être alignés au niveau de la phrase ou au-delà du niveau de la phrase puisqu'un graphe AMR correspond à une ou plusieurs phrases dans le corpus *Le Petit Prince*. À cette fin, nous utilisons un outil d'alignement de textes pré-entraîné, *LF Aligner* d'Andras Farkas [5] et nous post-éditons les erreurs d'alignement manuellement. À la suite de la projection des annotations, nous avons obtenu 1 562 graphes AMR pour le français. Nous appelons ce jeu de données FR_LPP_GOLD.

AMR-2.0 Nous obtenons également un deuxième jeu de données d'évaluation en traduisant le texte source anglais du jeu de test AMR-2.0 (LDC2017T10) [6] vers le français. Comme nous utilisons la traduction automatique DeepL [7] sans la post-éditer manuellement, nous appelons ce jeu de données FR_SILVER. Nous évaluons empiriquement la fiabilité des résultats sur ces données dans le paragraphe suivant.

Examen de fiabilité des données de test : FR_SILVER Nous avons examiné les 200 premiers [8] exemples de FR_SILVER et post-édité la traduction manuellement. Ainsi, 56 phrases ont été corrigées sur 200 lors de la post-édition (Correspondance Exacte : 0.72). Nous avons considéré ces 200 exemples post-édités comme la traduction de référence et évalué les traductions correspondantes de FR_SILVER par rapport à ces données avec le score BLEU (Papineni *et al.*, 2001). Le score BLEU de cet échantillon de traduction est égal à 0,89. Ensuite, nous avons examiné l'effet de l'utilisation de FR_SILVER sur l'évaluation d'analyses AMR. Pour cela, nous avons pris deux modèles différents, **Multilingue** et **Bilingue_FR** et avons évalué chaque modèle avec deux jeux de données : 200 exemples de FR_SILVER ainsi que ses correspondants post-édités. L'objectif de cette expérience est de comparer les deux scores Smatch et de considérer la différence comme un indicateur de la fiabilité de FR_SILVER en tant que jeu de test. Nous avons reproduit l'expérience en espagnol et en allemand en traduisant les données d'évaluation à l'aide de DeepL (ES_SILVER, DE_SILVER). Puis, nous avons évalué les modèles entraînés deux fois, sur les données d'évaluation originelles et les données automatiquement traduites (tableau 2). La différence entre les deux scores Smatch en allemand est égale à 4,5 % en moyenne pour les deux modèles et pour l'espagnol à 2,4 %. Pour le français, la différence est de ±1%, ce qui est dans la marge d'erreur. En raison de cet écart insignifiant pour le français, nous n'incluons pas ces résultats dans la discussion approfondie dans 4 et considérons FR_SILVER comme des données de test qui donnent une approximation relativement fiable pour évaluer nos modèles en français.

4. https://amr.isi.edu/download.html
5. Nous utilisons une version en ligne : http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit/.
6. Nous avons choisi la version 2.0 pour être cohérent avec les données d'évaluation AMR multilingues (Damonte & Cohen, 2018).
7. DeepL API sur https://www.deepl.com/
8. Les exemples redondants ou les phrases trop courtes telles que "SOB" et "EMR" ainsi que des interjections telles que "Haha", "Braawwk !" sont exclues.

	Bilingue_FR	Multilingue
DE	61,1	63,3
DE_SILVER	63,3	64,6
ES	66,4	66,8
ES_SILVER	67,7	64,9
FR_SILVER_POSTEDIT	58,2	55,6
FR_SILVER	58,2	55,0

TABLEAU 2 – Score Smatch avec les données de référence et les données silver.

4 Expériences

Pour l'expérimentation, nous suivons principalement les spécifications du modèle mBART_{ft} de Procopio *et al.* (2021). Nous renvoyons les lectrices et lecteurs à l'article original pour obtenir des informations générales et détaillées sur la reproduction des expériences. Sauf indication contraire, nous utilisons la même configuration que le modèle mBART_{ft}.

Données et pré-traitement Procopio *et al.* (2021) reproduisent l'environnement de la campagne d'évaluation CoNLL2019 (Koller *et al.*, 2019) et utilisent les données d'analyse sémantique UCCA qui sont distribuées aux participant·es de l'atelier. Ces données n'étant pas directement accessibles, nous avons téléchargé les données d'entraînement UCCA correspondantes directement sur le site officiel [9]. Ce corpus compte 8 775 paires phrase-graphe alors que pour mBART_{ft}, le corpus correspondant compte 6 572 paires. Au cours de la linéarisation des graphes UCCA, les auteurs utilisent l'étiquette T à la fois pour le nœud *Time* et le nœud *Terminal*. Cependant, pour distinguer ces deux concepts différents, nous utilisons T pour Time et Z pour Terminal. Nous employons l'API UCCA [10] pour la linéarisation et la délinéarisation des graphes UCCA. À la différence de mBART_{ft}, nous incluons le corpus parallèle français-anglais dans les données d'entraînement, plus précisément, *ParaCrawl* (Esplà-Gomis *et al.*, 2019). Nous appliquons le même filtrage de corpus parallèle que celui de mBART_{ft}, par exemple, le filtrage selon la longueur de la phrase ainsi que le ratio relatif du nombre de caractères entre la phrase d'entrée et la phrase cible.

Détails de l'entraînement et de l'évaluation Dans la version originale de mBART_{ft}, Procopio *et al.* (2021) utilisent 8 000 tokens comme taille effective de batch pendant l'entraînement. Pour nos expériences, nous utilisons des batches de 128 exemples. Pour l'échantillonnage des données, mBART_{ft} suréchantillonne les graphes sémantiques par rapport aux données de traduction multilingues. Nous appliquons la même probabilité que celle utilisée dans mBART_{ft}, soit 0,8 pour les graphes sémantiques et 0,2 uniformément répartie entre les corpus de traduction multilingues. Pour être plus précis, dans un cadre bilingue, nous attribuons 0,8 pour les graphes sémantiques et 0,2 pour le corpus parallèle EN↔FR. Dans le cadre multilingue, 0,8 pour l'analyse sémantique et 0,2 est uniformément répartie entre cinq corpus parallèles : EN↔ ES, EN↔ DE, EN↔ IT, EN↔ ZH, EN↔FR.

Résultats Nous présentons nos résultats dans le tableau 3 et incluons des résultats pour l'anglais, pour les 4 langues evaluées par Damonte & Cohen (2018), ainsi que pour le français. Nous incluons dans le

9. https://universalconceptualcognitiveannotation.github.io/
10. https://ucca.readthedocs.io/en/latest/api.html

tableau 3 les résultats d'autres modèles d'analyse sémantique multilingue : USeA (Orlando *et al.*, 2022), XL-AMR (Blloshmi *et al.*, 2020) et XLPT-AMR (Xu *et al.*, 2021). Nous utilisons Smatch (Cai & Knight, 2013) en tant que mesure d'évaluation pour l'analyse AMR. Un graphe AMR peut être exprimé comme un ensemble de triplets, dont chacun contient deux nœuds et un arc qui relie les deux. L'algorithme SMATCH calcule le chevauchement entre deux graphes en comptant les triplets en communs (F-score).
Le tableau 3 montre l'avantage d'utiliser des données bilingues pour une langue cible par rapport à l'utilisation de données multilingues dans les cas majoritaires. Pour les deux jeux de test en français, FR_LPP_GOLD et FR_SILVER, **Bilingue_FR** surpasse **Multilingue** de respectivement 1,8 et 1,9 score Smatch. Pour l'allemand, **Bilingue_DE** obtient un score Smatch supérieur de 1,5 par rapport à **Multilingue**. Pour l'espagnol, cependant, **Multilingue** surpasse **Bilingue_ES**. Une hypothèse pour expliquer ce résultat est la particularité du corpus parallèle utilisé pour entraîner ce modèle. MultiUN (Windsor *et al.*, 2010), le corpus parallèle utilisé pour l'espagnol, a un style d'écriture plus formel que Paracrawl qui est utilisé pour le français et l'allemand. La majorité des données AMR proviennent de forums ou de discussions en ligne dont le style d'écriture est relativement familier et moins formel. Par conséquent, **Multilingue** a probablement bénéficié des corpus français et allemand qui sont plus similaires aux données AMR, tandis que **Bilingue_ES** a probablement eu davantage de données moins représentatives.

	EN*	DE	ES	IT	ZH	FR_LPP_GOLD	FR_SILVER
$\text{Multilingue}_{\text{zero}}$	80,3	63,3	66,8	65,8	52,9	52,4	64,1
$\text{Bilingue\_FR}_{\text{zero}}$	80,0	61,1	66,4	65,9	48,4	**54,2**	**66,0**
$\text{Bilingue\_ES}_{\text{zero}}$	79,9	61,5	66,1	65,2	48,2	51,9	63,3
$\text{Bilingue\_DE}_{\text{zero}}$	80,2	64,8	65,0	64,2	48,9	50,0	61,8
$\text{USeA}_{\text{zero}}$ (Orlando *et al.*, 2022)	78,6	58,8	62,1	60,2	38,3	46,5	61,1
SGL_{zero} (Procopio *et al.*, 2021)	**81,2**	65,8	69,2	69,6	54,8	-	-
SGL (Procopio *et al.*, 2021)	-	69,8	**72,4**	**72,3**	**58,0**	-	-
$\text{XL-AMR}_{\text{zero}}$ (Blloshmi *et al.*, 2020)	-	32,7	39,1	37,1	25,9	-	-
XL-AMR (Blloshmi *et al.*, 2020)	-	53,0	58,0	58,1	41,5	-	-
XLPT-AMR (Xu *et al.*, 2021)	-	**70,4**	71,7	70,8	-	-	-

TABLEAU 3 – Score Smatch pour l'analyse AMR multilingue. Les scores sont tels que rapportés dans les articles originaux à l'exception des scores français pour USeA qui sont nouvellement ajoutés dans notre travail. Nous avons exploité l'API fournie d'USeA pour effectuer le test sur le français. ZERO indique les modèles zéro-ressource (aucun graphe AMR de la langue cible utilisé lors de l'apprentissage). L'évaluation sur l'anglais n'est pas zéro-ressource dans aucun cas et est ainsi marqué avec *.

Limites Alors que les données FR_SILVER donnent une bonne approximation pour évaluer le modèle français en l'absence de données annotées, elles peuvent conduire à un biais pour évaluer le modèle translingue. La traduction automatique a souvent tendance à être similaire à l'anglais en matière d'ordre de mots et de choix de termes. Par conséquent, utiliser FR_SILVER comme données d'évaluation peut biaiser en faveur du modèle translingue qui est entraîné principalement sur des données en anglais. Il convient également de noter que nous n'avons pas pu reproduire fidèlement les résultats rapportés par Procopio *et al.* (2021). En reproduisant le résultat d'évaluation de mBART_{ft}, c'est-à-dire sans les données du français, notre modèle a montré un écart allant de 1,1% pour l'analyse AMR anglaise à 5,3% pour l'analyse AMR italienne. En l'absence du code original, il est difficile d'enquêter sur l'origine exacte de cette lacune. L'une des expli-

cations probables réside dans notre propre implémentation ainsi que la configuration de l'expérimentation mentionnée dans la section 4, qui est légèrement différente de celle de Procopio *et al.* (2021). Notre objectif, cependant, est de comparer l'effet des données multilingues à celui des données bilingues dans l'analyse AMR translingue. Nous estimons que les différences dans les détails de mise en œuvre et les résultats de l'entraînement par rapport au travail original n'affectent pas de manière importante notre enquête comparative.

5 Conclusion

Dans cet article, nous présentons un parseur AMR français zéro-ressource entraîné par transfert translingue, ainsi qu'un premier jeu de données d'évaluation pour l'analyse AMR en français. Nos résultats montrent que l'utilisation de données bilingues est plus bénéfique que l'utilisation de données multilingues pour construire un parseur AMR français avec l'approche zéro-ressource. Dans la mesure où les données cibles sont soigneusement choisies, cette observation pourrait être généralisée aux analyses AMR dans d'autres langues.

Remerciements

Nous tenons à remercier les relecteurices anonymes ainsi que Adrien Pupier, Bastien Giordano, Kevin Cousot et Sylvain Verdy pour leurs suggestions constructives qui nous ont permis d'enrichir l'article. Nous tenons à remercier également le premier auteur de Procopio *et al.* (2021) pour des échanges informatifs qui nous ont permis de reproduire leur travail avec le plus de détails possible. Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2023-AD011012853R1 attribuée par GENCI.

Références

ABEND O. & RAPPOPORT A. (2013). Universal conceptual cognitive annotation (ucca). p. 228–238.

BANARESCU L., BONIAL C., CAI S., GEORGESCU M., GRIFFITT K., HERMJAKOB U., KNIGHT K., KOEHN P., PALMER M. & SCHNEIDER N. (2013). Abstract meaning representation for sembanking. p. 178–186.

BLLOSHMI R., TRIPODI R. & NAVIGLI R. (2020). Xl-amr : Enabling cross-lingual amr parsing with transfer learning techniques. p. 2487–2500 : Association for Computational Linguistics (ACL). DOI : 10.18653/V1/2020.EMNLP-MAIN.195.

CAI D., LI X., HO J. C.-S., BING L. & LAM W. (2021). Multilingual amr parsing with noisy knowledge distillation. p. 2778–2789 : Association for Computational Linguistics. DOI : 10.18653/V1/2021.FINDINGS-EMNLP.237.

CAI S. & KNIGHT K. (2013). Smatch : an evaluation metric for semantic feature structures.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2020). Unsupervised cross-lingual representation learning at scale. p. 8440–8451.

DAMONTE M. & COHEN S. B. (2018). Cross-lingual abstract meaning representation parsing. volume 1, p. 1146–1155 : Association for Computational Linguistics (ACL). DOI : 10.18653/V1/N18-1104.

DIXON R. (2009). *Basic linguistic theory volume 1 : Methodology*.

DIXON R. (2010). *Basic linguistic theory volume 2 : Grammatical topics*.

ESPLÀ-GOMIS M., FORCADA M. L., RAMÍREZ-SÁNCHEZ G. & HOANG H. (2019). Paracrawl : Web-scale parallel corpora for the languages of the eu. p. 118–119.

KOLLER A., OEPEN S. & SUN W. (2019). Graph-based meaning representations : Design and processing. p. 6–11 : Association for Computational Linguistics (ACL). DOI : 10.18653/V1/P19-4002.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

ORLANDO R., CONIA S., FARALLI S. & NAVIGLI R. (2022). Universal semantic annotator : the first unified api for wsd, srl and semantic parsing. p. 20–25.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2001). Bleu : a method for automatic evaluation of machine translation. p. 311 : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PROCOPIO L., TRIPODI R. & NAVIGLI R. (2021). Sgl : Speaking the graph languages of semantic parsing via multilingual translation. p. 325–337 : Association for Computational Linguistics (ACL). DOI : 10.18653/V1/2021.NAACL-MAIN.30.

TANG Y., TRAN C., LI X., CHEN P.-J., GOYAL N., CHAUDHARY V., GU J., FAN A., AI F. & AI A. (2021). Multilingual translation from denoising pre-training. p. 3450–3466. Multilingual finetuning » Bilingual finetuning.

UHRIG S., GARCIA Y., OPITZ J. & FRANK A. (2021). Translate, then parse ! a strong baseline for cross-lingual AMR parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, p. 58–64, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.iwpt-1.6.

VANDERWENDE L., MENEZES A. & QUIRK C. (2015). An amr parser for english, french, german, spanish and japanese and a new amr-annotated corpus. p. 26–30 : Association for Computational Linguistics (ACL). DOI : 10.3115/V1/N15-3006.

WANG Z., DAI Z., POCZOS B. & CARBONELL J. (2019). Characterizing and avoiding negative transfer. volume 2019-June, p. 11285–11294 : IEEE Computer Society. Introduction of the word "negative transfer / inference", DOI : 10.1109/CVPR.2019.01155.

WANG Z., LIPTON Z. C. & TSVETKOV Y. (2020). On negative interference in multilingual models : Findings and a meta-learning treatment. p. 4438–4450 : Association for Computational Linguistics (ACL). Negative inference, DOI : 10.18653/V1/2020.EMNLP-MAIN.359.

WINDSOR L. C., CUPIT J. G. & WINDSOR A. J. (2010). Automated content analysis across six languages. volume 14 : Public Library of Science. DOI : 10.1371/JOURNAL.PONE.0224425.

XU D., LI J., ZHU M., ZHANG M. & ZHOU G. (2021). Xlpt-amr : Cross-lingual pre-training via multi-task learning for zero-shot amr parsing and text generation. p. 896–907 : Association for Computational Linguistics (ACL). DOI : 10.18653/V1/2021.ACL-LONG.73.

ZHANG B., WILLIAMS P., TITOV I. & SENNRICH R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. p. 1628–1639. DOI : 10.18653/V1/2020.ACL-MAIN.148.

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

DWIE-FR : Un nouveau jeu de données en français annoté en entités nommées

Sylvain Verdy[1] Maxime Prieur[2, 3] Guillaume Gadek[3] Cédric Lopez[1]
(1) Emvista, 10 rue Louis Breguet, 34830 Jacou, France
(2) CNAM, 292 rue Saint-Martin, 75003 Paris, France
(3) Airbus Defence and Space, 1 Bd Jean Moulin, 78990 Elancourt, France
sylvain.verdy@emvista.com, maxime.prieur.auditeur@lecnam.net,
cedric.lopez@emvista.com, guillaume.gadek@airbus.com

RÉSUMÉ

Ces dernières années, les contributions majeures qui ont eu lieu en apprentissage automatique supervisé ont mis en evidence la nécessité de disposer de grands jeux de données annotés de haute qualité. Les recherches menées sur la tâche de reconnaissance d'entités nommées dans des textes en français font face à l'absence de jeux de données annotés "à grande échelle" et avec de nombreuses classes d'entités hiérarchisées. Dans cet article, nous proposons une approche pour obtenir un tel jeu de données qui s'appuie sur des étapes de traduction puis d'annotation des données textuelles en anglais vers une langue cible (ici au français). Nous évaluons la qualité de l'approche proposée et mesurons les performances de quelques modèles d'apprentissage automatique sur ces données.

ABSTRACT

DWIE-FR : A new French dataset annotated in named entities

In the recent years, major contributions have been made in the field of supervised machine learning, which increasingly empathize the need for large-scale high-quality annotated datasets. Research on the french named entity recognition task faces the lack of large-scale annotated datasets with many hierarchical entity classes. In this paper, we present an approach to obtain a dataset that relies on translation and annotation steps from English to a target language (French in our study). We evaluate the quality of this alignment and measure the performances obtained by machine learning models on an aligned dataset.

MOTS-CLÉS : TAL, reconnaissance d'entités nommées, jeu de données, traduction, alignement.

KEYWORDS: NLP, named entity recognition, dataset, translation, alignment.

1 Introduction

Les technologies d'apprentissage automatique ont connu une forte accélération et ont montré de nettes améliorations sur différentes tâches de compréhension des langues naturelles. Alors que l'effort est principalement porté sur les algorithmes d'apprentissage, les jeux de données demeurent rares pour le français. C'est notamment le cas pour la tâche de reconnaissance d'entités nommées (REN).

Les entités nommées, définies comme « toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus » (Ehrmann, 2008), et la reconnaissance de ces

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 63–72.

entités ont fait l'objet de nombreuses études et de campagnes d'évaluation ((Galliano *et al.*, 2009), CLEF (Ehrmann *et al.*, 2020), ETAPE (Galibert *et al.*, 2014), QUAERO (Rosset *et al.*, 2011)). La tâche consiste généralement à repérer et à classer les tokens automatiquement selon une typologie prédéfinie. Les typologies proposées reprennent généralement le triptyque « Personne », « Lieu », « Organisation » ou sont spécifiques à un domaine (par exemple la santé) et ne participent pas au développement de systèmes d'analyse de texte tous domaines confondus. Les expériences réalisées sur cette tâche sont donc limitées par l'absence de jeux de données en français qui soient à la fois de grande taille et annotés avec une vaste typologie.

Des scores excellents (car comparables à un annotateur humain) sont désormais obtenus sur l'anglais (Ye *et al.* 2022 obtient un micro F1 à 91.4 sur OntoNotes 5.0, un jeu contenant 18 classes différentes), soutenus par la disponibilité de jeux de données annotés, tels que FewNerd (Ding *et al.*, 2021), Conll-2003/2005 (Sang & De Meulder, 2003) ou encore DWIE (Zaporojets *et al.*, 2021). Sur le français, les F1-scores atteignent des résultats de l'ordre de 85.7 (Bannour *et al.*, 2022) pour un nombre de classes limité. Créer de nouveaux jeux de données se différenciant des précédents sur un certain nombre de critères devrait permettre d'améliorer les modèles.

Dans cet article, nous montrons d'abord à travers le recensement des jeux de données français (section 2) qu'un jeu de données volumineux et annoté avec de nombreuses classes manque au panorama des jeux de données. Nous proposons une traduction automatisée et maîtrisée du corpus anglophone DWIE (Zaporojets *et al.*, 2021), qui préserve les types d'entités (section 3) dont la qualité est maîtrisée et rend possible l'entraînement de modèles pour la reconnaissance d'entités nommées (section 4). Nous montrons la capacité d'apprentissage de modèles sur ce corpus et fournissons des métriques de qualité. Nous espérons que la mise à disposition de ce jeu de données permettra d'améliorer la qualité des systèmes francophones à court terme.

2 Travaux antérieurs

De nombreux jeux de données en français annotés en entités nommées ont été publiés. Certains sont commercialisés (par exemple ESTER [1]), d'autres sont inaccessibles (par exemple DAWT (Spasojevic *et al.*, 2017)) ou distribués à usage non commercial uniquement (par exemple WikiNeural (Tedeschi *et al.*, 2021a)). Enfin, certains sont annotés avec les URI DBpedia mais pas directement avec les types d'entités (par exemple (Hellmann *et al.*, 2013)) et sont plutôt destinés à une tâche de liage d'entités. Notons finalement que des approches permettent de générer des jeux de données annotés en entités nommées « à la volée » en fonction de certains critères (par exemple GeNER (Kim *et al.*, 2021)). Nous avons finalement recensé neuf jeux de données en français qui sont à la fois accessibles et libres d'utilisation (cf. Tab. 1). Huit des neuf jeux de données sont annotés avec un nombre de classes compris entre trois et quinze. Sur cet aspect, le jeu de données Wikipedia-ner (Lopez *et al.*, 2019) se distingue des autres puisqu'il contient 41 classes bien qu'il soit limité en taille (21855 tokens). Il apparaît ainsi qu'il n'existe aucun grand jeu de données en français annoté avec plusieurs dizaines de classes alors que de tels jeux de données existent pour l'anglais notamment, par exemple FewNerd (Ding *et al.*, 2021) ou encore DWIE (Zaporojets *et al.*, 2021). L'objectif de ce travail est de proposer un tel jeu de données (cf. section 3).

1. cf. http://catalog.elra.info/product_info.php?products_id=999

Nom	Tokens	Annotations	Classes	Références
MultiNERD	4 300 000	279 300	15	(Tedeschi & Navigli, 2022)
WikiNeural	3 240 000	231 000	4	(Tedeschi *et al.*, 2021b)
Le tour du monde en 80 jours	84 972	6 076	12	(Lopez *et al.*, 2019)
WiNER-Fr	322 931	24 144	7	(Dupont, 2019)
Wikipedia-ner	21 855	6 132	41	(Lopez *et al.*, 2019)
CAP Twitter	env. 60 000	6 562	13	(Lopez *et al.*, 2017)
Europeana-newpapers-ner	207 000	13 860	3	(Neudecker, 2016)
Quaero French Medical Corpus	72 183	16 233	10	(Névéol *et al.*, 2014)
French TreeBank	350 931	11 636	7	(Sagot *et al.*, 2012)
WikiNer-Fr	3 499 695	420 061	4	(Nothman *et al.*, 2008)

TABLE 1 – Jeux de données annotés en entités nommées pour le français

3 Création du jeu de données

3.1 Choix du jeu de données anglais

Le choix du jeu de données a été réalisé selon plusieurs indicateurs. Ces indicateurs ont été identifiés pour répondre à plusieurs de nos contraintes, à savoir le nombre de classes, le nombre de tokens et d'entités suffisantes pour réaliser un apprentissage des systèmes. Nous avons également souhaité que le jeu de données soit annoté manuellement avec un score d'évaluation inter-annotateurs reconnu par la communauté scientifique tel que le Kappa de Cohen (Cohen, 1960). Nous avons ainsi identifié deux jeux de données anglophones susceptibles de nous intéresser, Few-Nerd (Ding *et al.*, 2021) et DWIE (Zaporojets *et al.*, 2021). Ces deux jeux de données présentent une ontologie sur plusieurs niveaux avec un nombre important de classes, de tokens et d'entités. À la suite de cette étape d'identification, nous avons retenu celui qui présentait le meilleur score d'accord inter-annotateur (Kappa de Cohen à 0.87) : DWIE. DWIE est annoté avec 169 classes organisées dans une taxonomie à 4 niveaux.

3.2 Traduction du jeu de données en français

La première étape de notre approche pour la conversion d'un jeu de données consiste à traduire le texte source de DWIE. Nous avons effectué une comparaison de différents systèmes de traduction anglais-français en utilisant le jeu de données WMT14 EN-FR (Bojar *et al.*, 2014), en nous concentrant sur des modèles proches de l'état de l'art. Parmi ces modèles, DeepL [2] est l'un des plus performants mais ce dernier n'étant pas open-source, nous utilisons le modèle Ott *et al.* 2018 [3]. Ce modèle utilise une architecture Transformer dotée de 6 blocs encodeurs et décodeurs, et obtient un score BLEU (Papineni *et al.*, 2002) de 43.2 sur WMT14 EN-FR.

3.3 Annotation du jeu de données français par alignement

L'objectif de cette étape est d'annoter les entités du jeu de données français à partir du jeu de données anglais annoté DWIE. Nous avons expérimenté deux versions d'une approche d'alignement qui sera

2. https://www.deepl.com/translator
3. https://github.com/facebookresearch/fairseq/tree/main/examples/scaling_nmt

évaluée dans la section 4.

La première approche d'alignement est divisée en trois étapes. Tout d'abord, une identification des tokens par correspondance exacte des formes en anglais et en français permet d'annoter certaines entités avec un haut niveau de confiance. Pour les tokens restant à annoter, une mesure de distance sémantique (similarité cosinus) retournée par le modèle *Bert-base-multilingual-cased* (Devlin *et al.*, 2018) est utilisée pour déterminer le mot traduit le plus proche sémantiquement (« mot cible », dans la suite). Si la distance sémantique ou la distance lexicale entre le mot source et le mot cible est supérieure respectivement à 0,70 et 0,60 (seuils définis empiriquement) alors ce mot est annoté avec le type du mot source. Cette première approche a donné lieu à un jeu de données que l'on nommera dans la suite DWIE-FR-v1.

La seconde approche d'alignement est une extension à la première. Cette dernière est complétée par deux nouveaux modules. Le premier module consiste à établir une liste de plusieurs traductions candidates pour chaque entité de la phrase d'origine, puis à vérifier si l'une de ces traductions apparaît dans la phrase traduite. L'algorithme de recherche en faisceau utilisé pour la traduction par Vijayakumar *et al.* 2016 et réutilisé par Ott *et al.* 2019 génère justement des propositions de traduction. Ceci établit les correspondances avec les entités présentes dans la phrase traduite. Le module suivant met en œuvre des patrons d'annotation ; par exemple un token qui se situerait entre deux tokens de même type et qui est un article est annoté avec ledit type. Ces deux modules sont particulièrement utiles lorsque l'entité à traduire est composée de plusieurs tokens qui n'ont pas tous été annotés lors des étapes précédentes. Cette seconde approche résulte en un jeu de données que l'on nommera dans la suite DWIE-FR-v2.

Par ailleurs, nous utilisons l'aligneur *FastAlign* (Dyer *et al.*, 2013) afin de positionner les deux approches décrites vis à vis de cet aligneur très utilisé par la communauté scientifique. Le jeu de données obtenu avec cet aligneur est nommé "DWIE-FR FastAlign" dans la suite.

3.4 Protocole d'évaluation de DWIE-FR-v1 et DWIE-FR-v2

La qualité des jeux de données DWIE-FR-v1, DWIE-FR-v2 et DWIE-FR FastAlign obtenus par les approches présentées précédemment a été mesurée grâce à sept personnes qui ont manuellement validé ou invalidé les annotations sur des échantillons représentatifs. D'une part, ces évaluations mesurent l'impact de chaque méthode d'alignement sur le jeu de données produit ; l'objectif ici est d'évaluer à quel point les transferts de classes de la version anglaise vers la version traduite ont été correctement effectués. D'autre part, elles donnent une indication sur la qualité globale du jeu de données produit.

Pour chacun des trois jeux de données, un échantillon de 1000 tokens a été annoté par chacun des sept experts (chaque expert a annoté des échantillons différents) en respectant la typologie d'erreurs présentée dans le Tableau 3. Au total, 298 phrases ont été évaluées.

Analyse de l'évaluation Le Tableau 2 montre les résultats de cette évaluation en termes de précision, rappel et F1-score. Il apparaît que DWIE dans sa version originale obtient un F1-score de 99%, ce qui confirme la qualité de l'annotation réalisée par Zaporojets *et al.* 2021, dont le score inter-annotateurs annoncé par les auteurs est de 0.87 et conforte ce choix du jeu de données d'origine. Le rappel reste légèrement plus faible que la précision, ce qui est dû à certaines annotations manquantes notamment

pour la classe « rôle » (i.e. rôle des personnes, fonctions, métiers). Cette absence s'explique par la subjectivité d'un tel label.

L'excellente précision de DWIE est héritée par DWIE-FR V1 mais la diminution de 13% du rappel indique que cette première approche d'alignement n'est pas assez couvrante. Ce rappel est augmenté grâce à la seconde approche d'alignement puisque DWIE-FR V2 obtient un rappel de 93.5% tout en conservant une précision très haute (98.6%). Cette perte de 3.2% de rappel par rapport à DWIE s'explique notamment par la difficulté d'aligner des tokens lorsque la traduction a généré plus de tokens que le texte source et par un score de similarité sémantique qui n'atteint pas le seuil fixé. La méthode non-supervisé FastAlign obtient un meilleur rappel (95.4%). Cependant, sa précision reste trop basse (93.4%).

Finalement, nous retenons DWIE-FR V2 que nous nommerons dans la suite DWIE-FR, version rendue libre et accessible [4]. Nous considérons que DWIE-FR est d'une qualité équivalente (à 3.2% près) à son homonyme anglais. Le jeu de données est composé de 589 394 tokens dont 60 292 annotés en entités nommées avec 169 classes.

Jeu de données (version)	Précision	Rappel	F1-Score
DWIE	99.6	98.5	99.0
DWIE-FR V1	99.0	85.0	91.4
DWIE-FR V2	98.6	93.4	95.8
DWIE-FR FastAlign	93.4	95.4	94.38

TABLE 2 – Évaluation de l'annotation des jeux de données

Label d'erreur	Type d'erreur
Ce n'est pas une entité nommée (l'erreur vient du jeu anglais)	Faux positif
Ce n'est pas une entité nommée (l'erreur vient de l'alignement)	Faux positif
L'entité possède la mauvaise étiquette (l'erreur vient du jeu anglais)	Faux positif
L'entité possède la mauvaise étiquette (l'erreur vient de l'alignement)	Faux positif
Le token devrait être annoté (l'erreur vient de l'alignement)	Faux négatif
Le token devrait être annoté (l'erreur vient du jeu anglais)	Faux négatif

TABLE 3 – Typologie d'erreurs utilisée lors de l'annotation en vue de l'évaluation de l'alignement

4 Expérimentations

Deux expériences principales ont été réalisées à partir de DWIE-FR : une spécialisation de FlauBERT (Le *et al.*, 2019) et une spécialisation de CamemBERT (Martin *et al.*, 2019) (section 4.1), ainsi que l'évaluation des modèles obtenus sur trois autres jeux de données (section 4.2). Les modèles ont été entraînés sur 50 *epochs* à l'aide du *framework* Flair (Akbik *et al.*, 2019), avec un *learning rate* de $1e{-}5$, une taille de batch de 16 et la fonction de coût cross-entropique.

4. https://github.com/Emvista/DWIE-FR

4.1 Performance des modèles FR vs. EN

Dans un premier temps, nous avons spécialisé RoBERTa (Zhuang *et al.*, 2021) à partir de DWIE-EN afin de positionner les spécialisations de FlauBERT et CamemBERT à partir de DWIE-FR. Le tableau 4 indique les performances obtenues par les trois modèles pour chacun des quatre niveaux de la taxonomie.

Modèles	Niv. 1		Niv. 2		Niv. 3		Niv. 4	
	F1 Micro	F1 Macro	F1 Micro	F1 Macro	F1 Micro	F1 Macro	F1 Micro	F1 Macro
DWIE-EN								
RoBERTa	95.35	94.3	93.51	78.98	85.99	52.01	84.32	54.69
DWIE-FR								
CamemBERT-ner [5]	87.18	83.81	84.53	58.66	75.27	29.82	72.5	27.85
CamemBERT-base	87.06	83.68	84.05	59.14	73.48	24.52	71.12	24.95
FlauBERT-base-uncased	87.21	84.36	84.36	63.17	78.08	42.31	76.21	42.69

TABLE 4 – Résultats des modèles sur DWIE-EN et DWIE-FR

Le tableau 4 laisse apparaître que plus le niveau de la taxonomie est élevé et plus le score F1 Micro diminue. La F1 Micro diminue d'environ 10% à 15% alors que le F1 Macro diminue d'environ 40% à 45% entre le niveau 1 et le niveau 4. Ceci s'explique par le déséquilibre entre les classes de haut niveau. En effet, plus le nombre de classes augmente (ce qui va de pair avec la spécialisation des classes ; niveau 1 à 4), plus l'écart entre le F1 Micro et le F1 Macro se creuse. Cette observation est valable aussi bien pour la version anglaise que pour la version française, ce qui indique une difficulté à appréhender des taxonomies à plusieurs niveaux et justifie la publication de ce nouveau jeu de données afin d'encourager des recherches à ce sujet.

Par ailleurs, il apparait que Flaubert est plus robuste que Camembert lorsqu'ils sont confrontés aux niveaux les plus élevés de la taxonomie. La spécialisation de FlauBERT obtient globalement les meilleurs résultats et pourra servir de référence pour les prochaines évaluations.

4.2 Inférence sur d'autres jeux de données

La seconde expérience vise à étudier le comportement des modèles entrainés sur DWIE-FR sur trois autres jeux de données libres et accessibles : European Newspapers-FR, Jules Verne et Wikipedia NER (décrits en section 2). L'évaluation s'effectue uniquement sur les classes Personne, Lieu et Organisation qui sont les seules classes communes à ces jeux de données.

La reconnaissance d'entité nommée est évaluée de deux manières, la reconnaissance de l'entité dans son entièreté et une reconnaissance à l'échelle du mot. Nous évaluons les systèmes en calculant le F1-score Macro tel qu'indiqué dans l'équation 1, soit, la moyenne des F1-score obtenus pour chacune des trois classes.

$$F1_{Macro} = \frac{F1_{LOC} + F1_{PER} + F1_{ORG}}{3} \quad (1)$$

5. https://huggingface.co/Jean-Baptiste/camembert-ner

Modèles	European Newspapers		Wikipedia NER		Jules Verne	
	F1 (Tokens)	F1 (BIO)	F1 (Tokens)	F1 (BIO)	F1 (Tokens)	F1 (BIO)
DWIE-FR-v2						
CamemBERT-ner	**62.38**	**54.14**	**85.23**	**76.84**	**78.62**	**70.83**
CamemBERT-base	60.53	52.03	83.06	72.27	74.62	65.37
FlauBERT-base-uncased	51.40	44.56	81.03	71.13	66.38	55.19
DWIE-FR-FastAlign						
CamemBERT-ner	56.60	46.25	82.73	72.86	74.6	59.14
CamemBERT-base	55.72	44.95	82.64	69.88	70.83	58.02
Flaubert-base-uncased	57.35	46.05	78.82	68.89	72.81	60.95

TABLE 5 – Résultats des modèles sur European NewsPapers, Wikipedia-NER FR et Jules Verne

Les résultats obtenus et consignés dans le tableau 5 montrent que les modèles entraînés sur DWIE-FR sont assez robustes pour être utilisés en pré-entraînement sur des jeux de données différents. Nous avons remarqué que globalement FlauBERT obtient de moins bons résultats que CamemBERT sur les méthodes **DWIE-FR-v2** et **FastAlign**. L'hypothèse du sur-apprentissage semble être une piste à étudier pour expliquer ces résultats. Nous pouvons également voir que le pré-apprentissage de Camembert-ner sur Wikiner-fr dans une tâche de reconnaissance d'entité nommées, permet d'améliorer les performances sur ces corpus jusqu'à 5% du F1-score.

Camembert-ner a été entraîné sur des classes génériques avec le corpus Wikiner-fr. Ce pré-entraînement a pu permettre de capturer des informations sémantiques dans les représentations intermédiaires de Camembert. Nous pouvons émettre l'hypothèse que le pré-entraînement d'un modèle sur une tâche de reconnaissance d'enitités nommées aide à la généralisation dans d'autres corpus de REN. Cependant il faut considérer que chaque jeu de données possède une stratégie d'annotation différente impliquant un alignement d'annotation difficile.

5 Conclusion

Dans cet article, nous avons constaté un manque de jeux de données en français qui soient accessibles, libres, et annotés avec plusieurs dizaines de classes d'entités nommées. Une approche constituée d'une étape de traduction et d'une étape d'alignement d'étiquettes a été expérimentée. Les évaluations permettent de considérer que le jeu de données DWIE-FR obtenu en appliquant cette approche est d'aussi haute qualité que le jeu source en anglais DWIE à 3,2% près. Des premiers apprentissages ont eu lieu avec ce nouveaux jeu de données sur la base des modèles de langues français CammemBERT et FlauBERT.

Dans la suite, nous étudierons comment maintenir des scores élevés quel que soit le niveau de l'ontologie considéré. Nous étudierons également la différence remarquée entre FlauBERT et CamemBERT qui s'accentue lorsque l'on tend vers les niveaux les plus spécifiques de la taxonomie.

Références

AKBIK A., BERGMANN T., BLYTHE D., RASUL K., SCHWETER S. & VOLLGRAF R. (2019). FLAIR : An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, p. 54–59.

BANNOUR N., WAJSBÜRT P., RANCE B., TANNIER X. & NÉVÉOL A. (2022). Modèles préservant la confidentialité des données par mimétisme pour la reconnaissance d'entités nommées en français. *Actes de la journée d'étude sur la robustesse des systemes de TAL*, p.1̃2.

BOJAR O., BUCK C., FEDERMANN C., HADDOW B., KOEHN P., LEVELING J., MONZ C., PECINA P., POST M., SAINT-AMAND H., SORICUT R., SPECIA L. & TAMCHYNA A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 12–58, Baltimore, Maryland, USA : Association for Computational Linguistics. DOI : 10.3115/v1/W14-3302.

COHEN J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**(1), 37.

DEVLIN J., CHANG M., LEE K. & TOUTANOVA K. (2018). BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, **abs/1810.04805**.

DING N., XU G., CHEN Y., WANG X., HAN X., XIE P., ZHENG H.-T. & LIU Z. (2021). Few-nerd : A few-shot named entity recognition dataset. DOI : 10.48550/ARXIV.2105.07464.

DUPONT Y. (2019). Un corpus libre, évolutif et versionné en entités nommées du français. In *TALN 2019-Traitement Automatique des Langues Naturelles*.

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *North American Chapter of the Association for Computational Linguistics*.

EHRMANN M. (2008). *Les Entitées Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Paris Diderot University.

EHRMANN M., ROMANELLO M., BIRCHER S. & CLEMATIDE S. (2020). Introducing the clef 2020 hipe shared task : Named entity recognition and linking on historical newspapers. In *Advances in Information Retrieval : 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, p. 524–532 : Springer.

GALIBERT O., LEIXA J., ADDA G., CHOUKRI K. & GRAVIER G. (2014). The etape speech processing evaluation. In *LREC*, p. 3995–3999.

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*.

HELLMANN S., LEHMANN J., AUER S. & BRÜMMER M. (2013). Integrating nlp using linked data. In *The Semantic Web–ISWC 2013 : 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, p. 98–113 : Springer.

KIM H., YOO J., YOON S., LEE J. & KANG J. (2021). Simple questions generate named entity recognition datasets.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. *CoRR*, **abs/1912.05372**.

LOPEZ C., MEKAOUI M., AUBRY K., BORT J. & GARNIER P. (2019). Reconnaissance d'entités nommées itérative sur une structure en dépendances syntaxiques avec l'ontologie nerd. In *Extraction et Gestion des Connaissances : Actes de la conférence EGC*, volume 79, p. 81–92.

LOPEZ C., PARTALAS I., BALIKAS G., DERBAS N., MARTIN A., REUTENAUER C., SEGOND F. & AMINI M.-R. (2017). Cap 2017 challenge : Twitter named entity recognition. *arXiv preprint arXiv :1707.07568*.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

NEUDECKER C. (2016). An open corpus for named entity recognition in historic newspapers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4348–4352, Portorož, Slovenia : European Language Resources Association (ELRA).

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. *Proc of BioTextMining Work*, p. 24–30.

NOTHMAN J., CURRAN J. R. & MURPHY T. (2008). Transforming wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, p. 124–132.

OTT M., EDUNOV S., BAEVSKI A., FAN A., GROSS S., NG N., GRANGIER D. & AULI M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv :1904.01038*.

OTT M., EDUNOV S., GRANGIER D. & AULI M. (2018). Scaling neural machine translation. *CoRR*, **abs/1806.00187**.

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.

ROSSET S., GROUIN C. & ZWEIGENBAUM P. (2011). *Entités nommées structurées : guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

SANG E. F. & DE MEULDER F. (2003). Introduction to the conll-2003 shared task : Language-independent named entity recognition. *arXiv preprint cs/0306050*.

SPASOJEVIC N., BHARGAVA P. & HU G. (2017). Dawt : Densely annotated wikipedia texts across multiple languages. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, p. 1655–1662, Republic and Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee. DOI : 10.1145/3041021.3053367.

TEDESCHI S., MAIORCA V., CAMPOLUNGO N., CECCONI F. & NAVIGLI R. (2021a). Wikineural : Combined neural and knowledge-based silver data creation for multilingual ner. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 2521–2533.

TEDESCHI S., MAIORCA V., CAMPOLUNGO N., CECCONI F. & NAVIGLI R. (2021b). Wiki-NEuRal : Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics : EMNLP 2021*, p. 2521–2533, Punta Cana, Dominican Republic : Association for Computational Linguistics.

TEDESCHI S. & NAVIGLI R. (2022). Multinerd : A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics : NAACL 2022*, p. 801–812.

VIJAYAKUMAR A. K., COGSWELL M., SELVARAJU R. R., SUN Q., LEE S., CRANDALL D. & BATRA D. (2016). Diverse beam search : Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv :1610.02424*.

YE D., LIN Y., LI P. & SUN M. (2022). Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 4904–4917.

ZAPOROJETS K., DELEU J., DEVELDER C. & DEMEESTER T. (2021). Dwie : An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, **58**(4), 102563.

ZHUANG L., WAYNE L., YA S. & JUN Z. (2021). A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th chinese national conference on computational linguistics*, p. 1218–1227.

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

Evaluating the Generalization Property of Prefix-based Methods for Data-to-text Generation

Clarine Vongpaseut[1, 2] Alberto Lumbreras [1] Mike Gartrell[1] Patrick Gallinari[1, 3]
(1) Criteo AI Lab, Paris, France
(2) École des Ponts ParisTech, Champs-sur-Marne, France
(3) Sorbonne Universite, CNRS, ISIR, Paris, France
clarinevong@gmail.com, a.lumbreras@criteo.com, m.gartrell@criteo.com,
patrick.gallinari@sorbonne-universite.fr

RÉSUMÉ

Évaluation de la capacité de généralisation de méthodes *prefix-based* pour le *data-to-text*

Le *fine-tuning* est le paradigme courant pour adapter des modèles de language pré-entrainés à une tâche. Les méthodes de *fine-tuning* léger, comme le *prefix-tuning*, modifient uniquement un petit ensemble de paramètres, permettant de réduire les coûts d'entrainement. Ces méthodes atteignent des résultats comparables au *fine-tuning*. Toutefois, leurs performances se dégradent lorsqu'on s'éloigne des données d'entraînement. De plus, des travaux récents questionnent l'efficacité de ces méthodes selon la tâche d'application et la taille du modèle. Nous proposons dans ce papier d'évaluer la capacité de généralisation de méthodes *prefix-based* en fonction de la taille du modèle pré-entrainé, dans le cadre multi-domaine pour le *data-to-text* i.e. la conversion de données structurées en texte. Nous observons que leurs performances dépendent fortement de la taille du modèle.

ABSTRACT

Fine-tuning is the prevalent paradigm to adapt pre-trained language models to downstream tasks. Lightweight fine-tuning methods, such as prefix-tuning, only tune a small set of parameters which alleviates cost. Such methods were shown to achieve results similar to fine-tuning ; however, performance can decrease when the inputs get farther from the training domain. Moreover, latest works questioned the efficiency of recent lightweight fine-tuning techniques depending on the task and the size of the model. In this paper, we propose to evaluate the generalization property of prefix-based methods depending on the size of the pre-trained language model in the multi-domain setting on data-to-text generation. We found that their performance depends heavily on the size of the model.

MOTS-CLÉS : *Prefix-tuning*, Apprentissage multi-tâche, Capacité de généralisation, *Data-to-text*.

KEYWORDS: Prefix-tuning, Multi-task learning, Generalization property, Data-to-text.

1 Introduction

Fine-tuning is the prevalent approach to adapting pre-trained language models (PLMs) to various downstream tasks (Devlin *et al.*, 2019). However, fine-tuning is both computationally and memory expensive. Lightweight fine-tuning methods address this problem by freezing most of the PLMs parameters, e.g. fine-tuning the top layers, or by training only a smaller set of added parameters. A first method, adapter-tuning (Houlsby *et al.*, 2019), inserts task-specific layers between the layers of PLMs. Recently Li & Liang (2021) presented a second method, prefix-tuning, where a prompt

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 73–81.

is optimized as hidden-states, i.e. key-value for Transformer (Vaswani *et al.*, 2017). Their method outperformed adapter-tuning and also achieved comparable performance with fine-tuning on data-to-text benchmarks when using GPT-2 medium and large (Radford *et al.*, 2019). Lester *et al.* (2021) introduce a third lightweight fine-tuning method, prompt-tuning, in which continuous embeddings are optimized as soft prompts. Their experiments show that the performance gap between prompt-tuning and fine-tuning reduces with increase of model's size. In this paper, we focus on prefix-based methods.

Lightweight fine-tuning methods are also used for multi-task learning. HyperFormer (Mahabadi *et al.*, 2021) builds on adapter-tuning and uses hypernetworks to generate task and layer-specific adapter parameters, conditioned on task and layer embeddings. When published, HyperFormer achieved better performance on average on the GLUE benchmark (Wang *et al.*, 2018) compared to fine-tuning and adapter-tuning, when using T5-small and T5-base (Raffel *et al.*, 2020). He *et al.* (2022) propose HyperPrompt, a multi-task method based on prefix-tuning and HyperFormer's work, which outperforms HyperFormer. They showed that only tuning the added parameters of HyperFormer and HyperPrompt did not lead to similar performance when compared to tuning both added parameters and the PLM's parameters on SuperGLUE (Wang *et al.*, 2019), a more difficult NLU benchmark, when using T5-large. This observation questions the efficiency of lightweight fine-tuning since the performance of those methods seem to vary depending on the task difficulty. Clive *et al.* (2022) concatenate two prefixes : one for the task and one for the domain.

We consider here data-to-text generation tasks, consisting in generating fluent descriptions of data available in table or graph format and typically extracted from databases. We consider a challenging multi-domain setting where data is supposed to come from multiple sources, each with its own term and term-relations distributions. This could be considered as a specific multi-task problem where each domain corresponds to a task. The challenge here is to adapt PLMs in order to handle this multi-domain setting when current practice considers mono-domain settings. We then propose to evaluate the generalization property of prefix-based methods in the multi-domain setting, both in zero-shot and after few-shot fine-tuning on new target domains. With the question of efficiency depending on the model's size in mind, we compare results for T5-small and T5-base transformers.

2 Prefix-based methods

2.1 Prefix-tuning

In this section, we detail prefix-tuning (Li & Liang, 2021). We consider a PLM with frozen parameters ϕ. We use E, D and D_c to denote the three classes of attention present in each layer respectively, for the self-attention in the encoder and decoder and cross-attention in the decoder.

For each attention class, a set of prependable key-value pairs is learnt $P = \left\{(P_k^i, P_v^i)\right\}_{1 \leq i \leq N}$ with $P_{k,v}^i \in \mathbb{R}^{l \times h \times d_h}$, where N is the number of transformer blocks, l is the length of the prefix, h is the number of attention heads and d_h is the dimension of each head.

At the i-th transformer block, the additional key-value pairs are concatenated to the original key and value matrices, denoted K^i, V^i : $\quad K^{i\prime} = \left[P_k^i, K^i\right] \quad V^{i\prime} = \left[P_v^i, V^i\right]$. Multi-head attention is then computed using the new key-value and the original query Q^i.

The overall prefix, parameterized by θ, is denoted $P_\theta = \left\{P^E, P^D\ P^{D_c}\right\}$ and optimized through gradient descent : $\max_\theta \sum_{j=1}^{n} \log P(Y_j | X_j, P_\theta, \phi)$.

The prefix is not optimized directly. For each attention class, we learn $W \in \mathbb{R}^{l \times d}$ with d the model's

dimension and a two-layered feed-forward network with a bottleneck architecture, which takes as input W and outputs the key-value pairs for all transformer blocks $P = \left\{(P_k^i, P_v^i)\right\}_{1 \leq i \leq N}$. After training, the output of the MLPs can be saved and their weights can be dropped.

$$P_\theta = \left\{P^E,\, P^D\, P^{D_c}\right\} = \left\{MLP^E(W^E),\, MLP^D(W^D),\, MLP^{D_c}(W^{D_c})\right\}.$$

2.2 HyperPrompt

In this section, we present HyperPrompt (He *et al.*, 2022) for multi-domain learning. We consider a set of M domains $S_{train} = \{S^\tau\}_{1 \leq \tau \leq M}$ where $S^\tau = \{(X_j^\tau, Y_j^\tau)\}_{1 \leq j \leq n_\tau}$ is the τ-th domain and a PLM with frozen parameters ϕ. We introduce domain-conditioned parameters $\{\theta_\tau\}_{1 \leq \tau \leq M}$, which are optimized through gradient descent $\max_{\{\theta_\tau\}_{1 \leq \tau \leq M}} \sum_{\tau=1}^{M} \sum_{j=1}^{n_\tau} \log P(Y_j^\tau | X_j^\tau, \theta_\tau, \phi)$.

In HyperPrompt, domain-specific information is contained in hyperprompts, which are prependable key-value pairs only used in the self-attention of both the encoder and the decoder. Re-using previous notations, at the i-th transformer block, the original key-values K^i, V^i of the self-attention are augmented : $\quad K^{i\prime} = [P_{k,\tau}^i,\, K^i] \quad V^{i\prime} = \left[P_{v,\tau}^i,\, V^i\right]$

The different architectures to generate the hyperprompts are based on Mahabadi *et al.* (2021)'s work. For each domain, we learn a global prompt $W_\tau \in \mathbb{R}^{l \times d}$, a matrix containing domain-specific information, where l is the length of the prompt and d is the hidden size of the PLM's layers.

For each transformer block i, we have two local hypernetworks h_k^i and h_v^i that take a global prompt W_τ and output layer-specific and task-specific key-value.

$$P_{k,\tau}^i = h_k^i(W_\tau) = U_k^i(\sigma(D_k^i(W_\tau))) \tag{1}$$
$$P_{v,\tau}^i = h_v^i(W_\tau) = U_v^i(\sigma(D_v^i(W_\tau))) \tag{2}$$

The hypernetworks have a bottleneck architecture where $D_{k,v}^i \in \mathbb{R}^{d \times b}$ (resp. $U_{k,v}^i \in \mathbb{R}^{b \times h \times d_h}$) denotes the down-projection matrix (resp. the up-projection matrix), b is the bottleneck dimension and σ is a non-linear activation function.

HyperPrompt-Share In this method, at each transformer block i, all domains share the same two local hypernetworks h_k^i and h_v^i.

HyperPrompt-Separate In this method, at each transformer block i, each domain has its two own local hypernetworks h_k^i and h_v^i. Each domain-specific hyperprompt is trained independently, no knowledge is shared between domains.

HyperPrompt-Global In this method, we learn a task embedding $k_\tau \in \mathbb{R}^{t'}$ for each domain and a layer embedding $z_i \in \mathbb{R}^{t'}$ for each layer i. A projection network combines both layer and task embeddings into a layer-aware task embedding : $I_\tau^i = h(k_\tau, z_i) \in \mathbb{R}^t$. All tasks and all transformer blocks share two global hypernetworks H_k and H_v, which project I_τ^i into the weight matrices of the local hypernetworks (eqs. 1 and 2) :

$$(U_{k,\tau}^i, D_{k,\tau}^i) = H_k(I_\tau^i) = (W^{U_k}, W^{D_k}) I_\tau^i \tag{3}$$
$$(U_{v,\tau}^i, D_{v,\tau}^i) = H_v(I_\tau^i) = (W^{U_v}, W^{D_v}) I_\tau^i \tag{4}$$

2.3 HyperPrefix

Inspired by the two previous approaches, we propose to introduce a new multi-domain learning model, HyperPrefix. The domain-specific information is also contained in key-value pairs, which are prepended to original key, value matrices. The difference with HyperPrompt is how the prefix/hyperprompt is generated, here we use the same architecture as prefix-tuning.

For each attention class E, D and D_c, we learn a global prompt $W_\tau \in \mathbb{R}^{l \times d}$. All domains share the same hypernetwork H, a two-layered feed-forward network with a bottleneck architecture, which takes as input W_τ and generates key-value pairs for all transformer blocks.

$$P_\tau = \left\{(P^i_{k,\tau}, P^i_{v,\tau})\right\}_{1 \leq i \leq N} = H(W_\tau)$$

3 Experiments

3.1 Experimental setup

Pre-trained language models For our experiments, we use T5-small (60M parameters) and T5-base (220M parameters).

Models We compare fine-tuning, prefix-tuning, HyperPrefix and the different versions of HyperPrompt. Our implementation of the different models is based on Xie *et al.* (2022)'s implementation of prefix-tuning [1].

Datasets We evaluate the models on WebNLG datasets (Gardent *et al.*, 2017) (Shimorina & Gardent, 2018). WebNLG data consist of pairs of RDF triple sets (subject, predicate, object) and their associated reference text. The objective of data-to-text is then to generate the textual description associated to a set of triplets. WebNLG 2017 training samples are from 10 categories. In the test set, we have 5 additional unseen categories. We consider here each category as a domain (this is the closest instantiation of our multi-domain problem we have found in available data-to-text public benchmarks). We linearize the graph input, e.g. if the input is composed of two triples, as follows : $\text{subject}_1 : \text{predicate}_1 : \text{object}_1 \mid \text{subject}_2 : \text{predicate}_2 : \text{object}_2$

Metrics We report the following automatic metrics to evaluate the different models : BLEU (Papineni *et al.*, 2002), METEOR (Banerjee & Lavie, 2005) and TER (Snover *et al.*, 2006). We use (Post, 2018)'s implementation of BLEU and TER [2] and NLTK's implementation of METEOR [3].

Hyperparameters and training details All models are trained for 50 epochs, using Adafactor optimizer (Shazeer & Stern, 2018). The initial learning rate is set at 5e-5 and we use a linear learning rate scheduler. The batch size is set to 32. We apply early stopping based on the average development set metric. We use beam-search decoding with a beam size of 4. The prompt length l is set to 10. The bottleneck dimension b of all networks generating the prefix/prompt is $d/4$ where d the hidden size of the PLM's layers. For HyperPrompt-Global, the dimension of the task, layer and layer-aware task embeddings is 32, the hidden dimension of task-layer projection network h_t is 8.

1. https://github.com/hkunlp/unifiedskg
2. https://github.com/mjpost/sacrebleu
3. https://www.nltk.org/

3.2 Zero-shot learning

For the zero-shot experiments, we train the different models on WebNLG 2017. For prefix-tuning and fine-tuning, we treat the WebNLG 2017 dataset as a single domain. For the multi-domain learning models, we define each category as a domain. At testing time when the encountered category was not seen during training, the multi-domain models use the prefix/hyperprompt of the closest category. We use the Euclidean distance of GloVe embeddings (Pennington *et al.*, 2014) to measure the similarity between two categories as in (Clive *et al.*, 2022). We also test the importance of the prefix/hyperprompt in the decoder cross-attention by doing an ablation study. The scores are reported in Table 1.

	WebNLG								
	BLEU			METEOR			TER↓		
	S	U	A	S	U	A	S	U	A
SOTA (T5-large fine-tuned)	*65.82*	*56.01*	*61.44*	-	-	-	-	-	-
T5-small									
Fine-tuning	**62.93**	**44.60**	**54.83**	**63.52**	**53.55**	**58.75**	**52.86**	67.39	**59.45**
Prefix-tuning ✦	50.78	41.37	46.62	55.48	50.06	52.89	58.76	**65.22**	61.69
Prefix-tuning	49.94	40.32	45.69	54.56	49.19	51.99	58.79	65.31	61.75
HyperPrefix ✦	53.35	37.84	46.51	56.97	46.96	52.18	56.92	67.56	61.75
HyperPrefix	51.45	37.13	45.17	54.98	45.47	50.43	57.99	66.55	61.87
HyperPrompt-Global ✦	57.41	33.40	46.93	59.19	42.80	51.35	55.13	71.21	62.41
HyperPrompt-Global	54.53	32.29	44.83	57.20	43.08	50.45	56.03	69.76	62.25
HyperPrompt-Share ✦	54.01	34.02	45.28	56.14	43.47	50.07	56.68	70.01	62.73
HyperPrompt-Share	52.14	33.11	43.84	55.22	42.48	49.13	57.22	69.30	62.70
HyperPrompt-Sep ✦	55.79	30.15	44.60	57.29	40.40	49.20	56.63	74.39	64.68
HyperPrompt-Sep	54.70	29.36	43.71	56.55	39.84	48.55	56.58	73.56	63.82
T5-base									
Fine-tuning	**64.22**	**48.86**	**57.42**	**64.04**	55.04	**59.73**	**52.04**	63.56	**57.26**
Prefix-tuning ✦	-	-	-	-	-	-	-	-	-
Prefix-tuning	60.16	48.60	55.01	62.55	**55.36**	59.11	53.10	62.79	57.49
HyperPrefix ✦	-	-	-	-	-	-	-	-	-
HyperPrefix	62.01	47.70	55.67	62.34	54.52	58.59	52.89	**62.56**	57.28
HyperPrompt-Share ✦	61.15	41.09	52.30	61.55	51.19	56.59	53.41	66.74	59.45
HyperPrompt-Share	60.35	41.70	52.15	60.63	51.27	56.15	54.70	66.36	59.99
HyperPrompt-Sep ✦	58.55	37.72	49.43	58.84	47.32	53.33	55.73	67.69	61.15
HyperPrompt-Sep	58.23	37.03	48.90	59.27	46.80	53.30	55.92	68.81	61.76

TABLE 1 – Results on WebNLG test set, best results are marked in bold. T5-large fine-tuned results are from (Ribeiro *et al.*, 2021). ✦ indicates that we add prefix/prompt in all attention classes vs only in the self-attention. S, U and A refer to scores for the *seen*, *unseen* and *all* portions of the test set

Overall, the fine-tuned PLMs achieve the best results. The performance of prefix-based methods seems to heavily depend on the model's size. For T5-small, the scores of prefix-based methods are way below the results of T5-small fine-tuned. This difference between fine-tuning and prefix-based gets smaller when using T5-base, we go from a gap of between 8 and 11 pts of BLEU on the whole test set to a difference between 2 and 9 pts. These results point in the same direction as (Lester *et al.*, 2021), where the performance gap between prompt-tuning and fine-tuning narrows as the size of the model grows.

When comparing the different HyperPrompt methods, we can see that HyperPrompt-Global performs the best. Hyperprompt-Separate, on the other side, performs the worst, which is expected since no knowledge is shared between domains. Among the multi-domain learning methods, HyperPrefix works best on new domains. We note that all multi-domain models hallucinate and generate words linked to the mapped category. Qualitatively, prefix-based and fine-tuned models are able to copy subjects and objects that were not seen during training, but they struggle to correctly transcribe new relations when used in the zero-shot manner (examples in Figure 1).

Adding prefix/hyperprompt in the decoder cross-attention module led to a slight score improvement

Input : Ace Wilder : genre : Hip hop music | Hip hop music : stylistic_origin : Disco **Reference text** : Ace Wilder's musical genre is Hip hop music which has its origins in Disco.
HyperPrefix : Ace Wilder is a character in hip hop music which is stylistically influenced by disco. **Fine-tuned T5-base** : Ace Wilder is a member of the genre Hip Hop music, whose stylistic origin is Disco.

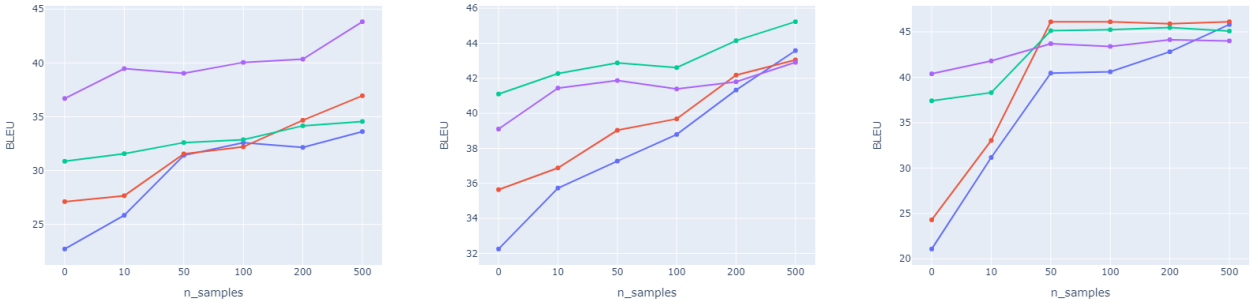
FIGURE 1 – Hallucinations produced by models with T5-base for a sample from the *Artist* category, which is not seen during training. Hallucination linked to the mapped category *ComicsCharacter* is highlighted in blue. Hallucinations linked to new predicates are emphasized in red.

with T5-small, whereas it is not significant for T5-base. In our following experiments, we only prepended prefix/hyperprompt in the self-attention module.

3.3 Few-shot domain adaptation

Since the tested models have good performance on seen domains, we evaluate if an already trained model on a set of domains can generalize to a new domain in the few-shot regime.

We consider a new category $C \in \{Artist, CelestialBody\}$ and subsample various number of examples (10, 50, 100, 200, 500) from WebNLG 2020 training set to constitute our training data. We use as validation set (resp. test set) the subset of all samples of category C from the WebNLG2020 validation dataset (resp. test set). We evaluate our models trained on WebNLG 2017 on the new domain after few-shot fine-tuning. For HyperPrompt-Share and HyperPrefix, we initialize the global prompt W_C with the global prompt of the closest category.



(a) *Artist*, T5-small (b) *Artist*, T5-base (c) *CelestialBody*, T5-base

FIGURE 2 – Few-shot domain transfer results : BLEU scores (avg. across 2 runs) of Hyperprompt-Share, HyperPrefix, prefix-tuning and fine-tuning models on two new domains *Artist* and *CelestialBody* after few-shot fine-tuning vs. # of training samples (0, 10, 50, 100, 200, 500)

We can observe in Figures 2(a) and 2(b) a difference in behaviour depending on the model's size. When using T5-small, the performance of prefix-based methods are way below regular fine-tuning. On the other hand, the BLEU scores of HyperPrompt-Share, HyperPrefix and prefix-tuning based on T5-base are able to reach fine-tuned T5-base when the number of samples for few-shot fine-tuning gets bigger.

4 Conclusion

Our study of prefix-based methods for Transformer shows that their performance depends heavily on the PLM's size. Scores of best prefix-based models were only comparable to fine-tuning for both trained and new categories when using T5-base. We also observe a difference of behaviour linked to the model's size for few-shot domain adaptation, where prefix-based models are able to reach results of fine-tuning as the number of training samples increases only with T5-base.

In the context of saving memory, additional experiments on few-shot domain adaption could be done : we could freeze the hypernetworks in HyperPrefix and HyperPrompt-Share and only train the new global prompts associated with the new domains. This type of few-shot domain adaption would be interesting since we would only need to save an additional matrix of dimension $(l \times d)$.

References

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan : Association for Computational Linguistics.

CLIVE J., CAO K. & REI M. (2022). Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, p. 363–382, Abu Dhabi, United Arab Emirates (Hybrid) : Association for Computational Linguistics.

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.

GARDENT C., SHIMORINA A., NARAYAN S. & PEREZ-BELTRACHINI L. (2017). Creating training corpora for nlg micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 179–188 : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1017.

HE Y., ZHENG S., TAY Y., GUPTA J., DU Y., ARIBANDI V., ZHAO Z., LI Y., CHEN Z., METZLER D., CHENG H.-T. & CHI E. H. (2022). HyperPrompt : Prompt-based task-conditioning of transformers. In K. CHAUDHURI, S. JEGELKA, L. SONG, C. SZEPESVARI, G. NIU & S. SABATO, Éds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 de *Proceedings of Machine Learning Research*, p. 8678–8690 : PMLR.

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.

LESTER B., AL-RFOU R. & CONSTANT N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 3045–3059.

LI X. L. & LIANG P. (2021). Prefix-tuning : Optimizing continuous prompts for generation. 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. 4582–4597.

MAHABADI R. K., RUDER S., DEHGHANI M. & HENDERSON J. (2021). Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 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. 565–576.

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.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543.

POST M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation : Research Papers*, p. 186–191, Belgium, Brussels : Association for Computational Linguistics.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners.

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*, **21**(1), 5485–5551.

RIBEIRO L. F., SCHMITT M., SCHÜTZE H. & GUREVYCH I. (2021). Investigating pretrained language models for graph-to-text generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, p. 211–227.

SHAZEER N. & STERN M. (2018). Adafactor : Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, p. 4596–4604 : PMLR.

SHIMORINA A. & GARDENT C. (2018). Handling rare items in data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, p. 360–370 : Association for Computational Linguistics.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas : Technical Papers*, p. 223–231, Cambridge, Massachusetts, USA : Association for Machine Translation in the Americas.

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 A., PRUKSACHATKUN Y., NANGIA N., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2019). Superglue : A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, **32**.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. (2018). Glue : A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 353–355.

XIE T., WU C. H., SHI P., ZHONG R., SCHOLAK T., YASUNAGA M., WU C.-S., ZHONG M., YIN P., WANG S. I., ZHONG V., WANG B., LI C., BOYLE C., NI A., YAO Z., RADEV D., XIONG C., KONG L., ZHANG R., SMITH N. A., ZETTLEMOYER L. & YU T. (2022). UnifiedSKG :

Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 602–631, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

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

Auto-apprentissage et renforcement pour une analyse jointe sur données disjointes : étiquetage morpho-syntaxique et analyse syntaxique

Fang Zhao Timothée Bernard
Laboratoire de linguistique formelle, Université Paris Cité
fang.zhao@etu.u-paris.fr, timothee.bernard@u-paris.fr

RÉSUMÉ

Cet article se penche sur l'utilisation de données disjointes pour entraîner un système d'analyse jointe du langage naturel. Dans cette étude exploratoire, nous entraînons un système à prédire un étiquetage morpho-syntaxique et une analyse syntaxique en dépendances à partir de phrases annotées soit pour l'une de ces tâches, soit pour l'autre. Deux méthodes sont considérées : l'auto-apprentissage et l'apprentissage par renforcement, pour lequel nous définissons une fonction de récompense encourageant le système à effectuer des prédictions même sans supervision. Nos résultats indiquent de bonnes performances dans le cas où les données disjointes sont issues d'un même domaine, mais sont moins satisfaisants dans le cas contraire. Nous identifions des limitations de notre implémentation actuelle et proposons en conséquence des pistes d'amélioration.

ABSTRACT

Self-training and reinforcement for joint analysis on disjoint data: part-of-speech tagging and syntactic parsing

In this exploratory study, we train a system to jointly perform POS tagging and syntactic dependency parsing on sentences annotated for only the former or the latter. We consider two methods: self-training and reinforcement learning, for which we define a reward function that can reward the system even for predictions for which no supervision is available. Our results indicate good performance when the disjoint data come from the same domain but are less satisfactory otherwise. We identify limitations of our current implementation and suggest possible improvements accordingly.

MOTS-CLÉS : apprentissage semi-supervisé, apprentissage par renforcement, multi-tâche, analyse jointe, étiquetage morpho-syntaxique, analyse syntaxique en dépendances, adaptation de domaine.

KEYWORDS: semi-supervised learning, reinforcement learning, multitask, joint analysis, POS tagging, syntactic dependency parsing, domain adaptation.

1 Introduction et travaux connexes

Nous cherchons dans cet article des solutions au potentiel manque de données nécessaires à l'entraînement de systèmes d'analyse jointe du langage naturel. Alors que, dans un système multi-tâche neuronal par simple partage de paramètres (Caruana, 1997; Zhang & Yang, 2017), les sorties prédites pour les différentes tâches sont calculées de manière indépendante après l'encodage de l'entrée, un système d'analyse jointe modélise de manière non triviale l'interaction entre les différentes tâches.

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 2 : travaux de recherche originaux – articles courts, pages 82–90.

Contrairement à un système multi-tâche simple, un tel système nécessite donc a priori pour son entraînement des données entièrement annotées, c'est-à-dire pour lesquelles les annotations couvrent, pour chaque texte, toutes les tâches modélisées. Le nombre de corpus ainsi annotés sur plusieurs niveaux d'analyse linguistique étant restreint, nous étudions la possibilité d'utiliser des jeux de données disjoints — c'est-à-dire tels que chacun n'est muni d'annotations que sur l'un des niveaux cibles — en comparant plusieurs techniques visant à palier à l'existence d'annotations manquantes. Notons que ce problème a été abordé notamment par Peng *et al.* (2018) dans le cadre d'un système par optimisation linéaire en nombres entiers.

Notre étude se fonde sur une variante du système de Bernard (2021), capable d'*intégrer* des tâches d'étiquetages lexicales ainsi que des tâches de création de dépendances bi-lexicales, c'est-à-dire de les traiter ensemble comme une tâche unique. Il s'agit d'un système par transition ayant la particularité qu'à chaque étape, jusqu'à une action par token peut être effectuée ; à tout moment, l'action d'un token peut indifféremment se rapporter à n'importe laquelle des tâches traitées, sans contrainte particulière. Chaque décision est basée sur toutes les structures précédemment prédites, ce qui permet une interaction complète entre les différentes tâches. Ce système a été choisi pour sa flexibilité (il est en particulier possible de l'entraîner sans modification sur des jeux de données disjoints, au prix des performances) et son mode d'entraînement : pré-entraîné de manière supervisée, il a été conçu pour ensuite apprendre à exploiter au mieux les interactions entre niveaux linguistiques lors d'une phase d'*apprentissage par renforcement* (Sutton & Barto, 2018). Contrairement à l'apprentissage supervisé standard qui consiste à entraîner explicitement un modèle à reproduire des annotations connues à l'avance, il s'agit d'entraîner le modèle sans le guider a priori mais en associant une *récompense* à chacune de ses actions. L'apprentissage par renforcement nécessite donc le calcul de ces récompenses, qui se fait généralement à partir d'annotations. Il est cependant possible d'apprendre une fonction de récompense ; c'est ce que fait notamment l'algorithme *apprentissage par renforcement inverse* (Ng & Russell 2000 ; voir aussi Finn *et al.* 2017). Ici, nous testons la possibilité d'utiliser les prédictions du modèle lui-même pour calculer les récompenses de ses actions pour lesquelles nous ne disposons pas d'annotation de référence.

Nous travaillons ici avec des données en anglais pour deux tâches classiques du traitement automatique des langues : l'analyse morpho-syntaxique et l'analyse syntaxique en dépendances. Bien que les corpus annotés en syntaxe soient généralement aussi annotés en morpho-syntaxe, l'inverse n'est pas le cas. De plus, ces expériences doivent être vues comme une première étape avant d'aborder le cas de tâches plus complexes, et la forte interdépendance entre analyse syntaxique et analyse morpho-syntaxique nous intéresse ici. Depuis l'article de Li *et al.* (2011), un certain nombre de travaux se sont penchés sur l'analyse jointe de ces deux tâches, mais toujours à partir de données entièrement annotées (notamment Hatori *et al.*, 2011; Bohnet & Nivre, 2012; Alberti *et al.*, 2015; Nguyen & Verspoor, 2018; Bernard, 2021).

Nous expérimentons deux méthodes. La première se concentre sur la phase d'apprentissage par renforcement et consiste à définir une fonction de récompense visant à encourager de manière pertinente le système à effectuer des prédictions même lorsque les annotations correspondantes ne sont pas disponibles. La seconde s'applique autant à la phase de pré-entraînement qu'à la phase de renforcement et est une forme d'*auto-apprentissage* (*self-learning* ou *-training* ; Nigam & Ghani 2000 ; à rapprocher, pour les modèles génératifs, de l'*algorithme espérance-maximisation*, ou EM ; Dempster *et al.* 1977) consistant, à un temps t, à utiliser le système pour prédire au moins une partie des structures d'annotations manquantes du corpus d'entraînement et ainsi pouvoir poursuivre l'apprentissage au temps $t + 1$ sur le jeu de données ainsi complété (mais donc aussi bruité).

Nous montrons l'efficacité de ces deux méthodes dans un scénario utilisant des données disjointes créées à partir d'un même corpus. Nous testons ensuite nos méthodes dans un scénario plus réaliste, utilisant des données issues de deux corpus de domaines différents. Il s'agit d'un cas de *glissement de domaine* (*domain shift* ; Ramponi & Plank 2020; Li 2012). Les résultats dans ce cadre-là sont pour l'instant moins probants, mais nous identifions un certain nombre de limitations de notre implémentation actuelle les expliquant et que nous corrigerons par la suite.

2 Modèle

Nous décrivons ici brièvement le modèle utilisé. Il ne diffère de celui de Bernard (2021) que par les tâches traitées (analyse morpho-syntaxique et analyse syntaxique en dépendances, sans analyse sémantique) et le processus d'entraînement (voir sections suivantes).

Le système fait intervenir quatre types d'actions. Pour un token de position i, effectuer (a) TAG-t consiste à assigner l'étiquette morpho-syntaxique t à i, (b) SYN-j-l à ajouter une dépendance syntaxique d'étiquette l de j vers i, (c) ROOT à définir i comme la racine de l'arbre syntaxique, et (d) HALT à ne rien faire.

À chaque étape de l'analyse d'une phrase, le système encode la phrase elle-même [1] ainsi que les prédictions effectuées aux étapes précédentes sous forme d'un vecteur par token. Chacun de ces encodages est converti en une distribution de probabilité sur l'ensemble des actions puis, pour chaque token, l'action de probabilité maximale est effectuée (une action par token est donc effectuée à chaque étape). Le système s'arrête lorsque pour chaque token est sélectionnée l'action HALT ; il n'est pas contraint à produire des structures d'annotations (ex : arbres syntaxiques) complètes. Cependant, en pratique, le taux d'analyse pour chaque tâche (le ratio du nombre de dépendances ou d'étiquettes prédites par rapport à celles annotées) dépasse toujours 99% lorsque l'entraînement s'effectue à partir de données entièrement annotées.

3 Données

Pour notre première expérience, nous utilisons la portion WSJ (*Wall Street Journal*) du corpus *Penn Treebank* (PTB ; Marcus *et al.* 1993). Nous utilisons comme arbres de dépendances de référence des conversions des arbres de constituants du PTB (de Marneffe *et al.*, 2006). Pour ces données, nous utilisons comme source la tâche 18 de la campagne SemEval 2015 (Oepen *et al.*, 2015), qui fournit aussi des graphes d'analyse sémantique de références, afin de pouvoir aisément mettre en place de futures expériences impliquant cette tâche plus complexe. Nous n'utilisons que les portions d'entraînement (sections 0 à 19 ; 33964 phrases) et de développement (section 20 ; 1692 phrases) : tous les résultats reportés dans ce texte sont calculés sur les données de développement ; nous réservons les données de test (section 21) pour une phase de comparaison ultérieure.

Pour étudier un scénario plus réaliste, impliquant un glissement de domaine, nous avons également démarré une expérience sur un jeu de données construit à partir de deux corpus de type *Universal Dependencies* (UD ; Nivre *et al.*, 2020), contenant eux aussi des phrases associées à des séquences

1. Des vecteurs GloVe (Pennington *et al.*, 2014) sont utilisés.

d'étiquettes morpho-syntaxiques et des arbres de dépendances syntaxiques. Le premier corpus, *Georgetown University Multilayer* (GUM ; Zeldes, 2017), comprend un douzaine de types de textes très variés, incluant des textes académiques, des billets de blog, des œuvres de fiction et des textes issus de médias sociaux. Le second corpus, *English Web Treebank* (EWT ; Silveira *et al.*, 2014), comprend des textes issus de diverses sources en ligne et regroupés en cinq catégories : *weblogs*, *newsgroups*, *emails*, *reviews* et *Yahoo ! answers*. Nous avons créé un jeu d'entraînement en sélectionnant aléatoirement 6911 phrases dans la portion d'entraînement de chacun des deux corpus pour un total de 13822 phrases, et de même avec un total de 2234 phrases pour le jeu de développement. Pour ce corpus aussi les résultats reportés sont calculés sur les données de développement.

Dans les expériences détaillées dans les sections suivantes, la configuration COMPLET correspond à un entraînement sur les données d'entraînement telles que décrites ci-dessus. Pour la configuration DISJOINT, par contre, pour une moitié des phrases d'entraînement les arbres syntaxiques sont ignorés et pour l'autre moitié les séquences d'étiquettes morpho-syntaxiques sont ignorées. Dans le cas du corpus PTB, cette division est effectuée au hasard. Dans le cas du corpus mixte GUM+EWT, nous ignorons les arbres syntaxiques des phrases issues du corpus EWT et inversement pour les séquences d'étiquettes morpho-syntaxiques.

4 Scénario idéal : corpus homogène

Rappelons que le modèle utilisé commence son entraînement par une phase de pré-entraînement, durant laquelle le système apprend à reproduire de manière supervisée les annotations de références connues. Dans la configuration DISJOINT, le système apprend donc à produire, pour chaque phrase, soit un arbre syntaxique, soit une séquence d'étiquettes morpho-syntaxiques, mais jamais les deux. Pour améliorer cette situation, nous implémentons un processus d'auto-apprentissage : trois fois par époque, le corpus d'entraînement est analysé par le système et les prédictions obtenues sont utilisées pour compléter les annotations et ainsi servir à l'apprentissage du système lui-même.

Après la phase de pré-entraînement commence la phase d'apprentissage par renforcement. Durant cette phase, le système analyse librement les phrases qu'on lui présente et est entraîné avec l'algorithme REINFORCE (Williams, 1992) à maximiser des *récompenses* associées à chaque action effectuée : positives pour les actions correspondant aux annotations de références, négatives pour celles contraires à celles-ci, nulles sinon (typiquement, lorsqu'il n'y a pas d'annotation correspondante). Ce système de récompenses est le système *zéro*. Un de ses effets sur notre corpus mixte GUM+EWT est qu'il n'incite pas le modèle à analyser les phrases de GUM en morpho-syntaxe ni celles de EWT en syntaxe. Afin d'encourager le système à analyser chaque phrase sur les deux niveaux linguistiques, nous implémentons le système de récompenses *auto* (pour « auto-renforcement » ; l'idée est qu'à l'issu du pré-entraînement, la plupart des prédictions du modèle sont correctes) : pour chaque token issu d'une phrase pour laquelle l'arbre syntaxique de référence n'est pas disponible, la première dépendance entrante prédite est considérée comme correcte et mène donc à une récompense positive, et de même pour les étiquettes morpho-syntaxiques [2]. Notons que l'auto-apprentissage est aussi possible durant la phase de renforcement. Cependant, lorsque l'auto-apprentissage complète entièrement les annotations, la distinction entre les systèmes *zéro* et *auto* disparaît.

2. Nous précisons « la première » car le système de transition du modèle lui permet de choisir une nouvelle dépendance entrante pour un token en ayant déjà une (effaçant alors la prédiction initiale et menant à une récompense négative d'après *auto*), et de même pour les étiquettes morpho-syntaxiques. Dans les présentes conditions, cependant, cette possibilité n'est pas utilisée de manière significative (Zhao, 2022).

La table 1 présente les performances sur le jeu de données PTB de différents modèles [3]. Les deux premières lignes concernent la configuration COMPLET et indiquent les bornes supérieures du modèle avec et sans renforcement. Les lignes suivantes concernent la configuration DISJOINT, avec cinq modèles se distinguant par un pré-entraînement simple (-AR) ou suivi d'un apprentissage par renforcement (+AR[zéro] ou +AR[auto] suivant le système de récompense utilisé) et de la complétion (+AA) ou non des annotations par auto-apprentissage (autant durant le pré-entraînement que la phase de renforcement si elle a lieu).

Données	Apprentissage	SYN	TAG	Chevauchement	Taux SYN	Taux TAG
COMPL	-AR*	0,912$_{(0,912/0,912)}$	0,971$_{(0,971/0,971)}$	1,000	1,000	1,000
	+AR$_{[zéro]}$	0,918$_{(0,919/0,918)}$	0,973$_{(0,973/0,973)}$	1,000	0,999	1,000
DISJ	-AR*	0,788$_{(0,918/0,691)}$	0,374$_{(0,940/0,235)}$	0,096	0,753	0,249
	-AR+AA	0,852$_{(0,883/0,823)}$	0,969$_{(0,969/0,969)}$	0,974	0,932	1,000
	+AR$_{[zéro]}$*	0,883$_{(0,896/0,870)}$	0,968$_{(0,969/0,967)}$	0,986	0,970	0,999
	+AR$_{[auto]}$	0,891$_{(0,891/0,891)}$	0,970$_{(0,970/0,970)}$	1,000	0,999	1,000
	+AR$_{[auto]}$+AA	0,898$_{(0,900/0,897)}$	0,971$_{(0,971/0,971)}$	0,999	0,997	1,000

TABLE 1 – Performances sur le PTB ; chaque valeur est une moyenne sur 9 exécutions. Format : F1$_{(Précision/Rappel)}$ pour la syntaxe (SYN) et la morpho-syntaxe (TAG). Pour chaque groupe de comparaison, les modèles sont comparés avec un modèle de référence (indiqué par *) et les différences significatives sont indiquées ainsi (p-valeur < 0,01 ; test de permutation de Pitman).

Nous constatons tout d'abord que dans la configuration DISJOINT, les performances du modèle -AR sont très faibles. Ce système produit des analyses plutôt justes (voir les valeurs de précision) mais très incomplètes : la colonne Taux SYN indique que seuls 75,3% des tokens se voient prédire une tête syntaxique ou le statut de racine, et la colonne Taux TAG indique que 24,9% des tokens se voient prédire une étiquette morpho-syntaxique. Le taux de chevauchement (un indice de Jaccard) nous indique que seulement 9,6% des tokens pour lesquels le système a fait au moins une prédiction a reçu une annotation pour chacune des deux tâches. Ces résultats ne sont pas particulièrement surprenants au vu du processus d'entraînement. Nous remarquons ensuite que l'utilisation de l'auto-apprentissage durant le pré-entraînement améliore grandement la situation, notamment en ce qui concerne l'étiquetage morpho-syntaxique (pour lequel les performances sont alors proches de la borne supérieure). Le renforcement, avec le système de récompense *zéro* et encore plus avec *auto*, s'avère particulièrement efficace, et les meilleures performances sont obtenus en combinant le renforcement avec l'auto-apprentissage.

5 Scénario plus réaliste : glissement de domaine

Let deux premières lignes de la table 2 concernent la configuration COMPLET du corpus mixte GUM+EWT. Les quatre lignes suivantes concerne une configuration similaire, MONO-CORPUS, dans laquelle les modèles sont aussi entraînées sur des données complètes, mais uniquement sur l'un des deux sous-corpus. Les cinq dernières lignes correspondent à la configuration DISJOINT. [4]

3. Les modèles n'étant pas contraints à effectuer des prédictions complètes, les mesures F1 sont naturelles pour chacune des tâches. Notons cependant que plus les taux d'analyse approchent 1, plus les F1 en syntaxe (SYN) et morpho-syntaxe (TAG) correspondent aux mesures habituelles : LAS (*labelled attachement score*) et exactitude.

4. La comparaison entre les configurations MONO-CORPUS est DISJOINT est intéressante mais non évidente. En effet, si les modèles MONO-CORPUS ne sont entraînés que sur l'un des sous-corpus (ce qui les place dans une situation de glissement

Données	Apprentissage	SYN (GUM)	TAG (GUM)	SYN (EWT)	TAG (EWT)	Chev.
GUM+EWT	-AR	0,810(0,810/0,809)	0,956(0,956/0,956)	0,792(0,793/0,791)	0,937(0,937/0,937)	1,000
	+AR[zéro]	0,830(0,841/0,820)	0,960(0,960/0,960)	0,809(0,817/0,800)	0,941(0,941/0,941)	0,994
GUM	-AR	0.783(0.784/0.783)	0.949(0.949/0.949)	**0.719(0.721/0.718)**	**0.904(0.904/0.903)**	1,000
	+AR[zéro]	0,817(0,832/0,802)	0,956(0,956/0,956)	**0,741(0,754/0,728)**	**0,908(0,908/0,908)**	0,994
EWT	-AR	**0.721(0.722/0.720)**	**0.931(0.931/0.931)**	0.766(0.767/0.764)	0.929(0.929/0.929)	1,000
	+AR[zéro]	**0,749(0,768/0,730)**	**0,936(0,936/0,936)**	0,798(0,812/0,785)	0,935(0,935/0,935)	0,987
GUM(SYN) + EWT(TAG)	-AR	0.685(0.730/0.646)	**0.176(0.913/0.108)**	**0.444(0.719/0.328)**	0.632(0.911/0.493)	0,037
	-AR+AA	0.735(0.760/0.712)	**0.938(0.939/0.938)**	**0.634(0.732/0.563)**	0.928(0.928/0.928)	0,936
	+AR[zéro]	0.695(0.757/0.644)	**0.281(0.914/0.177)**	**0.474(0.735/0.356)**	0.711(0.915/0.588)	0,155
	+AR[auto]	0,725(0,730/0,720)	**0,915(0,924/0,907)**	**0,679(0,682/0,675)**	0,903(0,916/0,890)	0,976
	+AR[auto]+AA	0,791(0,833/0,754)	**0.945(0,945/0,945)**	**0,741(0,782/0,705)**	0,934(0,934/0,934)	0,978

TABLE 2 – Performances sur l'UD ; chaque valeur est une moyenne sur 9 exécutions. Format : F1(Précision/Rappel) pour la syntaxe (SYN) et la morpho-syntaxe (TAG). Chev. indique le taux de chevauchement évalué sur la totalité du corpus mixte GUM+EWT. Les valeurs correspondant à des évaluations hors-domaines sont notées **ainsi**.

Nous remarquons ici aussi que le pré-entraînement seul (-AR) mène à des modèles qui n'analysent que très rarement leur entrée à la fois en syntaxe et en morpho-syntaxe. Contrairement à ce qui se passe sur le PTB, cependant, le renforcement avec le système de récompense *zéro* (+AR[zéro]) ne suffit pas à corriger cette faiblesse. L'utilisation du système de récompense *auto* améliore grandement cette situation, sans toutefois mener à des analyses toujours complète. De même, l'auto-apprentissage seul est moins efficace que sur le PTB. L'une des raisons est que notre implémentation actuelle ne force pas l'auto-apprentissage à compléter entièrement les annotations manquantes et que, si cela se produit néanmoins très souvent sur le PTB, cela est moins le cas sur le corpus mixte GUM+EWT (ce que l'on voit notamment en comparant les taux de chevauchement des modèles -AR+AA dans les deux tableaux). Cela explique aussi pourquoi l'utilisation du système de récompense *auto* plutôt que *zéro* (non inclus dans la table) fait une différence même en combinaison avec l'auto-apprentissage.

Les meilleurs résultats, obtenus pour le modèle entraîné en employant la technique d'auto-apprentissage en conjonction avec le système de récompense *auto* lors du renforcement, restent clairement en deçà de celles de la configuration COMPLET. Toutefois, ce modèle obtient des performances en étiquetage morpho-syntaxique en domaine (sur EWT) similaires à celles du modèle MONO-CORPUS entraîné sur EWT et est meilleur que lui hors-domaine (sur GUM). En analyse syntaxique, les résultats sont moins convaincants : le système est moins performant en domaine (sur GUM) que le modèle MONO-CORPUS entraîné sur GUM, et lui est équivalent hors-domaine (sur EWT).

6 Discussion

Nous avons présenté des travaux en cours sur la possibilité d'entraîner des systèmes d'analyse jointe en utilisant des jeux de données disjoints. L'une des méthodes étudiées consiste en l'utilisation d'apprentissage par renforcement avec un système de récompense encourageant le système à effectuer des prédictions même lorsqu'il n'existe pas d'annotations de référence. L'autre méthode, une

de domaine total lorsque évalués sur l'autre sous-corpus), l'utilisation de données de référence complètes lors de l'entraînement est un avantage notable.

forme d'auto-apprentissage, consiste en l'entraînement du modèle sur des données (plus ou moins partiellement) complétées par lui-même. (Nous prévoyons dans un jeu d'expériences plus complet de comparer ces méthodes à l'utilisation d'un système multi-tâche par simple partage de paramètres, construit autour d'un encodeur commun à différents sous-systèmes — un par tâche traitée.)

Sur un corpus homogène, nous avons constaté de nets gains de performances grâce à l'utilisation de ces deux méthodes. Cependant, lorsque les données d'entraînement pour chaque tâche proviennent de corpus distincts, leur efficacité est moindre. Il faut garder en tête les difficultés inhérentes au glissement de domaine (Ramponi & Plank, 2020; Li, 2012), mais nous pouvons aussi pointer du doigt un certain nombre de caractéristiques de l'implémentation actuelle qui interagissent mal avec la configuration DISJOINT. Comme nous travaillons uniquement sur des tâches impliquant un nombre connu de prédictions par token (une dépendance entrante et une étiquette ; contrairement à, par exemple, la tâche d'analyse sémantique traitée par le modèle Bernard 2021 que nous reprenons), il serait possible de forcer le système à effectuer des prédictions complètes. Cela aurait un intérêt en ce qui concerne les sorties effectives du système, mais potentiellement aussi en ce qui concerne la complétion des annotations par auto-apprentissage. À propos de l'auto-apprentissage en particulier, le système récupère, pour chaque phrase, ses propres prédictions concernant l'une des tâches, que nous lui faisons calculer à partir de la suite de tokens brute alors qu'il serait possible de lui fournir aussi les annotations concernant l'autre tâche en entrée. Notons enfin que la fonction de coût minimisée durant le pré-entraînement pénalise non seulement les prédictions fausses, mais aussi celles pour lesquelles il n'existe pas d'annotations. Une alternative — notre prochaine étape — est d'ignorer les probabilités assignées aux actions correspondant à ces dernières, ce qui pourrait avoir un fort impact sur la qualité des modèles pré-entraînés en configuration DISJOINT.

Remerciements

Ces travaux ont bénéficié d'un financement Émergence 2021 (projet SYSNEULING) de l'IdEx Université Paris Cité.

Références

ALBERTI C., WEISS D., COPPOLA G. & PETROV S. (2015). Improved Transition-Based Parsing and Tagging with Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1354–1359, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1159.

BERNARD T. (2021). Multiple tasks integration : Tagging, syntactic and semantic parsing as a single task. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 783–794, Online : Association for Computational Linguistics.

BOHNET B. & NIVRE J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 1455–1465, Jeju Island, Korea : Association for Computational Linguistics.

CARUANA R. (1997). Multitask Learning. *Machine Learning*, **28**(1), 41–75. DOI : 10.1023/A :1007379606734.

DE MARNEFFE M.-C., MACCARTNEY B. & MANNING C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy : European Language Resources Association (ELRA).

DEMPSTER A. P., LAIRD N. M. & RUBIN D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38. 54689.

FINN C., YU T., FU J., ABBEEL P. & LEVINE S. (2017). Generalizing skills with semi-supervised reinforcement learning. In *International Conference on Learning Representations*.

HATORI J., MATSUZAKI T., MIYAO Y. & TSUJII J. (2011). Incremental Joint POS Tagging and Dependency Parsing in Chinese. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, p. 1216–1224, Chiang Mai, Thailand : Asian Federation of Natural Language Processing.

LI Q. (2012). *Literature Survey : Domain Adaptation Algorithms for Natural Language Processing*. Rapport interne, Department of Computer Science, City University of New York.

LI Z., ZHANG M., CHE W., LIU T., CHEN W. & LI H. (2011). Joint Models for Chinese POS Tagging and Dependency Parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 1180–1191, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

NG A. Y. & RUSSELL S. J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, p. 663–670, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

NGUYEN D. Q. & VERSPOOR K. (2018). An Improved Neural Network Model for Joint POS Tagging and Dependency Parsing. In *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 81–91, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/K18-2008.

NIGAM K. & GHANI R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, CIKM '00, p. 86–93, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/354756.354805.

NIVRE J., DE MARNEFFE M.-C., GINTER F., HAJIČ J., MANNING C. D., PYYSALO S., SCHUSTER S., TYERS F. & ZEMAN D. (2020). Universal Dependencies v2 : An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4034–4043, Marseille, France : European Language Resources Association.

OEPEN S., KUHLMANN M., MIYAO Y., ZEMAN D., CINKOVA S., FLICKINGER D., HAJIC J. & URESOVA Z. (2015). SemEval 2015 Task 18 : Broad-Coverage Semantic Dependency Parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 915–926 : Association for Computational Linguistics. DOI : 10.18653/v1/S15-2153.

PENG H., THOMSON S., SWAYAMDIPTA S. & SMITH N. A. (2018). Learning Joint Semantic Parsers from Disjoint Data. 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. 1492–1502, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1135.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). GloVe : Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

RAMPONI A. & PLANK B. (2020). Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 6838–6855, Barcelona, Spain (Online) : International Committee on Computational Linguistics. DOI : 10.18653/v1/2020.coling-main.603.

SILVEIRA N., DOZAT T., DE MARNEFFE M.-C., BOWMAN S., CONNOR M., BAUER J. & MANNING C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 2897–2904, Reykjavik, Iceland : European Language Resources Association (ELRA).

SUTTON R. S. & BARTO A. G. (2018). *Reinforcement Learning : An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, second edition édition.

WILLIAMS R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8**(3-4), 229–256. DOI : 10.1007/BF00992696.

ZELDES A. (2017). The GUM corpus : Creating multilayer resources in the classroom. *Language Resources and Evaluation*, **51**(3), 581–612. DOI : http ://dx.doi.org/10.1007/s10579-016-9343-x.

ZHANG Y. & YANG Q. (2017). A survey on multi-task learning. DOI : 10.48550/ARXIV.1707.08114.

ZHAO F. (2022). Auto-correction dans un analyseur neuronal par transitions : un comportement factice ? In *Actes de la 29e conférence sur le Traitement Automatique des Langues Naturelles et des 24es Rencontres des Etudiants Chercheurs en Informatique et Traitement Automatique des Langues*, volume RECITAL, p. 20–32, Avignon, France : ATALA.