Qwant Research @DEFT 2019 : appariement de documents et extraction d'informations à partir de cas cliniques

Estelle Maudet, Oralie Cattan, Maureen de Seyssel, Christophe Servan

Qwant Research @DEFT 2019 : appariement de documents et extraction d'informations à partir de cas cliniques

Estelle Maudet, Oralie Cattan, Maureen de Seyssel, Christophe Servan
QWANT RESEARCH, 7 Rue Spontini, 75116 Paris, France
initiale.nom@qwant.com

RÉSUMÉ

Dans ce papier, nous présentons la participation de Qwant Research aux tâches 2 et 3 de l'édition 2019 du défi fouille de textes (DEFT 2019) portant cette année sur l'analyse de documents cliniques rédigés en français. La tâche 2 est une tâche de similarité sémantique qui demande d'apparier cas cliniques et discussions médicales deux à deux. Pour résoudre cette tâche, nous proposons une approche reposant sur des modèles de langue et évaluons l'impact de différents pré-traitements et de différentes techniques d'appariement sur les résultats. Pour la tâche 3, nous avons développé un système d'extraction d'information qui produit des résultats encourageants en terme de précision. Nous avons expérimenté deux approches différentes, l'une se fondant exclusivement sur l'utilisation de réseaux de neurones pour traiter la tâche, l'autre reposant sur l'exploitation des informations linguistiques issues d'une analyse syntaxique.

ABSTRACT

Document matching and information retrieval using clinical cases.

This paper reports on Qwant Research contribution to tasks 2 and 3 of the DEFT 2019's challenge, focusing on French clinical cases analysis. Task 2 is a task on semantic similarity between clinical cases and discussions. For this task, we propose an approach based on language models and evaluate the impact on the results of different preprocessings and matching techniques. For task 3, we have developed an information extraction system yielding very encouraging results accuracy-wise. We have experimented two different approaches, one based on the exclusive use of neural networks, the other based on a linguistic analysis.

MOTS-CLÉS : Similarité sémantique, extraction d'information, modèle de langues, modèle de vraisemblance de la requête, réseaux de neurones, analyse syntaxique.

KEYWORDS: Semantic similarity, information extraction, language model, query likelihood model, neural network, syntactic analysis.

1 Introduction

L'analyse et l'extraction d'information pertinentes au sein d'un corpus médical est une tâche qui peut se montrer particulièrement difficile en raison de l'extrême spécificité du domaine. L'édition 2019 du défi fouille de texte (DEFT) porte sur cette problématique (Grabar *et al.*, 2019), et met à disposition un corpus de cas cliniques français, eux-mêmes issu du corpus CAS (Grabar *et al.*, 2018).

Nos motivations pour participer cette année au DEFT sont multiples. L'accent sur l'aspect médical de

l'édition 2019 est particulièrement stimulant, du fait de l'impact médical que peuvent permettre les avancées de recherche dans ce domaine. De plus, la complexité du travail dans le domaine médical dû à sa spécificité, ses problématiques d'accès et la taille restreinte de ressources associées en font un défi particulièrement intéressant. Nous voulions également avoir la possibilité de nous confronter à d'autres équipes réfléchissant à des problématiques similaires. Enfin, il nous tient à coeur de pouvoir contribuer à la recherche dans le domaine français du traitement automatique des langues.

Nous avons participé à deux des trois tâches proposées dans le cadre de la campagne. La tâche 2, portant sur l'appareillement de cas cliniques et de discussions, est détaillée dans la Section 2. Nous décrivons ensuite dans la Section 3 notre contribution à la tâche 3, dont le but est d'extraire des informations démographiques et cliniques sur les cas cliniques.

2 Tâche 2 : Mise en correspondance des cas cliniques et discussions par vraisemblance de la requête

L'objectif de la tâche 2 est de faire un appariement entre un cas clinique et la discussion correspondante. Le corpus d'entraînement contient 290 discussions et 290 cas cliniques. Tandis que chaque cas clinique est unique, plusieurs discussions sont identiques. Le corpus de test contient 214 discussions et cas cliniques, présentant les mêmes caractéristiques que le corpus d'entraînement.

2.1 Approches

L'approche utilisée dans le cadre de la tâche 2 est la même que celle proposée par Ponte & Croft (1998) pour le calcul de similarité basé sur des modèles de langue.

L'idée principale est de générer un modèle de langue par discussion et de mesurer leur proximité avec chacun des cas cliniques. Le plus proche étant celui qui sera appareillé. La mesure utilisée est la perplexité calculée suivant l'équation 1 :

$$PPL = \hat{P}(w_1, ..., w_m)^{-\frac{1}{m}} \tag{1}$$

L'approche étant basé sur la forme de surface des mots, nous avons appliqués différents pré-traitements et étudié l'impact de ces derniers sur les résultats. Nous avons également étudié l'effet de différentes méthodes d'appariement des données.

2.2 Pré-traitement des données

Les processus de pré-traitements ont été appliquées de la même manière aux cas cliniques et aux discussions. Le texte est tout d'abord systématiquement converti en minuscules et tokenisé à l'aide de notre outil interne Qnlp-toolkit [1]. Différentes approches supplémentaires ont également été appliquées en fonction des expériences : la racinisation, la suppression de mots vides et le désabrègement.

1. https ://github.com/QwantResearch/qnlp-toolkit

Racinisation Afin de ne pas désavantager au sein des modèles de langue des mots partageant la même racine mais différant dans leur forme finale, nous avons racinisé dans cette approche l'ensemble du corpus, utilisant l'algorithme de Snowball (Porter, 2001).

Suppression des mots vides Les mots très courants (souvent appelés « mots vides »), tels que « le », « et » ou « de », sont généralement ignorés dans les recherches, car ils ne contiennent habituellement pas autant d'information que les autres mots recherchés permettant l'appariement des cas et discussions Par exemple, une situation dans laquelle conserver les mots vides est importante est la correspondance de l'information stylistique des cas et discussions (si les cas et discussions matchant avaient systématiquement été rédigés par la même personne).

Désabrègement Compte tenu de l'abondance des sigles, acronymes, symboles et autres abrègements rencontrés dans les cas cliniques et dans les discussions, il nous a semblé intéressant de procéder à leur désabrègement automatique. Dans le domaine médical l'abrègement est un procédé de construction lexicale très utilisé pour des raisons mnémotechnique et d'économie du langage qui permet de réduire les longues compositions de mots (souvent savants).

Il arrive fréquemment que ces abrègements soient définis et repris tout au long du texte. Afin de lever toute ambiguïté lexicale en écartant au maximum les cas d'abrègements polysémiques (e.g. IVG peut à la fois être utilisé pour signifier « insuffisance ventriculaire gauche » ou « interruption volontaire de grossesse »), nous avons constitué un lexique des formes étendues, élaboré à partir des sigles, acronymes et de leurs définitions rencontrés en contextes à partir de l'ensemble du corpus. La mise en correspondance de la forme développée et de son abrègement a été réalisé selon plusieurs règles de recherche et nous avons compléter notre lexique initial par des ressources terminologiques recensant les abrègements reconnus par la communauté.

Ainsi à partir du corpus comprenant les cas cliniques et les discussions s'y rapportant, 227 abrègements ont été relevés, ce qui a engendré un nombre de substitutions égal à 9016.

2.3 Appariements

Les appariements peuvent se faire de différentes manières. Nous avons choisi deux manières de faire : appareiller les cas cliniques avec les discussions et appareiller les discussions aux cas cliniques (notées respectivement *c2d* et *d2c*). Dans première méthode, les modèles de langue sont entraînés sur les cas cliniques, et nous estimons la perplexité de chaque modèle sur les discussions. Dans la seconde méthode, nous faisons l'inverse : nous entraînons les modèles de langue sur les discussions et nous estimons les scores de perplexité sur chacun des cas cliniques.

Nous avons également testé deux techniques permettant le choix des meilleures paires en fonction du score de perplexité. La première (« non-exclusive ») consiste à choisir indépendemment et pour chaque LM le texte avec la perplexité la plus basse. Cela signifie que le même texte peut être appareillé à plusieurs LMs. Dans la seconde technique (« exclusive »), un cas ne peut être appareillé qu'une seule fois avec une discussion (et vice-versa). Pour ce faire, nous choisissons de façon itérative la paire de cas/discussion générant la perplexité la plus faible sur toutes les paires possibles, avant de supprimer ces cas et discussions de la liste de choix futurs. Cela exige que les scores de perplexités soient comparable, ce qui est le cas ici car le même vocabulaire a été utilisé pour générer tous les

modèles de langue.

2.4 Apprentissage des modèles de langue

Les différents modèles de langue ont été appris à l'aide de l'outil *SRILM* (Stolcke, 2002), selon trois ordres différents : unigrammes, bigrammes et trigrammes. Un vocabulaire unique a été conservé pour les différents modèles en conservant les N mots les plus fréquents du corpus d'apprentissage. Afin de permettre une comparaison entre les différentes perplexités (voir Section 2.3), tous les modèles ont été créés utilisant un vocabulaire commun, correspondant à l'ensemble des mots existants dans le corpus (cas cliniques et discussions).

2.5 Expériences & Discussions

Nous avons testé différentes combinaisons des techniques de pré-traitement et d'appareillement présentées ci-dessus. Toutes les expériences ont été effectuées sur l'entièreté du corpus d'apprentissage, soit 290 paires de discussions et cas cliniques. Les scores obtenus sur le corpus d'évaluation, sont également présentées.

2.5.1 Effet du pré-traitement de texte

Les résultats présentés dans le tableau 1 mettent en exergue l'effet de différentes techniques de pré-traitement sur les corpus de cas et discussions. Dans ce tableau, nous avons comparé uniquement les différentes approches en fonction de la méthode d'appariement *d2c* décrite dans la section 2.3. La racinisation est notée *rac*, la suppression des mots-vides *mv* et le désabrègement *des*.

Le système initial qui ne comporte aucun pré-traitement atteind un score de 61,38 en répcision et rappel. Les pré-traitement classiques de racinisation et de suppression de mots-vides améliorent logiquement les scores de près de 11 points. le processus de désabrègement automatique, seul, offre des scores de précision et de rappel de 80,69, soit près de 19 points d'amélioration. Lorsqu'on combine les deux pré-traitements, malheureusement, les améliorations ne se cumulent pas. Au contraire, on observe une légère contre-performance de 0,7 points par rapport au meilleur système.

Pré-traitement	Apprentissage	
	Pr	Rp
Initial	61,38	61,38
rac+mv	72,41	72,41
des	**80,69**	**80,69**
rac+mv+des	80,00	80,00

TABLE 1 – Scores de précision et rappel pour les données de la tâche 2, en fonction du pré-traitement testé. Tous les LMs sont d'ordre 2, et l'appariement s'est fait de façon exclusive, en direction d2c. *rac : racinisation ; mv : mots vides supprimés ; des : désabbréviation.*

Ordre	Pr	Rp
1-gram	77,24	77,24
2-gram	**80,68**	**80,68**
3-gram	79,31	79,31

TABLE 2 – Scores de précision et rappel pour les données de la tâche 2, en fonction de l'ordre du LM. L'appariement s'est fait de façon exclusive, en direction d2c. Le corpus d'apprentissage a été pré-traité uniquement avec les abbréviations normalisées (des).

2.5.2 Effet de l'ordre du modèle de langue

Utilisant des modèles de langue nous nous sommes intéressés à l'impact de l'ordre de ces derniers. Le tableau 2 présente les résultats obtenus. Nous avons utilisé une configuration *d2c* et avec le pré-traitement de désabréviation. On peut constater que les modèles de langue d'ordre 2 et 3 obtiennent de meilleurs résultats que les modèles d'ordre 1 (près de 3 points de mieux).

2.5.3 Impact des techniques d'appariement

Nous avons également testé les différentes techniques d'appariement introduits dans la Section 2.3. La Table 3 souligne l'amélioration apportée par la technique d'exclusivité, avec un score de précision et de rappel systématiquement plus haut que pour les mêmes systèmes n'utilisant pas cette technique. En effet, sans ce procédé, il est probable que si un cas est relativement général dans les termes qui le compose, il soit appareillé à une grande majorité de discussions (ou vice-versa). Puisque la tâche 2 nécessite qu'un texte ne soit appareillé qu'une seule fois, nous avons donc choisi d'utiliser cette technique dans nos soumissions.

L'importance de la direction utilisée pour l'appariement (*d2c* ou *c2d* - voir Section 2.3), est aussi mise en exergue dans les résultats présentés Table 3. Il semble ainsi que mesurer la proximité de modèles de langue estimés sur les discussions par rapport à chacun des cas cliniques (d2c) produise des résultats plus probants que dans le cas inverse. Les résultats peuvent s'expliquer par la plus grande longueur des discussions par rapport aux cas (environ 393 mots en moyenne pour les cas versus 919 mots pour les discussions), les discussions permettant ainsi d'estimer des modèles de langue plus variés. Une expérience intéressante pour le futur serait cependant de combiner les deux approches, et sélectionner les meilleurs scores sur toutes les paires possibles, avec les deux directions c2d et d2c.

Direction	NE/E	Pr	Rp
c2d	NE	40,34	40,34
c2d	E	71,03	71,03
d2c	NE	48,96	48,96
d2c	E	**80,69**	**80,69**

TABLE 3 – Scores de précision et rappel pour les données de la tâche 2, en fonction de la méthode d'appariement. L'appariement s'est fait de façon exclusive (E) ou non-exclusive (NE), dans les deux directions (*d2c* et *c2d*). Le corpus d'apprentissage a été pré-traité uniquement avec les abbréviations normalisées (des).

2.5.4 Résultats soumis

La Table 4 récapitule les résultats obtenus sur les trois contributions que nous avons soumis pour la tâche 2 de DEFT 2019. En accord avec les conclusions tirées Section 2.5.3, nous utilisons l'approche d2c et la technique d'exclusion pour les trois soumissions.

La première version testée (*run-1*) correspond aux meilleurs résultats obtenus lors de tous nos essais sur le corpus d'apprentissage. Le texte a été racinisé, désabrégé et les mots vides supprimés. Les modèles de langue sont d'ordre 1, et ont été crées sur les discussions (direction *d2c*). Pour la deuxième soumission (*run-2*), nous avons choisi l'approche qui serait théoriquement la meilleur en se basant exclusivement sur les conclusions tirées lors des expériences sur le corpus d'apprentissage. Nous avons ainsi utilisé un modèle de langue d'ordre 2 et le seul pré-traitement effectué est le désabrègement. Enfin, la troisième version (*run-3*) est plus expérimentale. Nous avons en effet décidé de jouer sur l'ordre du modèle de langue (choisissant un modèle d'ordre 3, et d'utiliser outre cette variable les mêmes caractéristiques que pour la version 1 (désabrègement, racinisation, suppression des mots vides, direction *d2c* et technique d'exclusion).

Version	Ordre	Pré-traitement	Apprentissage		Évaluation	
			Pr	Rp	Pr	Rp
run-1	1-gram	rac+mv+des	**81,72**	**81,72**	**80,84**	**80,84**
run-2	2-gram	des	80,69	80,69	71,03	71,03
run-3	3-gram	rac+mv+des	80,34	80,34	79,91	79,91

TABLE 4 – Scores de précision et rappel pour les résultats soumis pour la tâche 2. L'appariement s'est systématiquement fait de façon exclusive, en direction *d2c*.

Les approches 1 et 3 produisent les résultats attendus sur les données de test, proches de ceux obtenus sur le corpus d'entraînement. La seconde version (run-2) cependant, produit sur le corpus de test des résultats bien plus bas que ceux du corpus d'apprentissage. Puisque nous savons que l'impact de l'ordre des modèles de langue est restreint, il est probable que ces résultats viennent du pré-traitement choisi. Cela peut souligner la possible importance de la racinisation et de la suppression de mots-vides lors de travail sur des corpus de taille restreinte.

Il est également intéressant d'observer que l'approche donnant lieu aux meilleurs résultats utilise des modèles de langue d'ordre 1. Ce type de modèle de langue (ou modèle « sac de mots »), qui ne prend en compte que la fréquence des mots, ignorant leur ordre, peut donc suffire pour ce type d'exercice. En effet, la taille extrêmement restreinte des cas et discussions sur lesquelles les modèles de langues ont été estimés ne permettent aucun gain d'information en utilisant des modèles d'ordre plus élevés.

3 Tâche 3 : Extraction d'information sur des cas cliniques

La tâche 3 est une tâche d'extraction d'information de type démographique et médicale. Ses objectifs concernent l'identification de cinq types d'informations correspondant au moment du dernier élément clinique rapporté dans le cas clinique : l'âge et le genre de la personne dont le cas est décrit, l'origine (motif de la consultation ou de l'hospitalisation) et l'issue parmi cinq valeurs possibles (guérison, amélioration, stable, détérioration, décès). Pour tous ces cas, il est possible qu'une ou plusieurs des

informations soient manquantes. Dans cette situation, la valeur est 'NUL'.

3.1 Approches

Nous présentons dans la suite les deux principales approches utilisées pour tenter de résoudre cette tâche : une approche neuronale et une approche hybride qui se fonde sur une analyse linguistique du corpus. L'évaluation de la pertinence des différentes variantes proposées sera ensuite donnée.

3.1.1 Pré-traitements des données

La tokénisation et la suppression de la casse ont été réalisées à l'aide de notre outil interne Qnlp-toolkit [2].

La lemmatisation permet d'obtenir la forme canonique des mots. Elle trouve son intérêt dans le cadre de ce travail car elle permet de débarrasser des mots les marques d'inflexion telles que celles de genre (masculin, féminin), de pluriel ou de conjugaison. Les mots ou plus précisément les chaînes de caractères peuvent ainsi être comparés à un niveau plus fin. La lemmatisation d'un verbe est la forme à l'infinitif de ce verbe, celle d'un nom, adjectif ou déterminant, sa forme au masculin singulier. Elle est effectuée en utilisant des règles de correspondance à partir des données de WordNet (Fellbaum, 1998).

La lemmatisation a été appliquée uniquement dans le cadre de l'approche hybride. En effet, les modèles neuronaux, se basent sur des représentations distribuées des mots et nécessitent donc peu de pré-traitement. En général une tokénisation suffit.

3.1.2 Approche neuronale

Nous avons utilisé un modèle neuronal supervisé pour l'étiquetage d'empan du cas clinique avec une segmentation BIO, dans le but d'étiqueter le texte relatif aux informations à extraire. L'approche décrite par Ma & Hovy (2016) fut choisie car ce modèle nécessite un nombre de données moindre pour obtenir des résultats à l'état-de-l'art en étiquetage morpho-syntaxique et en reconnaissances d'entités nommées. Les résultats sont rendus possible grâce à l'utilisation de représentations pré-entraînées de mots et de caractères ainsi qu'à la combinaison d'un réseau de neurones récurrent (*Bi-directional Long-Short Term Memory*, Bi-LSTM), un réseau de neurones convolutionnel (CNN) et un champ markovien conditionnel (CRF) tels que présentés dans la figure 1.

Pour créer le corpus d'apprentissage, nous avons manuellement étiqueté les 290 cas cliniques selon 4 types d'étiquette : Age, Genre, Admission et Issue.

Dans le but d'améliorer les performances du modèle et pour palier au faible nombre de données, nous avons généré automatiquement des introductions de cas typiques. Cette génération fut opérée grâce à une grammaire hors-contexte ainsi qu'un certain nombre de ressources linguistiques, telles que des listes de symptômes et de maladies. Ainsi les cas suivants ont pu être générés :

— *mlle a a été mis sous observations suite à hématome temporo-pariétal post-traumatique* ;
— *mme u , 10 ans , a été traité pour une cancer épidermoïde* ;

2. https ://github.com/QwantResearch/qnlp-toolkit

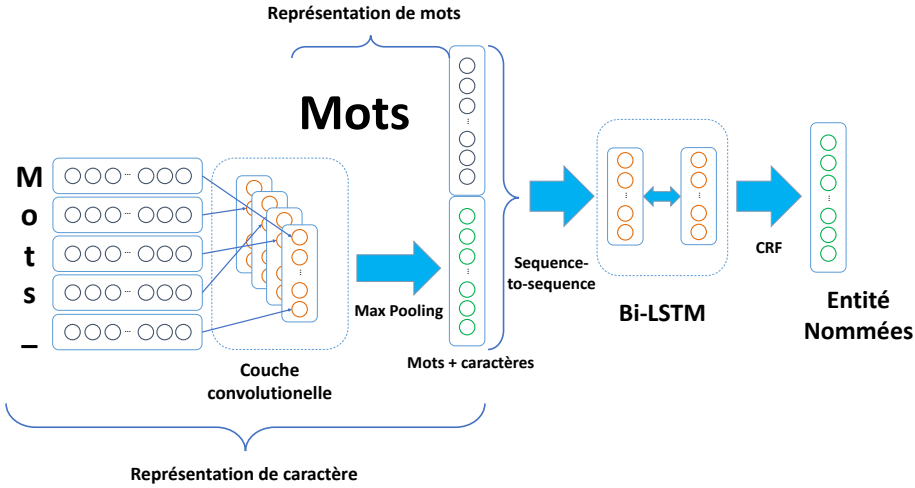

FIGURE 1 – Modèle neuronal utilisé pour l'extraction d'entités nommées fondé sur l'approche proposée par Ma & Hovy (2016).

— *un jeune homme agé de 20 ans avec comme antécédent un perforation digestive instrumentale , a été mis sous observations suite à une surinfection kt jugulaire*.

En parallèle, nous avons appris des représentations de mots sur des corpus plus larges pour les utiliser dans le modèle. Nous avons extrait l'ensemble des pages du Wikipédia français appartenant au portail de la Médecine [3] ainsi que le corpus EMEA contenant des documents PDF de l'agence européenne de Médecine [4]. Sur ce corpus aggrégé, nous avons appris des représentations de mots en utilisant le modèle d'apprentissage de représentations de mots de FastText [5] proposé par Bojanowski *et al.* (2017).

Après l'étiquetage du texte, nous inférerons les informations demandées, notamment concernant l'âge, le genre, ainsi que l'issue. En effet, une fois l'obtention d'un empan, nous devons déduire l'âge en années. A partir d'un certain nombre de règles heuristiques on infère la valeur pour un empan tel que *"18 mois"*. On obtient alors 1 an. De la même manière, un certain nombre d'heuristiques ont été utilisées pour inférer le genre du patient à partir de l'empan correspondant. Concernant l'admission, aucune modification n'a été appliquée à l'empan retourné par le modèle.

L'identification de l'issue est quant à elle, vu comme un problème de classification multi-classes où l'on considère les cinq valeurs possibles de l'issue (guérison, amélioration, stable, détérioration, décès) plus une sixième, utilisée pour nous permettre de considérer les cas où la valeur est NUL. Cette classification repose un modèle neuronal proposé par Joulin *et al.* (2017) réalisée avec fastText [6].

3. Le portail médecine regroupe les articles appartenant au domaine médical https://fr.wikipedia.org/wiki/Portail:M%C3%A9decine.
4. Le corpus EMEA est accessible à l'adresse http://opus.nlpl.eu/EMEA.php.
5. FastText : https://fasttext.cc
6. fastText : https://fasttext.cc/

3.1.3 Approche hybride : modèles neuronales et analyses linguistiques

Une seconde approche, hybride, est utilisée afin d'extraire les unités lexicales correspondant aux âges, genres et origine recherchés, à partir de l'analyse syntaxique en dépendance des cas cliniques et l'utilisation de patrons lexico-syntaxiques prédéfinis.

Pour déterminer l'origine, l'analyse syntaxique de la phrase dans laquelle elle se trouvait nous a permis de repérer et d'extraire un ensemble de lexies verbales apparaissant de façon récurrente et se révélant toutes concerner la tâche de prise en charge (présenter, hospitaliser, admettre, consulter, adresser, etc.). A partir de cette observation, nous avons procéder à l'extraction des compléments verbaux en position postverbale (syntagmes nominaux et syntagmes prépositionnels) entretenant une relation directe avec les verbes récurrents identifiés. L'âge et l'origine sont extraits de manière assez similaire. On observe par exemple 148 occurrences de l'adjectif '[âla]gé'. A partir de son noeud dans l'arbre, il est aisé d'extraire le syntagme nominal dépendant dont la fonction est sujet. L'identification du genre peut par la suite être réalisée en se référant aux valeurs des traits morpho-syntaxiques (features) de ses constituants. L'âge correspond au numéral dépendant indirect qu'il reste à extraire.

En ce qui concerne les issues, leurs identifications s'effectuent en utilisant le même classifieur que celui présenté précédemment. Un traitement spécifique est réalisé en amont pour identifier les cas de décès à partir d'une liste des lexies spécifiques au champ sémantique lié au décès, dans leurs versions lemmatisées.

Pour déterminer l'âge et le genre à partir du dernier événement clinique ayant motivé la prise en charge, nous avons tout d'abord observé leurs fréquences d'apparitions dans le texte. On a ainsi observé que l'âge et le genre apparaissait en moyenne à la Xème phrase (age :X, genre :X) tandis que l'origine apparaissait en moyenne à la 3,8ème phrase.

3.2 Expériences & Discussions

3.2.1 Corpus d'apprentissage

En premier lieu, nous présentons les résultats des différentes approches sur le corpus d'entraînement afin de comparer et sélectionner les approches les plus prometteuses. Le corpus d'apprentissage comprend 290 cas cliniques.

Étiquetage automatique Nous présentons tout d'abord les résultats de l'étiquetage automatique du texte. Nous avons retranché dix pour-cent de cas pour évaluation. La table 5 présente les scores F-measure pour chacun des types d'étiquette. Trois cas sont présentés, tout d'abord un apprentissage du modèle sur les données d'apprentissage uniquement. Puis l'utilisation de représentations de mots pré-entraînées sur des corpus de données médicales (PRE). Et enfin l'ajout de données générées à l'aide d'une grammaire hors-contexte (GEN).

La combinaison la plus efficace est celle qui regroupe utilisation des représentations de mots pré-entraînées et données générées automatiquement. Le nombre de données du corpus d'apprentissage étant relativement restreint, tout ajout de connaissances extérieures permet d'augmenter sensiblement les performances du modèle.

Nous décidons donc de conserver le modèle avec utilisation de représentations de mots pré-entraînées

Modèle / F-measure	Age	Genre	Issue	Admission	All
Etiquetteur	84.21	50.00	51.72	28.57	53.81
Etiquetteur + PRE	87.72	59.26	47.27	36.73	58.60
Etiquetteur + GEN + PRE	90.00	53.33	58.62	43.64	61.80

TABLE 5 – Scores de F-measure sur l'étiquetage automatique de la tâche 3 pour un modèle appris sur 90% du corpus d'entraînement et tester sur le reste. PRE désigne l'utilisation de représentations de mots pré-entraînées sur des corpus de données médicales. GEN correspond à l'extension des données par des données générées automatiquement à l'aide d'une grammaire hors-contexte.

ainsi que les données générées automatiquement.

Classification de l'issue Les résultats de la classification de l'issue sont ensuite présentés dans la table 6. L'évaluation s'est faite par validation croisée avec 10 plis. Dans notre cas, nous considérons le résultat NUL comme un label comme les autres.

Portion du cas clinique	Précision	Rappel	F-Measure
Cas clinique entier	0.4482	0.4459	0.4471
Deux dernières phrases du cas clinique	0.4448	0.4448	0.4448
Empan de l'issue obtenu par étiquetage manuel	0.6215	0.6215	0.6215
Empan de l'issue obtenu par étiquetage automatique	0.4222	0.4222	0.4222

TABLE 6 – Scores de précision, rappel et F-measure sur l'issue en validation croisée. On compare les résultats sur différentes portions de texte du cas clinique. On considère l'entièreté du cas clinique ainsi que les deux dernières phrases du cas clinique.

On observe que le meilleur score est celui obtenu en prédisant l'issue uniquement sur le texte s'y référant. Malheureusement, lorsque l'empan est obtenu par étiquetage automatique et non pas manuel, la qualité est grandement dégradée. Cela est dû à la faible qualité de l'étiquetage automatique de l'issue présenté dans le paragraphe précédent.

Pour la phase de test, on décide de conserver le cas où l'on considère le cas clinique en entier ainsi que celui basé sur les empans sélectionnés seulement.

3.2.2 Corpus de test

Après avoir choisi les systèmes les plus prometteurs à partir des résultats obtenus sur le corpus d'entraînement, nous avons pu évaluer nos approches bout-à-bout sur le corpus de test.

Age et genre Les résultats concernant l'âge et le genre sont présentés dans la table 7. L'approche neuronale d'étiquetage de mots et l'approche par arbre syntaxique sont comparées. On observe que l'approche par étiquetage neuronal fonctionne mieux que l'approche lexico-syntaxique. En effet, malgré le faible nombre de données à l'origine (290 données d'apprentissage), l'extension des cas

Approche	Age		Genre	
	Précision	Rappel	Précision	Rappel
Etiquetteur + PRE + GEN (*run 1*)	**0.9748**	**0.9023**	**0.9421**	**0.9465**
Analyse lexico-syntaxique	0.9709	0.8558	0.8969	0.8906

TABLE 7 – Scores de précision et rappel pour l'extraction de l'âge et du genre sur le corpus de test. Etiquetteur + PRE + GEN correspond au système envoyé au soumission en première tentative (*run 1*). L'analyse lexico-syntaxique utilise une approche par arbre syntaxique pour extraire les informations.

cliniques par génération automatique ainsi que le recours aux représentations de mots pré-entraînées permettent d'obtenir une bonne généralisation, et donc des résultats satisfaisants pour l'étiquetage automatique.

Approche	macro Précision	macro Rappel	micro Précision	micro Rappel	*micro overlap accuracy*
Etiquetteur + PRE + GEN (*run 1*)	0.7850	0.5786	0.6582	0.6398	0.5889

TABLE 8 – Scores de micro et macro précision et rappel et *micro overlap accuracy* pour l'extraction de l'admission sur le corpus de test à partir d'un système étiqueteur neuronal avec représentations de mots pré-entraînées et génération de données d'apprentissage.

Admission Les résultats relatifs à l'admission sont présentés dans la table 8. Les scores issus de l'étiqueteur neuronal sont prometteurs et les différences entre précision et rappel (macro et micro) laissent supposer que le modèle retourne un empan de texte trop précis. Nous n'avons pas pu obtenir de résultats concluants par analyse lexico-syntaxique car, contrairement à l'âge et au genre, les cas cliniques ne suivent pas un schéma suffisamment récurrent pour obtenir une bonne extraction.

Approche	Issue	
	Précision	Rappel
Cas clinique entier	0.5285	0.5199
Empan de l'issue obtenue par étiquetage (*run 1*)	0.5198	0.4918
Traitement linguistique de *décès* et 4 dernières phrases	**0.5985**	**0.5831**

TABLE 9 – Scores de précision et rappel pour l'issue.

Issue Enfin, les résultats de l'issue sont présentés dans la table 9. Dans les deux premiers cas, nous avons utilisé uniquement un modèle de classification pour prédire l'ensemble des classes. Nous avons tout d'abord évalué un modèle apprenant et prédisant sur l'entièreté du cas clinique. Nous avons comparé les résultats avec l'apprentissage et la prédiction du modèle sur les données étiquetées automatiquement. Enfin, nous avons utilisé un traitement linguistique pour extraire les cas de décès et appris un modèle pour les issues restantes sur les 4 dernières phrases des cas cliniques. Après

annotation manuelle du corpus d'entraînement pour étiquetage, nous avions en effet fait plusieurs observations relatives à l'issue. D'une part, les cas de décès se prêtaient mieux à une approche lexicale avec un vocabulaire très spécifique. Et d'autre part, dans la majorité des cas, l'empan de texte renvoyant à l'issue apparaissait vers la fin du cas clinique. Cette dernière approche est celle qui retourne les meilleurs résultats avec un score de 0.60 et 0.58 en précision et rappel respectivement.

4 Conclusion

Dans cet article, nous avons présenté notre participation aux tâches 2 et 3 de l'évaluation du DEFT 2019, respectivement une tâche de similarité sémantique et une tâche d'extraction d'information.

Pour la tâche 2, nous avons présenté une méthode de similarité sémantique à base de modèles de langues. Nous avons évalué l'impact des pré-traitements et des techniques d'appareillement sur les résultats. Nous avons réussi à obtenir des résultats prometteurs en se restreignant uniquement aux données fournies.

Concernant la tâche 3, nous avons présenté deux approches différentes. La première est une approche neuronale avec étiquetage automatique du cas clinique. Nous avons utilisé des données extérieures pour améliorer les résultats de cette approche. La seconde, quant à elle, repose sur une approche linguistique d'arbre syntaxique pour l'extraction d'information. Dans les deux cas, une phase de classification a été appliquée concernant l'issue.

Références

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

FELLBAUM C. (1998). *WordNet : An electronic lexical database*. MIT Press.

GRABAR N., CLAVEAU V. & DALLOUX C. (2018). Cas : French corpus with clinical cases. In *LOUHI 2018 : The Ninth International Workshop on Health Text Mining and Information Analysis*.

GRABAR N., GROUIN C., HAMON T. & CLAVEAU V. (2019). Recherche et extraction d'information dans des cas cliniques. présentation de la campagne d'évaluation deft 2019. In *Actes de DEFT*.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431, Valencia, Spain : Association for Computational Linguistics.

MA X. & HOVY E. (2016). End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, p. 1064–1074.

PONTE J. M. & CROFT W. B. (1998). A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, p. 275–281.

PORTER M. F. (2001). Snowball : A language for stemming algorithms.

STOLCKE A. (2002). Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.