HAL open science

Apprentissage de plongements de mots sur des corpus en langue de spécialité : une étude d'impact

Valentin Pelloin, Thibault Prouteau

To cite this version:

Valentin Pelloin, Thibault Prouteau. Apprentissage de plongements de mots sur des corpus en langue de spécialité : une étude d'impact. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), 2020, Nancy, France. pp.164-178. hal-02786198v1

HAL Id: hal-02786198

https://hal.science/hal-02786198v1

Submitted on 7 Jun 2020 (v1), last revised 2 Feb 2021 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Apprentissage de plongements de mots sur des corpus en langue de spécialité : une étude d'impact

Valentin Pelloin Thibault Prouteau
LIUM, Avenue Laennec, 72085 LE MANS, France
valentin.pelloin.etu@univ-lemans.fr,
thibault.prouteau.etu@univ-lemans.fr

RÉSUMÉ

Les méthodes d'apprentissage de plongements lexicaux constituent désormais l'état de l'art pour la représentation du vocabulaire et des documents sous forme de vecteurs dans de nombreuses tâches de Traitement Automatique du Langage Naturel (TALN). Dans ce travail, nous considérons l'apprentissage et l'usage de plongements lexicaux dans le cadre de corpus en langue de spécialité de petite taille. En particulier, nous souhaitons savoir si dans ce cadre, il est préférable d'utiliser des plongements préappris sur des corpus très volumineux tels Wikipédia ou bien s'il est préférable d'apprendre des plongements sur ces corpus en langue de spécialité. Pour répondre à cette question, nous considérons deux corpus en langue de spécialité : OHSUMED issu du domaine médical, et un corpus de documentation technique, propriété de SNCF. Après avoir introduit ces corpus et évalué leur spécificité, nous définissons une tâche de classification. Pour cette tâche, nous choisissons d'utiliser en entrée d'un classifieur neuronal des représentations des documents qui sont soit basées sur des plongements appris sur les corpus de spécialité, soit sur des plongements appris sur Wikipédia. Notre analyse montre que les plongements appris sur Wikipédia fournissent de très bons résultats. Ceux-ci peuvent être utilisés comme une référence fiable, même si dans le cas d'OHSUMED, il vaut mieux apprendre des plongements sur ce même corpus. La discussion des résultats se fait en interrogeant les spécificités des deux corpus, mais ne permet pas d'établir clairement dans quels cas apprendre des plongements spécifiques au corpus.

ABSTRACT

Learning word embeddings on domain specific corpora : an impact study

Word embedding approaches are state of the art in Natural Language Processing (NLP). In this work, we focus on learning word embeddings for small domain-specific corpora. In particular, we would like to know whether word embeddings learnt over large corpora such as Wikipedia perform better than word embeddings learnt on domain specific corpora. In order to answer this question, we consider two corpora : OHSUMED from the medical field, and SNCF, a technical documentation corpus. After presenting the corpora and evaluating their specificity, we introduce a classification task. We use word embeddings learnt on domain-specific corpora or Wikipedia as input for this task. Our analysis demonstrates that word embeddings learnt on Wikipedia achieve excellent results, even though, in the case of OHSUMED, domain specific word embeddings perform better.

MOTS-CLÉS : langue de spécialité, plongements de mots, catégorisation de documents.

KEYWORDS: domain specific, word embeddings, documents categorization.

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 164–178.
HAL : hal-02786198.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Les approches contemporaines en traitement automatique des langues privilégient une représentation des mots au travers de *plongements de mots* (ou *plongements lexicaux*) décrits notamment par Mikolov *et al.* (2013) et Bojanowski *et al.* (2017). Les plongements lexicaux sont des représentations vectorielles compactes des mots, qui encapsulent les liens sémantiques et syntaxiques présents dans les textes des corpus sur lesquels ils ont été appris. L'apprentissage de plongements de mots s'effectue le plus souvent sur un corpus comportant une grande quantité de documents en langue générale comme Wikipédia (Huang *et al.*, 2012; Liu *et al.*, 2015; Yamada *et al.*, 2018).

Dans notre cas, nous nous concentrons sur la langue de spécialité. La langue de spécialité est celle qui est utilisée pour rendre compte de connaissances sur un sujet ou une discipline spécialisée (Charnock, 1999; Van Der Yeught, 2016; Schmitt, 2002; Condamines, 1997; Chujo & Utiyama, 2006; Paltridge & Starfield, 2016). Par conséquent, le lexique employé est la plupart du temps imposé par la discipline abordée et le vocabulaire employé diffère donc de celui que l'on retrouve dans un corpus non spécialisé. Ainsi, il semble nécessaire, pour représenter efficacement le vocabulaire et les documents de tels corpus, d'apprendre les plongements lexicaux sur ces mêmes corpus. Cependant, la taille de ces corpus est très nettement inférieure à celle de corpus comme Wikipédia, ce qui peut nuire à l'apprentissage des plongements qui repose sur les cooccurrences du vocabulaire. Or, plus le nombre de cooccurrences est faible, moins l'algorithme dispose d'informations pour apprendre une représentation du lexique. De ce fait, on cherche à savoir quelle représentation adopter lorsque l'on traite des données en langue de spécialité : est-il préférable d'utiliser des plongements lexicaux appris sur une grande quantité de données non spécialisés ou sur une faible quantité de données spécialisées ?

Plusieurs études s'intéressent à l'apprentissage de plongements lexicaux sur un corpus spécialisé du domaine médical (El Boukkouri *et al.*, 2019; Lee *et al.*, 2020) en adaptant des modèles appris sur des données en langue générale au domaine médical. Wang *et al.* (2018) étudient l'apprentissage de plongements lexicaux sur des corpus médicaux et concluent que les plongements lexicaux appris sur un corpus spécialisé capturent mieux les liens sémantiques et syntaxiques que ceux appris sur un corpus non *spécialisé*. Cet article s'inscrit dans la continuité de ces travaux en proposant une étude sur deux corpus en langue de spécialité : OHSUMED, un corpus issu du domaine médical, et SNCF, un corpus *ad hoc* de documents techniques. À la différence des approches proposées par (El Boukkouri *et al.*, 2019; Lee *et al.*, 2020) nous décidons de comparer les performances des modèles appris sur des données en langue générale et spécialisée ainsi qu'une spécialisation aussi bien à une tâche de classification de document qu'au domaine spécialisé.

Dans un premier temps, nous tentons d'estimer la spécificité de ces deux corpus automatiquement, à travers plusieurs indicateurs extraits de la littérature et décrits Section 2.1. Ensuite, Section 2.2, nous détaillons le protocole d'apprentissage des plongements de mots sur les corpus spécialisés OHSUMED et SNCF, et sur les corpus non spécialisés. Dans cette même section, nous présentons le protocole mis en oeuvre pour évaluer les performances des différents plongements appris. Ce protocole repose sur une tâche de classification en classes multiples par un réseau neuronal convolutif. Puis, nous décrivons les corpus OHSUMED et SNCF Section 3. Enfin, nous détaillons Section 4 les résultats avant de conclure et de discuter des perspectives de ce travail en cours.

2 Méthodologie

L'objectif de ce travail est de savoir, si utiliser des plongements de mots préappris sur des corpus très volumineux tels que Wikipédia est préférable à l'utilisation de plongements lexicaux appris sur nos

corpus en langue de spécialité.

Avant de détailler nos expérimentations sur les plongements lexicaux dans le cadre d'une utilisation sur un corpus spécialisé, nous commençons par définir ce qu'est un corpus spécialisé et présentons quelques indicateurs de la spécialité d'un corpus.

2.1 Estimer la spécificité

Le contenu des textes d'un corpus spécialisé correspond à un domaine particulier et s'adresse aux initiés dudit domaine. De plus, le domaine impose le vocabulaire et le style employé (Charnock, 1999; Van Der Yeught, 2016).

Analyse du destinataire. Les corpus spécialisés ont la particularité d'être à destination d'un groupe limité d'individus. Dans le cas du corpus SNCF les documents sont destinés aux agents de l'entreprise. Concernant le corpus OHSUMED, les documents sont des *abstracts* d'articles scientifiques sur des maladies, ceux-ci sont destinés à des médecins et chercheurs en médecine.
Cette observation faite, on peut s'interroger sur le niveau d'expertise de la cible et son impact sur le lexique employé. Les documents à destination d'un public expert contiennent-ils plus de termes apparentés au jargon que ceux destinés à des néophytes ? Nous présentons ci-après plusieurs indicateurs sur le lexique.

Analyse du lexique. D'après Cressot & James (1996), le style d'un corpus peut se résumer aux choix réalisés par les auteurs lors de l'écriture des documents du corpus, ainsi, la lexicologie, la grammaire et la syntaxe peuvent entre autres permettre de décrire ce style. Nous étudions le style des documents en choisissant une approche lexicale restreinte au vocabulaire employé et à la taille des phrases.

Nous calculons pour un corpus *spécialisé* (resp. non *spécialisé*) la couverture de son vocabulaire par un dictionnaire en langue commune (Éq. 1 et 2). Cela constitue un indicateur permettant d'évaluer à quel point le vocabulaire d'un corpus supposé *spécialisé* est différent du vocabulaire d'un corpus non *spécialisé*. Cette méthode est un premier estimateur facile à calculer. Néanmoins, il ne prend pas en compte la terminologie employée pouvant avoir un sens ou une connotation différente dans la langue de spécialité (Cabré, 2002).

Pour calculer la *couverture*, nous nous dotons d'un dictionnaire de mots pour la langue commune. Dans notre cas, nous utilisons les livres français de la bibliothèque du projet Gutenberg [1], composée de 60.000 livres numérisés. Un dictionnaire a été construit à partir de ces textes par Pythoud (1998). Concernant le vocabulaire anglais, nous utilisons le vocabulaire fourni par le modèle de langage EN_CORE_WEB_LG de Honnibal & Montani (2017) présent dans SPACY. La mesure de couverture peut se calculer à deux niveaux : soit en ne considérant qu'une seule occurrence de chaque mot (vocabulaire, Éq. 1), soit en considérant chaque occurrence d'un mot (Éq. 2). Le ratio de couverture pour un corpus C par rapport à un dictionnaire D est calculé comme suit :

1. https://www.gutenberg.org

$$\text{Couv}_{\text{voc}}(C|D) = \frac{|V_C \cap V_D|}{|V_C|} \quad (1) \qquad \text{Couv}_{\text{occ}}(C|D) = \frac{\sum_{w_i \in V_C \cap V_D} \#_{w_i} C}{\#C} \quad (2)$$

avec V_X le vocabulaire du corpus X, $\#X$ le nombre d'unités lexicales du corpus X et $\#_m X$ le nombre d'occurrences de l'unité lexicale m dans le corpus X.

Cet indicateur permet de comparer deux corpus selon leur couverture lexicale. Ainsi, si la couverture au niveau vocabulaire ou au niveau occurrence d'un corpus C_A est plus élevée que celle d'un autre corpus C_B, cela indique que le corpus C_B fait appel à moins de mots non *spécialisés* que le corpus C_A.

Étudier la couverture du vocabulaire par un dictionnaire donne une première idée du lexique utilisé dans un corpus. Néanmoins, cette mesure demeure relativement naïve puisqu'elle est dépendante du dictionnaire utilisé. On propose donc d'étudier la diversité du lexique qui peut être estimée à partir du *type-token ratio* (TTR) : le ratio entre la taille du vocabulaire d'un document et le nombre d'unités lexicales. Cependant, si le nombre d'unités lexicales d'un corpus croît linéairement selon sa taille, la taille du vocabulaire suit une loi différente, la *Heap law* (Herdan, 1960). Le TTR ne permet ainsi pas de comparer deux corpus de taille différente. Pour corriger cela, l'indice *measure of textual lexical diversity* (MTLD) permet d'estimer la diversité lexicale des documents d'un corpus indépendamment de la taille de ces corpus (Mccarthy, 2006; McCarthy & Jarvis, 2010; Torruella & Capsada, 2013). Pour ce faire, le document est divisé en n échantillons d'unités lexicales consécutives de façon à obtenir un seuil de TTR fixé (usuellement $TTR = 0.72$). La valeur de MTLD est ensuite le ratio entre n et la taille $|D|$ du document. Plus l'indice de diversité lexicale est important, plus le lexique employé dans les documents est large.

Analyse de la lisibilité. La lisibilité est définie comme la facilité avec laquelle un lecteur peut comprendre un texte écrit. Son contenu (complexité du vocabulaire et de la syntaxe) et sa présentation permettent de la mesurer. Plusieurs mesures de lisibilité de textes existent. Ces métriques n'utilisent pas les mêmes informations pour qualifier la lisibilité d'une phrase. Nous avons donc décidé d'utiliser les métriques SMOG et ARI qui sont complémentaires : cela permet de mesurer l'influence de la longueur des mots et des phrases sur la lisibilité.

La mesure de lisibilité SMOG définie par GH (1969) pour caractériser la *readability*, c.-à-d. la difficulté qu'aura un lecteur à comprendre un document, fait l'hypothèse qu'il est possible d'estimer la complexité du document à partir du nombre de mots polysyllabiques. SMOG repose sur le nombre de mots contenant trois syllabes ou plus dans chaque échantillon (Éq. 3). Cette mesure correspond au nombre théorique d'années d'éducation nécessaire pour comprendre un texte. Nous calculons le score SMOG en sélectionnant aléatoirement 1000 échantillons de 30 phrases dans le corpus. Le résultat obtenu est la moyenne des scores SMOG pour ces 1000 échantillons.

$$\text{SMOG}(S) = 1.0430\sqrt{|\text{polysyllabes}_S|} + 3.1291 \quad (3)$$

avec $S \subset C$ un échantillon du corpus C pour lequel nous calculons le score SMOG, polysyllabes_S les mots avec 3 syllabes ou plus dans S.

D'une langue à l'autre, il existe des variations dans l'utilisation des mots polysyllabiques. Par exemple, Contreras *et al.* (1999) montrent que sur des corpus parallèles en français et en anglais, le score SMOG est toujours plus faible en anglais. À partir de ce constat, ils introduisent un cadre permettant de comparer le score SMOG obtenu sur deux corpus qui ne sont pas écrits dans la même langue. Ainsi, pour comparer un score SMOG obtenu sur un corpus français avec celui d'un corpus anglais, on applique la formule 4.

$$\text{SMOG}(En) = -1,35 + 0,17 \times \text{SMOG}(Fr) \tag{4}$$

La mesure de lisibilité ARI telle que définie par Senter & Smith (1967) fournit, comme SMOG, une estimation du nombre d'années d'éducation nécessaire pour comprendre un document. Cette métrique est calculée à partir du nombre de caractères, de mots, et de phrases pour un corpus C :

$$\text{ARI}(C) = 4.71 \left(\frac{\text{caractères}_C}{\text{mots}_C} \right) + 0.5 \left(\frac{\text{mots}_C}{\text{phrases}_C} \right) - 21.43 \tag{5}$$

Cette mesure a été construite pour des documents en anglais, les résultats pour les documents du corpus SNCF en français sont donc donnés à titre indicatif.

2.2 Apprentissage de plongements lexicaux et classification

Plongements lexicaux. Pour chaque corpus spécialisé (OHSUMED et SNCF) ou corpus non spécialisé, on apprend un ensemble de plongements de mots à l'aide de l'algorithme CBOW, tel que présenté par Mikolov *et al.* (2013). Chaque mot est représenté par un vecteur de dimension 300. Ces vecteurs sont construits à partir d'une fenêtre de contexte de 5 mots, avec un minimum de 1 ou 2 occurrences de chacun des mots du vocabulaire des corpus, selon la taille du corpus utilisé. Nous utilisons un échantillonnage négatif (*negative sampling*). Les autres paramètres sont ceux recommandés et initialisés par défaut dans l'implémentation Gensim (Řehůřek & Sojka, 2010). On désigne dans la suite de l'article les plongements appris sur les corpus non spécialisés comme plongements *génériques*.

Classification. Les documents des corpus OHSUMED et SNCF à notre disposition sont catégorisés en classes spécifiant le thème du document en question (voir Section 3). L'objectif est d'apprendre un classifieur capable de retrouver ces classes automatiquement. Nous choisissons un classifieur de type réseau de neurones convolutif particulièrement adapté à la classification de documents représentés par des plongements lexicaux. Nous faisons l'hypothèse que le réseau de neurones obtiendra de meilleures performances si les plongements en entrée sont les plus adaptés à la tâche. Ainsi, nous comparons Section 4.4 les résultats obtenus avec le classifieur en fonction des plongements utilisés : *spécialisés* ou *génériques*. L'architecture de ce réseau est détaillée ci-dessous. Les corpus sont séparés en corpus d'apprentissage (TRAIN), en corpus de développement afin d'adapter les hyperparamètres (DEV), et en corpus d'évaluation (TEST). Cette séparation est réalisée de façon stratifiée : les proportions de chacune des classes sont préservées dans les corpus de TRAIN, DEV et TEST.

Architecture des modèles. L'architecture des modèles de classification appris est similaire à celle introduite par Kim (2014) et décrite Figure 1. Chaque document est représenté par la concaténation verticale des plongements des mots du document. Cette représentation est de taille fixe, il s'agit des 3 000 premiers mots de chaque document. Si un document est plus long que cette limite il sera tronqué, s'il est plus court, alors la matrice sera complétée par des 0 (*padding*). Ensuite, l'architecture intègre deux blocs de convolution. Chacun de ces blocs contient : une convolution, une normalisation par lot (*batch normalisation*), une couche d'activation ReLU (Hahnloser *et al.*, 2000) et une couche de *max-pooling*. En sortie des convolutions, le réseau est constitué de trois couches linéaires (*feed forward*), suivies d'une fonction *softmax* afin d'obtenir une distribution de probabilité pour les classes possibles. La dernière couche linéaire comporte autant de neurones que le niveau de classification comporte de classes. Les deux premières comportent respectivement 250 et 100 neurones. Un abandon (*dropout*, (Hinton *et al.*, 2012)) de $p = 0.2$ est réalisé sur ces deux couches. Nous utilisons un optimiseur ADAM (Kingma & Ba, 2015). Afin de réduire le taux de surapprentissage, nous réalisons une régularisation des poids (*weight decay*, Weigend *et al.* (1991)), ainsi qu'un *early stopping* qui arrête l'apprentissage lorsque l'erreur sur le corpus de DEV augmente. Enfin, le pas d'apprentissage est réduit au fil du temps.

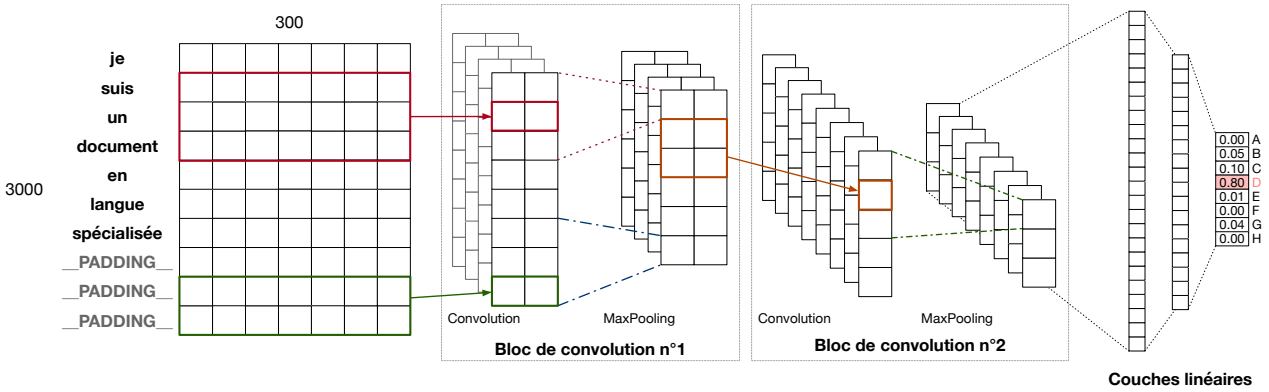

FIGURE 1 – Architecture de nos modèles de classification.

Lors de l'étape de *rétropropagation* (*backpropagation* en anglais) du gradient, deux choix sont possibles : faire la mise à jour des poids sur les couches linéaires et sur les blocs de convolution sans toucher aux plongements (non *trainable*), ou faire la mise à jour des poids jusqu'à la couche des plongements de mots afin de spécialiser ces plongements pour la tâche (*trainable*). Ces deux méthodes de *rétropropagation* sont utilisées dans nos expérimentations.
Nous apprenons un modèle pour chaque corpus, en utilisant en entrée soit les plongements appris sur le corpus *spécialisé*, soit les plongements appris sur Wikipédia (*générique*), et avec les deux modes de *rétropropagation*. Cela représente ainsi 12 modèles distincts, dont nous avons adapté les hyperparamètres à l'aide du corpus DEV.

Métriques utilisées. Nous utilisons le F1-score, qui est la moyenne harmonique du rappel et de la précision. Notre classification étant en classes multiples, nous reportons la moyenne des résultats obtenus individuellement pour chaque classe. Les deux méthodes les plus courantes sont les moyennes *Micro* et les moyennes *Macro*. Les *Micro*-moyennes accordent autant de poids à chacun des documents, tandis que les *Macro*-moyennes ne pondèrent que par le nombre de classes. Ainsi, dans le cas d'un corpus comportant des classes disproportionnées, le modèle obtiendra facilement une *Micro*-moyenne équivalente à la proportion de la classe la plus probable. Nos classes étant disproportionnées, nous utilisons la mesure *Macro*-moyenne.

Significativité des résultats. Notre apprentissage est non déterministe. Ainsi, pour garantir la pertinence des résultats reportés, les expérimentations sont réalisées 10 fois pour chaque configuration. Nous calculons ensuite des intervalles de confiances à 95% pour chaque métrique pour toutes les expériences. Au total, nous avons construit 120 modèles de classification pour l'ensemble de nos corpus, 10 pour chacune des 12 configurations décrites dans le paragraphe précédent.

3 Corpus utilisés

Nous disposons de deux corpus de documents en langue de spécialité annotés en classes : le corpus SNCF (Section 3.1) et le corpus OHSUMED (Section 3.2). Nous utilisons le corpus Wikipédia comme un corpus en langue non spécialisée (Section 3.3).

3.1 Le corpus SNCF

Il s'agit d'un corpus de documents textuels introduit par Dugué *et al.* (2019), en français, privé, appartenant à SNCF. Les documents ont été rédigés par des agents SNCF ou des acteurs du domaine ferroviaire. Ce corpus contient 7 255 documents techniques sur deux domaines : les ressources humaines (définition des différents régimes indemnitaires ou de congés, gratifications, bilan social, etc.) et les opérations (description des manoeuvres dans les gares, consignes de desserte des trains, service de circulation, etc.). Initialement au format PDF, ces documents ont été convertis en documents texte à l'aide de l'outil PDFTOTEXT. Au total, 19 millions d'unités lexicales sont présentes au sein de ce corpus, pour un vocabulaire de 77 mille mots.

Afin de mieux structurer son fonds documentaire, SNCF possède un système de classification hiérarchique de ces documents. Dans nos expérimentations, nous considérons deux niveaux de classification différents. Le premier niveau contient 8 classes, indiquant le thème général du document. Nous notons que la répartition des documents selon ces huit classes est déséquilibrée. En effet, 4 de ces 8 classes représentent 99.8% des documents. Le second niveau de classification, quant à lui, caractérise le sous-thème du document. Celui-ci est composé de 49 classes, avec une distribution très déséquilibrée des documents au sein de ces classes : 97.2% des documents sont représentés par seulement 11 classes.

Nous divisons le corpus SNCF en 3 sous-corpus de façon stratifiée : entraînement TRAIN, développement DEV et TEST, selon les proportions respectives suivantes : 0.60, 0.20, 0.20.

3.2 Le corpus OHSUMED

Le corpus OHSUMED (Hersh *et al.*, 1994) est un corpus public extrait de la base de données MEDLINE, disponible librement et composé de documents textuels en anglais. Ce corpus contient 23 166 abstracts d'articles scientifiques relatifs à différents types de maladies (maladies cardiovasculaires, infections bactériennes, maladies du système digestif, etc.). Ces abstracts ont été rédigés par et à destination de médecins et chercheurs en médecine.

Le corpus contient 23 classes, 16 classes couvrent 90% des documents du corpus. Au total, 4 millions d'unités lexicales sont présentes dans le corpus pour un vocabulaire de 49 mille mots.
Les données du corpus OHSUMED sont séparées en 3 sous-corpus : entraînement TRAIN, développement DEV et TEST selon les proportions décrites dans Joachims (1998) : 0.45, 0.05, 0.50.

3.3 Corpus génériques

Les documents des corpus spécialisés à notre disposition sont en français (SNCF) et anglais (OHSUMED). Ainsi, il est nécessaire de disposer également de deux corpus non *spécialisés* composés pour l'un de documents en français et pour l'autre de documents en anglais. Nous avons pour cela utilisé deux corpus provenant de *dumps* Wikipédia. Un *dump* de Wikipéda est une sauvegarde complète de toute la base de données du site à un instant t. Ces données sont par la suite prétraitées, afin de ne garder que le texte. Ces corpus sont, en comparaison avec les corpus SNCF et OHSUMED, considérés en langue courante : leur contenu se veut non *spécialisé*, et n'est pas restreint à un domaine en particulier.

Le premier *dump* provient de la version anglaise de Wikipédia, enregistrée en mars 2006. Il s'agit du corpus TEXT8 [2]. Celui-ci n'est composé que des 100 premiers méga-octets du *dump* original au format texte. Il contient ainsi 17 millions d'occurrences de mots, et un vocabulaire de 253 mille mots. Le second provient d'un *dump* du Wikipédia français de 2015. Ce corpus s'appelle FR_WIKI_NONLEM (Fauconnier, 2015). Il contient 600 millions d'occurrences de mots, avec un vocabulaire de 191 mille mots.

Afin d'uniformiser les noms de ces corpus, nous nommons par la suite le corpus non spécialisé anglais Wiki-EN et le corpus non spécialisé français Wiki-FR.

4 Expériences et Résultats

4.1 Mesure de la couverture du corpus par dictionnaire

Les dictionnaires choisis pour étudier le vocabulaire d'après l'approche décrite Section 2.1 sont : un dictionnaire français (Pythoud, 1998) pour SNCF, et un dictionnaire anglais issu du modèle de langage EN_CORE_WEB_LG de SPACY (Honnibal & Montani, 2017) pour le corpus OHSUMED. Le dictionnaire français est extrait du corpus Gutenberg et possède un nombre de mots dans son vocabulaire trois fois supérieur au dictionnaire en anglais issu du modèle de langage de SPACY.

	Dictionnaire Français		Dictionnaire Anglais	
	SNCF	Wiki-FR	Ohsumed	Wiki-EN
Couv_{occ}	93.2%	93.8%	96.8%	98.8%
Couv_{voc}	32.6%	72.7%	60.4%	54.8%

TABLE 1 – Couverture du vocabulaire des corpus par deux dictionnaires *génériques*.

Les comparaisons des couvertures des unités lexicales des corpus spécialisés et corpus non spécialisés par rapport aux dictionnaires sont données dans la Table 1.
La même expérience a été réalisée en utilisant le vocabulaire appris lors de la construction des plongements sur les corpus non spécialisés. Ce vocabulaire ne contient que les mots apparaissant au moins 2 fois pour le corpus Wiki-FR, 3 pour Wiki-EN. Ces résultats sont présentés dans la Table 2.

2. Matt Mahoney 2006 ; About the Test Data – http://mattmahoney.net/dc/textdata.html

	Plongements Wiki-FR		Plongements Wiki-EN	
	SNCF	Wiki-FR	Ohsumed	Wiki-EN
Couv_{occ}	91.6%	99.9%*	93.3%	98.9% *
Couv_{voc}	34.0%	88.3%*	47.7%	39.4%*

TABLE 2 – Couverture du vocabulaire des corpus par les plongements de mots *génériques*. Les valeurs notées d'un astérisque* sont directement liées au nombre minimum d'occurrences défini pour qu'un mot soit comptabilisé dans l'espace de représentation.

D'après les calculs de couverture, le corpus OHSUMED est bien couvert (vocabulaire et nombre d'occurrences) par le dictionnaire et les plongements lexicaux *génériques* au regard des résultats obtenus pour le corpus non spécialisé (Wiki-EN). Il semble donc que le corpus OHSUMED comporte principalement du vocabulaire commun. Cette part plus importante de vocabulaire commun peut-être liée au fait que le corpus est composé d'abstracts scientifiques contenant les termes principaux utilisés dans chaque article. Concernant le corpus SNCF, la couverture du vocabulaire est faible au regard des résultats obtenus pour le corpus non spécialisé (Wiki-FR) aussi bien lors de la comparaison au dictionnaire que celle avec les plongements lexicaux. Le vocabulaire SNCF semble donc *spécialisé*. Par ailleurs, on observe que le vocabulaire du corpus non spécialisé français (Wiki-FR) est mieux couvert que celui du corpus non spécialisé anglais (Wiki-EN), alors que la couverture en occurrences est plus forte sur le corpus non spécialisé anglais (Wiki-EN) (Table 1). Cela peut provenir de la différence de taille des deux corpus.

Exemples de mots hors vocabulaires les plus fréquents. Les mots hors vocabulaires du dictionnaire les plus fréquents dans le corpus SNCF sont surtout des acronymes, *spécialisés* du domaine SNCF : *AC*, *PN*, *SGTC*, etc. Ceux non couverts issus du corpus Wiki-FR portent eux sur des noms propres : *France*, *Europe*, *Charles*, *Paul*, etc. En ce qui concerne le corpus OHSUMED, les mots non couverts par le dictionnaire anglais portent sur des noms de maladies ou de termes médicaux : *postoperatively*, *immunohistochemical*, ou encore *histopathologic*. Les unités lexicales non couvertes du corpus Wiki-EN sont en grande majorité des commandes LaTeX. Cela semble montrer un problème de nettoyage des données du corpus anglais *générique* utilisé.
Nous retrouvons globalement les mêmes termes lors de l'analyse en utilisant un dictionnaire extrait des plongements lexicaux. Ces unités lexicales non couvertes sont donc en majorité des termes propres au domaine du corpus *spécialisé* en question.

4.2 Mesure de diversité lexicale

Corpus	MTLD
SNCF	39.7
Wiki-FR	42.9
Wiki-EN	54.2
OHSUMED	66.5

TABLE 3 – Scores de diversité lexicale MTLD pour chaque corpus.

Nous calculons dans un second temps le score de diversité lexicale *MTLD* (Tab. 3) décrit dans la Section 2.1 : plus ce score est élevé, plus le vocabulaire employé dans les documents du corpus est varié. Tout d'abord, la diversité lexicale du corpus SNCF est faible, ce corpus devrait donc être plus aisément couvert par le dictionnaire. Or, nous montrons (Tab. 1) que le vocabulaire du corpus SNCF n'est pas bien couvert par le dictionnaire français (pourtant 3 fois plus grand que celui de l'anglais). Cela semble indiquer que de nombreux termes spécialisés sont employés dans le corpus SNCF. Dans le cas du corpus OHSUMED, la diversité lexicale est importante. Néanmoins, la couverture du vocabulaire du corpus OHSUMED par le dictionnaire anglais est bonne. Le vocabulaire employé est donc peut-être moins *spécialisé* que celui des documents du corpus SNCF. Le corpus SNCF semble donc plus *spécialisé* en termes de lexique

qu'OHSUMED, de par son vocabulaire moins couvert et moins diversifié.

4.3 Mesure de lisibilité des corpus

Les scores de lisibilité ont été calculés à l'aide de la bibliothèque python PY-READABILITY-METRICS [3]. La mesure SMOG définie Section 2.1 pour le corpus SNCF est ensuite convertie pour obtenir le score équivalent en anglais (Éq. 4). Les indices SMOG et ARI sont présentés pour les deux corpus dans la table 4. Le corpus SNCF et le corpus OHSUMED obtiennent des scores SMOG et ARI aussi élevés, ce qui indique qu'ils sont difficilement lisibles. Le score SMOG pour le corpus SNCF est converti en anglais (score brut en français : $\text{SMOG}(\text{FR}) = 22$). La mesure ARI ne possède pas de coefficients de conversion du français vers l'anglais. Nous présentons ainsi les résultats bruts pour ARI. À titre de comparaison, nous donnons les scores SMOG et ARI pour la Déclaration universelle des droits de l'homme (*DUDH*) ainsi que SMOG pour plusieurs ouvrages en anglais. On observe que les phrases sont en moyenne plus longues dans le cas du corpus SNCF (≈29 mots/phrase) mais les mots contiennent en moyenne moins de syllabes (≈1.5 syllabe/mot). Le nombre de syllabes est en moyenne plus important pour le corpus OHSUMED (≈1.9 syllabe/mot) ce qui explique que le score SMOG plus important. Les scores de lisibilité semblent montrer que les corpus SNCF et OHSUMED sont bien spécialisés par le style utilisé dans la rédaction des documents qui les composent (longueur des phrases et nombre de syllabes des termes utilisés).

4.4 Résultats de classification

Nous présentons dans les Figures 2 et 3 les résultats de la classification en *Macro*-F1-score en au cours de l'apprentissage sur le corpus SNCF, tandis que la Figure 4 concerne la classification sur le corpus OHSUMED.
Ces résultats sont présentés sur les corpus de DEV afin de présenter l'évolution de l'apprentissage et non les résultats finaux. En fin d'apprentissage, les résultats obtenus sur les corpus DEV sont proches, et suivent la même tendance que ceux obtenus sur les corpus TEST.

Corpus	SMOG	ARI
SNCF	15.6*	16.4
OHSUMED	17	15.6
DUDH †	13	9.3
Don Quichote‡	11.24	-
La Bible‡	9.35	-
Blanche-Neige‡	6.72	-

TABLE 4 – Scores SMOG et ARI pour les corpus SNCF (FR) et OHSUMED (EN). *Conversion du score FR-EN, † résultats issus de Jakobsen & Skardal (2007), ‡ résultats issus de Contreras *et al.* (1999).

4.4.1 Sur le corpus SNCF

Nous pouvons tout d'abord observer qu'apprendre un classifieur en utilisant les plongements de mots *spécialisés*– ceux appris sur SNCF – sans faire la *rétropropagation* sur ces plongements conduit aux plus mauvais résultats (courbes discontinues avec des marqueurs losanges). Ceux-ci convergent moins vite, et obtiennent des résultats inférieurs aux autres.
Ensuite, les classifieurs entraînés avec les plongements *génériques* (appris sur le corpus Wikipédia français, marqueurs en forme de croix) convergent globalement rapidement, mais pas aussi bien que les modèles de classification appris avec les représentations spécialisées qui bénéficient de l'option *trainable*. Ces modèles convergent plus rapidement vers un meilleur *Macro*-F1-score. Dans le cas du premier niveau de classification, ce modèle n'apporte cependant pas de résultats finaux meilleurs que ceux appris avec les plongements lexicaux non *spécialisés* (marqueurs en forme de croix), ceux-ci

3. https://pypi.org/project/py-readability-metrics/

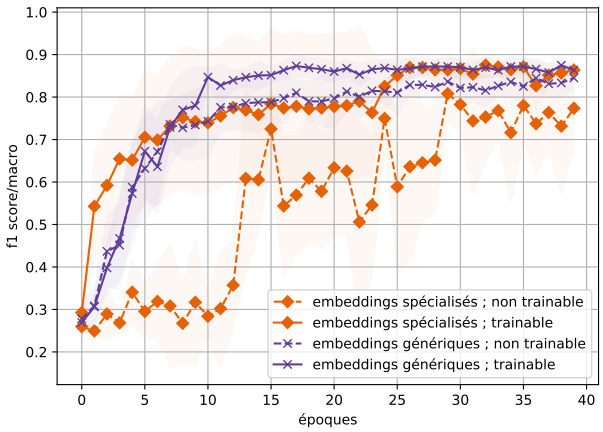

FIGURE 2 – *Macro*-F1-score sur le corpus DEV au cours de l'apprentissage du premier niveau de la classification du corpus SNCF. Intervalle de confiance à 95%.

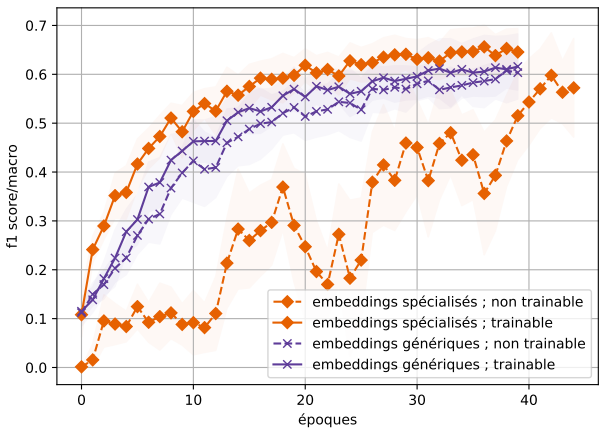

FIGURE 3 – *Macro*-F1-score sur le corpus DEV au cours de l'apprentissage du second niveau de la classification du corpus SNCF. Intervalle de confiance à 95%.

sont équivalents.
Il faut donc faire la rétropropagation de l'erreur en *trainable*. Apprendre des représentations *spécialisées* n'apporte pas de gain notable, il semble donc préférable d'utiliser des plongements pré-appris.

4.4.2 Sur le corpus OHSUMED

Les modèles appris avec les plongements de mots *spécialisés* obtiennent des résultats significativement meilleurs que ceux obtenus avec des plongements *génériques*. Cette différence est bien plus prononcée sur le corpus OHSUMED qu'elle ne l'est sur le corpus SNCF. Pourtant, le corpus OHSUMED est cinq fois plus petit que le corpus SNCF, et apprendre une représentation correcte du vocabulaire de ce corpus semble intuitivement plus difficile que pour le corpus SNCF.
D'après l'indicateur de *couverture* (Section 2.1), le corpus SNCF semble plus *spécialisé* que le corpus OHSUMED, et pourtant c'est pour OHSUMED qu'il est plus pertinent d'apprendre des plongements *spécialisés*. Ces résultats vont donc à l'encontre des intuitions que nous pouvons avoir à l'issue de la Section 2.1. En revanche, ces résultats confirment qu'il est préférable de mettre à jour les plongements de mots lors de l'apprentissage de la tâche de classification (courbes continues *trainable*).

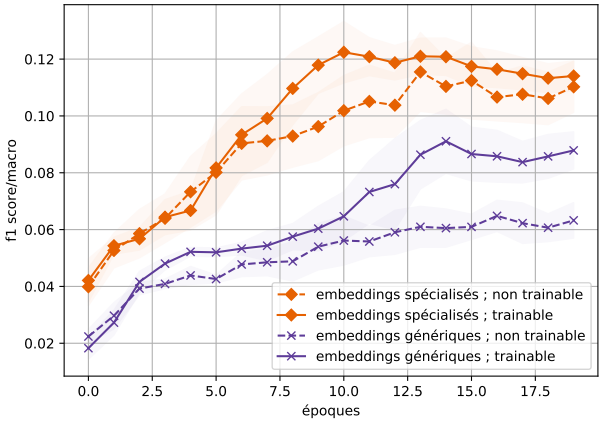

FIGURE 4 – *Macro*-F1-score sur le corpus DEV au cours de l'apprentissage de la classification du corpus OHSUMED. Intervalle de confiance à 95%.

5 Conclusion et perspectives

Dans cet article, nous cherchons à déterminer quels types de plongements utiliser lorsque l'on considère des documents issus d'un corpus en langue de spécialité. En effet, il est possible d'utiliser des plongements préappris sur un corpus plus volumineux tel que Wikipédia, ou bien d'apprendre des plongements sur le corpus *spécialisé* de taille réduite. Pour mener à bien nos expérimentations, nous utilisons deux corpus en langue de spécialité, OHSUMED et SNCF. Nous détaillons dans un

premier temps quelques indicateurs pour tenter d'estimer le degré de spécialité de ces corpus, par ex. la *couverture* du vocabulaire, la diversité du lexique et des scores de lisibilité. On constate que le vocabulaire du corpus SNCF semble particulièrement *spécialisé*, il est très peu couvert par les dictionnaires alors qu'il est moins varié. Le corpus OHSUMED est lui plutôt bien couvert. Par ailleurs, nous montrons avec les scores de lisibilité que ces corpus semblent complexes, c.-à-d. construits avec des phrases longues et de nombreux mots polysyllabiques, en particulier pour OHSUMED. Enfin, le corpus OHSUMED est cinq fois plus petit que le corpus SNCF au regard du nombre d'unités lexicales, il semble donc plus difficile d'obtenir des plongements de bonne qualité sur ce corpus.
Nous apprenons ensuite à catégoriser les documents de ces corpus en utilisant en entrée de nos classifieurs soit des plongements appris sur des corpus Wikipédia (*générique*), soit des plongements appris sur le corpus considéré *spécialisé*. Il s'agit de voir quels types de plongements sont les plus efficaces pour résoudre ce problème. Les résultats confirment les travaux de Wang *et al.* (2018) sur les corpus médicaux : dans le cas d'OHSUMED, les résultats sont meilleurs avec des plongements appris sur ce même corpus. Pourtant, ce corpus est plus petit que SNCF et son vocabulaire mieux couvert avec un dictionnaire classique. En revanche, dans le cas de SNCF, les plongements appris sur ce dernier ne sont pas plus performants que ceux appris sur Wikipedia. Le corpus est pourtant moins bien couvert, et de plus grande taille.

Il est donc difficile de conclure avec ces résultats préliminaires. Nous pouvons tout de même dire que des plongements appris sur un corpus non spécialisé permettent d'obtenir une *baseline* de bonne qualité, même si dans le cas d'OHSUMED, réapprendre les plongements améliore les résultats. Par ailleurs, il est nécessaire de faire la *rétropropagation* jusqu'à la couche de plongements lexicaux. Spécialiser les plongements pour la tâche garantit de bons résultats pour la classification. Enfin, la *couverture*, la *lisibilité* ou la *taille* du corpus ne semblent pas être des indicateurs suffisants pour pouvoir décider s'il faut ou non apprendre des plongements *spécialisés*. Par exemple, de façon contre-intuitive, même si l'on dispose de peu de cooccurrences sur OHSUMED, il vaut mieux apprendre des plongements sur ce corpus pour classifier ses documents. Et même si le vocabulaire de SNCF semble plus *spécialisé*, contre-intuitivement, des plongements appris sur un corpus non spécialisé obtiennent de bons résultats.

En perspective, il s'agirait bien entendu de reproduire cette étude sur d'autres corpus. Dans notre étude, nous disposons de deux corpus en langue différente, et cela peut fragiliser nos conclusions : les dictionnaires sont différents, les corpus Wikipédia également. De plus, pour mieux caractériser la spécificité des corpus, nous pourrions utiliser d'autres indicateurs. Par exemple, la couverture des acronymes et abréviations à l'aide d'un lexique, la distribution des étiquettes morphosyntaxiques (POS), ou encore la distribution des transitions entre ces étiquettes semblent être des estimateurs utiles (Campbell & Johnson, 2001). Par ailleurs, il serait également intéressant d'étudier les effets possibles de la terminologisation en nous basant sur des ontologies spécialisées au domaine des corpus traités. Enfin, il serait également intéressant de comparer les espaces appris par la méthode de plongements afin d'étudier de manière qualitative ce qui différencie les plongements appris sur les corpus *spécialisés*, des plongements appris sur les corpus non *spécialisés*.

6 Remerciements

Ces travaux ont été financés dans le cadre du partenariat entre SNCF I&R et le LIUM. Nous remercions particulièrement L. Lefeuvre de nous avoir autorisé à mener ces travaux et l'équipe du LIUM (N. Dugué, N. Camelin et J. Wottawa) pour leurs conseils avisés.

Références

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. DOI : 10.1162/tacl_a_00051.

CABRÉ M. T. (2002). Terminologie et dictionnaires. *Meta*. DOI : 10.7202/002182ar.

CAMPBELL D. A. & JOHNSON S. B. (2001). Comparing syntactic complexity in medical and non-medical corpora. In *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, p. 90–94 : American Medical Informatics Association.

CHARNOCK R. (1999). Les langues de spécialité et le langage technique : considérations didactiques. *ASp*. DOI : 10.4000/asp.2566.

CHUJO K. & UTIYAMA M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, **34**(2), 255–269. DOI : 10.1016/j.system.2005.12.003.

CONDAMINES A. (1997). Langue spécialisée ou discours spécialisé ? In L. LAPIERRE, I. OORE & H. RUNTE, Éds., *Mélanges de linguistique offerts à Rostislav Kocourek*, p. 171–184. Les presses d'Alfa. HAL : halshs-01380935.

CONTRERAS A., GARCÍA-ALONSO R., ECHENIQUE M. & DAYE-CONTRERAS F. (1999). The SOL formulas for converting SMOG readability scores between health education materials written in Spanish, English, and French. *Journal of Health Communication*. DOI : 10.1080/108107399127066.

CRESSOT M. & JAMES L. (1996). *Le Style et ses Techniques*. Presses Universitaires de France - PUF.

DUGUÉ N., CAMELIN N., LEFEUVRE L., LI X., REUTENAUER C. & VAUDAPIVIZ C. (2019). Apprentissage et évaluation de plongements lexicaux sur un corpus SNCF en langue spécialisée. In *Extraction et Gestion des Connaissances*, Metz, France. HAL : hal-01982661.

EL BOUKKOURI H., FERRET O., LAVERGNE T. & ZWEIGENBAUM P. (2019). Embedding Strategies for Specialized Domains : Application to Clinical Entity Recognition. p. 295–301. DOI : 10.18653/v1/p19-2041.

FAUCONNIER J.-P. (2015). French Word Embeddings.

GH M. (1969). SMOG grading : A new readability formula. *Journal of Reading*.

HAHNLOSER R., SARPESHKAR R., MAHOWALD M., DOUGLAS R. & SEUNG H. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, **405**, 947–51. DOI : 10.1038/35016072.

HERDAN G. (1960). *Type-token mathematics : A textbook of mathematical linguistics*, volume 4. Mouton.

HERSH W., BUCKLEY C., LEONE T. J. & HICKAM D. (1994). OHSUMED : An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994*, SIGIR '94, p. 192–201, New York, NY, USA : Springer-Verlag New York, Inc. DOI : 10.1007/978-1-4471-2099-5_20.

HINTON G. E., SRIVASTAVA N., KRIZHEVSKY A., SUTSKEVER I. & SALAKHUTDINOV R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arxiv.org*.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

HUANG E. H., SOCHER R., MANNING C. D. & NG A. Y. (2012). Improving word representations via global context and multipleword prototypes. In *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*.

JAKOBSEN T. & SKARDAL T. (2007). Readability index. *Agder University*.

JOACHIMS T. (1998). Text categorization with support vector machines : Learning with many relevant features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. DOI : 10.1007/s13928716.

KIM Y. (2014). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1746–1751, Doha, Qatar : Association for Computational Linguistics. DOI : 10.3115/v1/D14-1181.

KINGMA D. P. & BA J. L. (2015). Adam : A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). BioBERT : A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240. DOI : 10.1093/bioinformatics/btz682.

LIU Q., JIANG H., WEI S., LING Z. H. & HU Y. (2015). Learning semanticword embeddings based on ordinal knowledge constraints. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*. DOI : 10.3115/v1/p15-1145.

MCCARTHY P. M. (2006). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International Section A : Humanities and Social Sciences*.

MCCARTHY P. M. & JARVIS S. (2010). MTLD, vocd-D, and HD-D : A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*. DOI : 10.3758/BRM.42.2.381.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

PALTRIDGE B. & STARFIELD S. (2016). English for specific purposes. In *Handbook of Research in Second Language Teaching and Learning*, volume 3, p. 56–67. Taylor and Francis. DOI : 10.4324/9781315716893.

PYTHOUD C. (1998). Français-GUTenberg : un nouveau dictionnaire français pour ISPELL. Problèmes résolus et intégration de contributions extérieures. *Cahiers GUTenberg*. DOI : 10.5802/cg.237.

SCHMITT D. (2002). Learning Vocabulary in Another Language. I.S.P. Nation. *ELT Journal*. DOI : 10.1093/elt/56.1.91.

SENTER R. J. & SMITH E. A. (1967). *Automated readability index*. Rapport interne, CINCINNATI UNIV OH.

TORRUELLA J. & CAPSADA R. (2013). Lexical Statistics and Tipological Structures : A Measure of Lexical Richness. *Procedia - Social and Behavioral Sciences*. DOI : 10.1016/j.sbspro.2013.10.668.

VAN DER YEUGHT M. (2016). Protocole de description des langues de spécialité. *Recherche et Pratiques Pedagogiques en Langues de Specialite - Cahiers de l'APLIUT*. DOI : 10.4000/apliut.5549.

WANG Y., LIU S., AFZAL N., RASTEGAR-MOJARAD M., WANG L., SHEN F., KINGSBURY P. & LIU H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics*. DOI : 10.1016/j.jbi.2018.09.008.

WEIGEND A. S., RUMELHART D. E. & HUBERMAN B. A. (1991). Generalization by Weight-Elimination with Application to Forecasting. In R. P. LIPPMANN, J. E. MOODY & D. S. TOURETZKY, Éds., *Advances in Neural Information Processing Systems 3*, p. 875–882. Morgan-Kaufmann.

YAMADA I., ASAI A., SAKUMA J., SHINDO H., TAKEDA H., TAKEFUJI Y. & MATSUMOTO Y. (2018). Wikipedia2Vec : An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia.

ŘEHŮŘEK R. & SOJKA P. (2010). Software framework for topic modelling with large corpora. p. 45–50. DOI : 10.13140/2.1.2393.1847.