Apport des termes complexes pour enrichir l'analyse distributionnelle en domaine de spécialité

Mérième Bouhandi
LS2N, Université de Nantes
merieme.bouhandi@ls2n.fr

RÉSUMÉ

L'essor et les performances des modèles de sémantique distributionnelle sont principalement dus à l'accroissement de la quantité de données textuelles disponibles ainsi qu'à la généralisation des méthodes neuronales pour la construction de ces modèles. La qualité des représentations distribuées est souvent corrélée à la quantité de données disponibles et les corpus spécialisés, généralement d'une taille modeste, se trouvent de ce fait pénalisés. Alors que la plupart des modèles de sémantique distributionnelle traitent de mots isolés, nous partons de l'hypothèse que l'exploitation des termes, notamment complexes, est essentielle notamment en langue de spécialité car ils sont porteurs d'une dimension sémantique supplémentaire. Ainsi, nous évaluons une méthode de généralisation des contextes distributionnels par un mécanisme d'inclusion lexicale reposant sur les termes complexes. Nos différentes représentations distributionnelles sont ensuite confrontées à une tâche d'extraction de concepts médicaux à partir des rapports médicaux proposée par l'édition 2010 du challenge i2b2.

ABSTRACT

Multi-words terms impact in improving domain-specific distributed representations

The success of distributional semantic models (DSM) is mainly due to the increase in the amount of available textual data and the democratisation of neural methods for building these models. The quality of distributed representations is often correlated with the quantity of data available, and domain-specific corpora are generally very small in size. While most DSMs deal with words, the contribution of terms, especially multi-word terms, is essential in domain-specific language because they carry an additional semantic dimension. However, they generally remain poorly represented in the context vectors and are often excluded when measuring similarity. Thus, we evaluate a method of generalising distributional contexts by performing lexical inclusion, using relations acquired in the corpora. The different distributional representations of our terms are then tested on a task of medical concepts classification from clinical reports proposed by the 2010 edition of the i2b2/VA Challenge.

MOTS-CLÉS : sémantique distributionnelle, inclusion lexicale, plongements de mots, extraction d'information.

KEYWORDS: distributional semantics, lexical inclusion, word embeddings, information extraction.

1 Introduction

L'analyse distributionnelle est un domaine renouvelé depuis la généralisation des approches neuronales qui sont venues concurrencer les approches traditionnelles à base de sacs de mots. Cette analyse repose sur l'hypothèse que les mots se trouvant dans des contextes d'apparition similaires tendent à

avoir un sens proche. Selon l'hypothèse harrissienne, le degré de recouvrement des contextes de deux mots permet de déterminer leur proximité sémantique. L'idée sous-jacente aux méthodes d'analyse distributionnelle est donc de regrouper les mots supposés sémantiquement proches.

Ces approches, qui se sont largement démocratisées notamment grâce à l'accroissement de la quantité de données textuelles disponibles sur le net, permettent de faire reposer le calcul du sens sur l'analyse statistique des contextes des mots dans un corpus et de ne pas avoir besoin de recourir à des sources extérieures de connaissance. Ces ressources, constituées, par exemple, d'archives de journaux, de publications glanées sur les réseaux sociaux, recommandations de films ou de restaurants ou de textes généraux issus du Web a permis la généralisation des méthodes distributionnelles et l'application de ces méthodes à grande échelle et sur une multitude de tâches. Les performances de ces méthodes se sont également nettement améliorées avec l'utilisation des méthodes neuronales pour la construction des modèles (Mikolov *et al.*, 2013; Bojanowski *et al.*, 2016).

Deux approches principales se côtoient. La première, dite approche explicite, est une approche par sac de mots. Celle-ci consiste en la création d'une matrice pondérée de mots cibles et des contextes distributionnels qui leur sont associés suivie ou non d'une méthode de réduction de dimension. La seconde, dite approche prédictive, consiste en l'utilisation de méthodes neuronales pour créer des représentations denses de faible dimension, en apprenant à prédire de façon optimale les contextes distributionnels autour d'un mot cible. Cependant, la qualité des représentations distribuées obtenues à travers ces méthodes est souvent corrélée au volume de données disponibles et des corpus généralistes de très grande taille sont utilisés pour générer ces matrices.

Or, dans le cas d'un domaine de spécialité, les corpus sont généralement de taille très modeste et ces méthodes se retrouvent dès lors moins efficaces. De plus, les corpus en langue de spécialité, qui véhiculent un savoir propre au domaine en question, présentent des différences fondamentales avec les domaines généralistes, autant d'un point de vue structurel que dans leur élaboration linguistique, étant donné l'importance des termes pour dénoter les notions du domaine. Des adaptations des méthodes disponibles sont donc alors nécessaires.

Ainsi, nous partons de l'hypothèse que l'exploitation des termes, notamment complexes, est essentielle notamment en langue de spécialité (Morin, 2007). Alors que la plupart des modèles de sémantique distributionnelle traitent de mots isolés, ces unités terminologiques sont porteuses d'une dimension sémantique supplémentaire et les relations classiques qu'elles ont entre elles peuvent généralement être trouvées dans une ressource terminologique propre au domaine. Les termes complexes se classent dans deux principales catégories (Périnet, 2015) : des multi-termes (polylexicaux) tel que *"muscle ischio-jambier"* ou des termes morphologiquement complexes, tel *"cardiovasculaire"*. Les constituants de ces termes sont liés entre eux par une variété de relations : modification (ex. *"muscle ischio-jambier"*), complémentation (ex. *"perforation intestinale"*), composition morphologique (ex. *"cardiovasculaire"*) ou encore affixation (ex. *"hyperthyroïdie"*) ou suffixation (ex. *"parasitose"*).

L'article se présente comme suit. Dans un premier temps (section 2), nous allons voir en quoi consistent ces méthodes distributionnelles à base de sacs de mots et neuronales et les différentes manières de les mettre en œuvre, en prenant en compte les spécificités des corpus en langue de spécialité. Dans un second temps (section 3), nous présenterons quelques expériences exploratoires que nous avons effectuées dans le but d'étudier l'effet sur les représentations distribuées de la prise en compte des termes complexes ainsi que de l'inclusion lexicale dans les vecteurs de contexte. Finalement (section 4), nous discutons des perspectives de travaux futurs.

2 Méthodes distributionnelles

Les méthodes d'analyse distributionnelle s'appuient sur l'hypothèse distributionnelle selon laquelle les mots se trouvant dans des contextes d'apparition similaires tendent à être sémantiquement proches. Le contexte, qui peut être *"une proposition, une phrase, un paragraphe, un document est souvent réduit aux mots qui le constituent"* (Zweigenbaum & Habert, 2006).

Il existe plusieurs manière de définir cette notion de contexte dans le cadre de l'analyse distributionnelle. Tout d'abord, deux type de modèles existent. Certains sont basés sur le document comme l'analyse sémantique latente (LSA, *Latent Semantic Analysis*) (Deerwester *et al.*, 1990). Ces méthodes partent du principe que le document entier fait partie du contexte d'un mot étudié m, par observation des occurrences du dit mot dans le contexte du document. Les autres méthodes sont basées sur les mots : le contexte d'un mot m est alors représenté par les mots se trouvant dans son voisinage immédiat. Lequel peut être de nature graphique (les contextes sont récupérés à l'aide d'une fenêtre d'une taille donnée, centrée sur le mot m) (Manning *et al.*, 2008; Périnet, 2015; Jurafsky & Martin, 2018) ou syntaxique (les relations de dépendance syntaxique déterminent le contexte de m) (Fabre *et al.*, 2014). Des expériences ont montré que les modèles à base de dépendances syntaxiques permettaient souvent d'acquérir des relations paradigmatiques alors que les modèles graphiques à base de co-occurrents tendent à extraire des concepts associés (Fabre *et al.*, 2014; Périnet, 2015). Combiner ces modèles pourrait alors permettre de tirer parti des avantages de chacun d'entre eux.

Il existe deux représentations computationnelles principales pour ces modèles distributionnels : les méthodes utilisant des graphes et les méthodes vectorielles. Les méthodes vectorielles reposent sur la représentation des mots dans un espace vectoriel en fonction de leurs propriétés distributionnelles. Une matrice pondérée de termes $\times$ termes ou de termes $\times$ contextes est générée et chaque vecteur représente alors à la fois les informations contextuelles et les données statistiques distributionnelles (Périnet, 2015). Ces représentations sont particulièrement prisées car elles ont l'avantage de permettre un calcul facile de la similarité sémantique entre deux termes : il suffit alors de calculer la distance entre leurs vecteurs de contexte.

2.1 Approches *classiques* par sacs de mots

Les méthodes distributionnelles classiques, dites approches en sacs de mots, reposent sur la création et l'exploitation d'une matrice de co-occurrences (Jurafsky & Martin, 2018; Manning *et al.*, 2008; Périnet, 2015). Cette matrice de *termes* $\times$ *contextes* est obtenue à travers l'acquisition dans le corpus des occurrences des mots cibles et de leurs contextes. Ces matrices ont l'avantage de permettre un calcul facile de la similarité sémantique entre deux termes. Celle-ci est calculée en prenant en compte la distance entre les deux vecteurs de contexte, en mesurant leur angle par exemple ou en comparant leur nombre de contextes communs par rapport à l'ensemble des contextes de ces mots.

Le simple calcul des occurrences des mots ne suffit généralement pas. Un mot avec une fréquence élevée peut ne pas être très discriminant en tant que contexte pour un mot cible, mais en prenant en compte uniquement son nombre d'occurrences une importante toute particulière lui sera accordée. La table 1 illustre la matrice de termes $\times$ contextes pour le texte *"The patient is an 81 year old female with a history of cerebrovascular accident , atrial fibrillation , hypothyroidism and dementia"*.

Différents moyens de pondérer cette matrice existent. L'un d'entre eux est l'information mutuelle ponctuelle ou PMI (Church & Hanks, 1990), qui permet de quantifier la probabilité que deux mots w (*"word"*) et c (*"context"*) apparaissent ensemble (distribution jointe) comparée à la probabilité qu'ils

	patient	year	old	female	history	cerebrovascular accident	atrial fibrillation	hypothyroidism	dementia
patient	1	1	1						
history			1	1	1	1	1		
cerebrovascular accident				1	1	1	1	1	
atrial fibrillation					1	1	1	1	1
hypothyroidism						1	1	1	1
dementia							1	1	1

TABLE 1: Matrice de co-occurences des mots cibles et de leurs contextes

soient tout à fait indépendants (distribution indépendante) :

$$PMI_{w,c} = log_2 \frac{P_{w,c}}{P_w P_c}$$

Les valeurs de la $PMI_{w,c}$ peuvent être positives ou négatives. Cependant, il est plus naturel de calculer la similarité des deux mots que leur "dissimilarité". La $PPMI_{w,c}$ ou information mutuelle ponctuelle positive permet de pallier ce problème en remplaçant les valeurs négatives par 0.

$$PPMI_{w,c} = \begin{cases} PMI_{w,c} & \text{si } PMI_{w,c} > 0 \\ 0 & \text{sinon} \end{cases}$$

Un autre problème se pose alors : les mots particulièrement rares tendent à avoir des valeurs de $PPMI_{w,c}$ particulièrement élevées. On peut réduire ce phénomène en remplaçant P_c par $P_{\alpha c}$ (Levy *et al.*, 2015). Généralement, la valeur de α est définie à $0,75$.

$$P_{\alpha c} = \frac{count(c)^\alpha}{\sum_c count(c)^\alpha} \qquad PMI_{\alpha w,c} = log_2 \frac{P_{w,c}}{P_w P_{\alpha c}}$$

$$PPMI_{\alpha w,c} = \begin{cases} PMI_{\alpha w,c} & \text{si } PMI_{\alpha w,c} > 0 \\ 0 & \text{sinon} \end{cases}$$

À l'issue de cette construction de matrice de co-occurrences pondérée, nous obtenons pour chacun de nos mots cibles un vecteur de contextes qui le représente.

Le principal problème rencontré avec l'utilisation de ces représentations vectorielles est ce que l'on appelle "*the curse of dimensionality*". En effet, et à plus forte raison dans les domaines de spécialité, peu de mots co-occurrents avec beaucoup de contextes et cela résulte en des matrices de très grande dimension et très éparses, avec beaucoup d'éléments à zéro. Au-delà des problèmes liés au temps de calcul de ces matrices, les matrices creuses sont moins efficaces que les matrices denses pour capturer certaines relations (la synonymie notamment).

Plusieurs méthodes ont été proposées pour pallier ce problème de matrice creuse dont la Décomposition en Valeurs Singulières (SVD, *Singular Value Decomposition*) (Jurafsky & Martin, 2018; Manning *et al.*, 2008). Cette méthode, initialement appliquée à des matrices de termes $\times$ documents

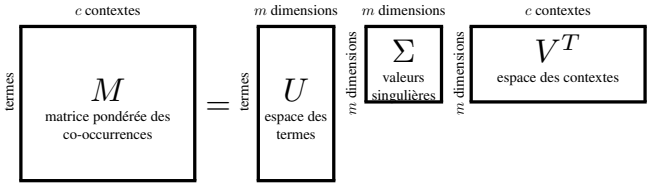

FIGURE 1: Décomposition de la matrice de termes × contextes en ses valeurs singulières

dans le cadre de la méthode LSA et dans un contexte de recherche d'information. Elle se présente comme suit : une matrice de termes × contextes M peut être décrite telle que $M = U\Sigma V^T$ (figure 1).

On peut ensuite ne conserver qu'un nombre k ($k \leq m$) de dimensions (figure 2). Ces k top dimensions correspondent alors aux k plus importantes valeurs singulières et un produit des matrices U, Σ et V^T renvoie une bonne approximation, dense de surcroît, de notre matrice creuse M de départ.

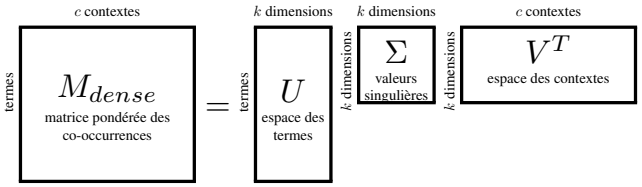

FIGURE 2: Décomposition de la matrice de termes × contextes en ses valeurs singulières en ne gardant les top k plus importantes valeurs singulières

2.2 De la sémantique distributionnelle *classique* aux réseaux de neurones : plongements de mots ou *word embeddings*

Une autre approche consiste à utiliser des réseaux de neurones. Bien qu'il existe de nombreuses sortes de plongements de mots (Bengio *et al.*, 2003; Collobert & Weston, 2008), nous nous ne concentrons ici que sur deux de ces méthodes. L'outil Word2vec (Mikolov *et al.*, 2013) permet de calculer des représentations vectorielles en utilisant deux architectures différentes : le modèle CBOW (*Continuous Bag of Words*) et le modèle Skip-Gram. CBOW entraîne le réseau de neurones pour prédire un mot w en fonction de son contexte dans la phrase. Pour la phrase *"the quick brown fox jumps over the lazy dog"* et une fenêtre de taille 2, le modèle va prédire le mot *"brown"* à partir de ses contextes w_{-1} (*"quick"*) et w_1 (*"fox"*). Skip-Gram prédit les contextes en fonction d'un mot w. Ainsi, pour la même phrase, le modèle essaye de prédire w_{-1} (*"quick"*) et w_1 (*"fox"*) à partir du mot cible w (*"brown"*).

FastText (Bojanowski *et al.*, 2016) permet de faire même chose en s'appuyant également sur les artchitectures Skip-Gram et CBOW mais en utilisant comme contexte l'information contenue au niveau des *n-grams* de caractères à la place des mots complets. Avec un découpage en $n = 6$, *"coronarography"* devient donc : *"<coron", "corona", "oronar", "onarog", "narogr", "narogr", "arogra", "rograp", "ograph", "graphy", "raphy>"*.

Cette méthode est particulièrement intéressante, et à plus forte raison dans les domaines de spécialité comme la médecine, où les mots sont souvent morphologiquement complexes ("*radio-*", "*cardio-*", "*-ite*", "*hyper-*", "*hemo-*", etc). Cela permet de rapprocher des mots ayant une composition similaire, *"coronarography"* et *"mammography"* par exemple, sur la base de leur composition (ici, les deux mots partagent le suffixe *"graphy"* qui fait référence à un mode de représentation écrite ou imagée).

2.3 Les plongements de mots ou *word embeddings* contextuels

Word2Vec et *FastText* ont pour inconvénient qu'ils n'apprennent qu'une représentation par mot – même si ce dernier est polysémique. De nouvelles approches sont apparues récemment pour construire des *word embeddings* contextuels. Nous présentons ici deux méthodes : *ELMo* et *BERT*.

ELMo (Peters *et al.*, 2018) (*Embeddings from Language Models*) est un modèle pré-entrainé dont l'architecture est basée sur des LSTM bidirectionnels. Les couches de niveau supérieur du modèle capturent les aspects contextuels des embeddings de mots tandis que les couches de niveau inférieur en capturent les aspects plutôt syntaxiques tout en prenant en compte la polysémie d'un mot compte-tenu de son contexte. De plus, le modèle apprend des représentations sur les caractères et permet, à l'instar de FastText, d'avoir un modèle robuste capable de gérer le problème des mots hors vocabulaire. Cependant, de par son architecture, à chaque étape de l'entraînement, *ELMo* se voit contraint de concaténer la "vue" arrière et la "vue avant".

BERT (Devlin *et al.*, 2018) (*Bidirectional Encoder Representations from Transformers*) est également un modèle de langue profond qui a rapidement fait office d'état de l'art dans de nombreuses tâches de traitement du language. Il s'appuie sur des couches de *transformers*. Il est donc directement bidirectionnel, ce qui lui donne un avantage par rapport à *ELMo*.

2.4 Représentation des termes complexes

Toutes ces approches sont généralement utilisées pour extraire les contextes de termes simples. Pour la représentation distribuée de représentation de termes complexes ou d'expression multi-mots, l'état de l'art propose d'additionner les représentations des mots qui les composent (Mitchell & Lapata, 2010). D'autres méthodes existent, notamment la fixation des termes complexes dans le corpus, avant entraînement, avec des résultats moins intéressants (Hazem & Daille, 2018).

2.5 Mesures de similarité

Après avoir extrait et pondéré les contextes pour chacun des termes, il convient alors de comparer les couples de mots cibles en calculant un score de similarité sémantique entre leurs vecteurs de contexte. Pour cela, il existe plusieurs mesures de similarités. Le choix d'une bonne mesure dépend de plusieurs facteurs, notamment le type de corpus utilisé, le type de relations visées et la dispersion des données. Deux mesures sont communément utilisées pour l'analyse distributionnelle (Périnet, 2015; Jurafsky & Martin, 2018) : l'indice de Jaccard et le cosinus.

L'indice de Jaccard compare le nombre de contextes communs de deux mots cibles m et n à l'ensemble des contextes de ces deux mots $C(m)$ et $C(n)$. Ainsi :

$$jaccard(m,n) = \frac{|C(m) \cap C(n)|}{|C(m) \cup C(n)|}$$

Le cosinus permet de déterminer la similarité sémantique de deux vecteurs en calculant leur angle. Cette méthode de calcul de similarité ne prend en compte ni la distance, ni la longueur des vecteurs et contrairement à l'indice de Jaccard, elle permet la prise en compte de la fréquence d'apparition des mots dans le vecteur de contexte :

$$cosine(m,n) = \frac{\sum_{i=1}^{N} m_i n_i}{\sqrt{\sum_{i=1}^{N} m_i^2}\sqrt{\sum_{i=1}^{N} n_i^2}}$$

2.6 Évaluation des représentations distribuées

Il est difficile d'évaluer ces représentations distribuées notamment parce qu'elles capturent une notion très large de proximité sémantique (Fabre *et al.*, 2014; Périnet, 2015). Les méthodes d'évaluation consistent souvent à comparer les relations obtenues entre le termes aux termes et relations extraites d'une ressource terminologique propre au domaine, à mettre à l'épreuve les représentations sur une tâche donnée où l'étude des relations entre les mots est nécessaire ou, par une très coûteuse évaluation intrinsèque, soumettre les relations qui en sont issues au jugement humain. Malgré cela, il y a un souvent un décalage entre les ressources prévues pour l'évaluation et les résultats fournis.

On sait cependant que la qualité des représentations distribuées est souvent corrélée à la quantité de données disponible ; en effet, cela permet aux modèles d'être moins affectés par le problème de dispersion des données. Or, à l'exception de quelques cas particuliers, les corpus de spécialité sont généralement de très petite taille – souvent de l'ordre du million de mots. De plus, nous partons de l'hypothèse que l'apport des termes complexes lors de la construction de nos vecteurs de contexte est important car ils sont porteurs d'une dimension sémantique supplémentaire (Morin, 2007). Cependant, selon Périnet (2015), les termes complexes sont parfois trop peu représentés dans les vecteurs de contexte. De ce fait, ils se retrouvent alors généralement écartés du calcul de similarité.

3 Approche poursuivie

Pour explorer cette problématique, nous partons de l'approche de Périnet (2015) qui consiste en la réduction de la dimensionnalité de la matrice de contexte par abstraction des contextes distributionnels. La méthode d'abstraction utilisée, l'inclusion lexicale, est une méthode de généralisation sémantique qui consiste à remplacer un terme complexe par sa tête. Dans le cas de Périnet (2015), cela se fait à partir de relations extraites automatiquement du corpus. L'inclusion lexicale permet de garder l'idée principale du terme complexe considéré en se débarrassant de ses spécificités. Par exemple *"examen médical"* devient *"examen"* et *"radiographie du col du fémur"* de vient *"radiographie"*. Ainsi, on peut dire que si un terme est inclus dans un autre, il existe une relation d'hyperonymie entre eux (Grabar & Zweigenbaum, 2004). Cela permet de réduire la dimensionnalité de la matrice tout en augmentant la fréquence d'apparition de certains contextes et donc d'augmenter la similarité entre les termes.

Nos représentations distribuées, construites sur un corpus spécialisé de petite taille, seront ensuite confrontées à une tâche de classification.

3.1 Données utilisées

L'évaluation de nos représentations distribuées a été réalisée sur la base du corpus proposé par l'édition 2010 du i2b2/VA *Natural Language Processing Challenges for Clinical Records*. Ce challenge comprend une tâche axée sur l'extraction de concepts médicaux à partir des rapports médicaux anonymisés. Initialement, 394 rapports annotés pour le corpus d'entraînement, 477 pour le test et 877 non annotés avaient été remis aux participants (Uzuner *et al.*, 2011). Cependant, une partie de ce corpus, fournie par *Partners HealthCare* et *Beth Israel Deaconess Medical Center*, n'est plus disponible. Nous avons donc utilisé une sous-partie de ce corpus : 170 rapports annotés pour le corpus d'entraînement et 256 pour le test, ainsi que 267 non annotés (392k mots, dont 34k mots uniques).

Cette édition du challenge proposait 3 tâches autour de l'extraction de concepts médicaux des corpus : i) classification des concepts extraits, ii) des assertions sur les concepts extraits et iii) des relations

entre les concepts extraits. Nous nous intéresserons à la tâche 1 dont voici un extrait de la référence :

```
c="concept text" offset||t="concept label"
c="prostate cancer" 48:3 48:4||t="problem"
c="chest x-ray" 155:1 155:2||t="test"
c="chemotherapy" 32:4 32:4||t="treatment"
```

Une partie du corpus de l'édition 2012 (1k documents) du i2b2/VA Natural Language Processing Challenges for Clinical Records a été rajoutée lors de l'apprentissage des représentations distribuées, dans le but de les enrichir et d'obtenir des représentations plus significatives.

L'utilisation des abréviations et acronymes médicaux ainsi que les fautes de saisie sont difficiles à corriger ou à normaliser sans avoir des connaissances préalables du domaine, ce qui rend difficile le prétraitement des rapports médicaux. Un premier nettoyage à été réalisé avant le passage des documents dans l'extracteur terminologique YaTea (Aubin & Hamon, 2006) pour la sélection des termes caractéristiques du domaine, notamment la normalisation de certains mots très fréquents et écrits sous diverses formes (table 2).

Terme	Dr.	M.d.	w/	w/o	w/out	or.
Normalisation	Doctor	Doctor	with	without	without	operation room

TABLE 2: Quelques exemples de normalisations effectuées

Les documents ont ensuite été segmentés en phrases, puis un étiquettage morpho-syntaxique a été effectué avec l'outil TreeTagger, suivi par la lemmatisation et l'extraction des termes simple ou complexes significatifs du document à l'aide de YaTea (la table 3 une comparaison entre concepts proposés par l'évaluation et concepts extraits avec l'outil YaTea). Ces termes seront ensuite utilisés lors de la construction de la matrice de contextes. Les mots vides (prépositions, articles, pronoms, etc.) ont été écartés pour ne conserver que les mots pleins et les nombres (tous normalisés à 0).

i2b2 (2010)	YaTea	i2b2 (2010) ∩ YaTea
19410	11806	1941

TABLE 3: Concepts proposés par l'évaluation et concepts extraits avec l'outil YaTea

3.2 Extration des relations entre les termes à l'aide de patrons lexicaux

Pour un terme donné, on peut tirer certaines de ses relations classiques d'une ressource terminologique. Ceci étant dit, construire des bases de relations manuellement est coûteux et à plus forte raison pour les domaines de spécialité, car cela demande une connaissance préalable du domaine. Les patrons lexicaux ou lexicaux-syntaxiques exploitent une analyse syntaxique préalable du corpus pour constituer automatiquement un réseau lexical sur la base de relations définies bien connues et caractéristiques de la langue ou du domaine.

Nous utilisons donc ici des patrons lexicaux pour la détection de l'inclusion lexicale. Les termes analysés sont les termes générés par YaTea (table 4).

3.3 Expériences et résultats

La première partie du travail a consisté en la construction de nos vecteurs de contexte en utilisant les méthodes PPMI et SVD présentées plus haut. Cette combinaison permet à la fois de pondérer les

Terme complet	community hospital emergency room	chest pressure	blood pressure	chest x-ray
Terme tronqué	hospital emergency room	pressure	pressure	x-ray

TABLE 4: Exemple de sortie du système d'extraction d'inclusion lexicale

vecteurs de contextes et de réduire la dimensionnalité des matrices de co-occurrences (Jurafsky & Martin, 2018). Nous procédons ensuite à l'application de l'inclusion lexicale sur notre corpus : Les termes complexes dans le texte sont remplacés par leurs têtes dans nos contextes grâce aux relations extraites du corpus. Nous comparons ainsi :

— plusieurs dimensions différentes : sans réduction de dimension, puis réduction à 50, 200, 300 dimensions
— corpus avec et sans abstraction de contextes par inclusion lexicale

La seconde partie du travail a consisté en la construction de nos vecteurs de contexte avec des méthodes neuronales Wordvec et FastText. Initialement prévu pour des corpus volumineux, nous avons testé plusieurs paramètres afin d'observer le comportement de nos petits corpus dans cet espace. L'inclusion lexicale est moins triviale sur les modèles neuronaux. En effet, on ne peut pas directement agir sur les contextes et appliquer sur ceux-ci l'inclusion lexicale. Pour contourner cette difficulté, nous avons tout simplement ajouté au corpus de base les phrases sur lesquelles l'inclusion lexicale a été effectuée. Par exemple, si l'on considère le corpus suivant : *"He was admitted to the **community hospital emergency room** where he got a **chest x-ray**"*, notre corpus devient : *"He was admitted to the **community hospital emergency room** where he got a **chest x-ray**. He was admitted to the **hospital emergency room** where he got a **x-ray**".*

Nous comparons ainsi :

— deux modèles d'embeddings différents (Word2Vec et FastText)
— deux architectures différentes : Skip-Gram et CBOW
— plusieurs dimensions différentes : 50, 200, 300
— corpus avec et sans abstraction de contextes par inclusion lexicale

Pour les deux approches (sac de mots et neuronale), nous utilisons comme autres paramètres une fenêtre restreinte de 5 mots et retirons du vocabulaire tout terme apparaissant moins de 4 fois.

3.4 Qualité des représentations obtenues

À l'issue de la construction des vecteurs de contexte et une fois la similarité entre les différents termes calculée, une liste de termes considérés par les modèles comme sémantiquement proches est alors générée. Plusieurs types de relations existent entre ces termes dont des *relations classiques* (Périnet, 2015) : l'hyperonymie (relation entre deux termes du plus général au plus spécifique : *organe/coeur*), la méronymie (un terme désigne une partie d'un second terme, comme : *bras/corps*), la synonymie (des termes similaire : *myocarde/muscle cardiaque* ou ayant un sens proche comme *triste/déprimé*), l'antonymie (des termes contraires : *présent/absent*, *chaud/froid*) ainsi que des liens morphologiques (des relations au niveau de la forme des mots : *changeable/buvable, opéraient/administraient, etc...*). On y trouve aussi des *relations non classiques*, telles que les relations propres au domaine ("*don d'organe*", "*transporteur*"), l'hyponymie ou co-hyponymie (relation entre deux termes du plus spécifique au plus général (*coeur/organe* ou deux termes ayant le même hyponyme *coeur/foie*), des relations nom-verbes (*"infection/se propager"*), des collocations (*"entraîner des conséquences"*), etc.

Pour examiner les représentations que nous avons obtenues, nous comparons un terme sélectionné *"drug abuse"* et ses 7 voisins les plus proches pour chacun des modèles.

PPMI et SVD Comme on peut le voir dans la table 5, les voisins du terme *"drug abuse"* sont assez pertinents : plusieurs relations propres au domaine avec *"prostitution"*, *"recreational"*, *"ivdu"* (*"Intravenous Drug Use"*), l'hyperonyme *"abuse"*, des co-hyponymes avec *"tobacco abuse"* et *"alcohol abuse"* voire même une certaine notion d'antonymie avec *"non-smoker"*.

Modèle	Plus proches voisins						
$PPMI_{\alpha}SVD - Full$	prostitution	smoke history	recrudescence	non-compliance	tobacco abuse	prisoner	tobacco
$PPMI_{\alpha}SVD - 50$	recreational	abuse	domestic	non-smoker	violence	ethanol	tobacco abuse
$PPMI_{\alpha}SVD - 200$	recreational	ethanol	prostitution	0-to-0-pack-year	etoh	abuse	tobacco
$PPMI_{\alpha}SVD - 300$	recreational	prostitution	ethanol	alcohol	0-to-0-pack-year	tobacco	alcohol abuse
$PPMI_{\alpha}SVD_{IL} - Full$	prostitution	smoke history	recrudescence	prisoner	non-compliance	alcohol abuse	aya ma den erinmarg hospital
$PPMI_{\alpha}SVD_{IL} - 50$	recreational	abuse	nonsensitization	non-smoker	ethanol	ivdu	0-to-0-pack-year
$PPMI_{\alpha}SVD_{IL} - 200$	ivdu	recreational	ethanol	prostitution	etoh	ethanol	alcohol
$PPMI_{\alpha}SVD_{IL} - 300$	prostitution	recreational	ivdu	ethanol	alcohol	illicit	etoh

TABLE 5: Mots les plus proches de "*drug abuse*", classés par ordre de proximité, en utilisant la PPMI, puis en combinant PPMI et SVD

Embeddings neuronaux Nous observons dans la table 6 deux tendances principales. Avec Word2Vec, et à plus forte raison avec CBOW, les termes les plus proches du mot cible sont principalement des termes monolexicaux, présentant des relations propres au domaine, comme *"crack"*, *"alcoholism"*, mophologiquement complexes, comme avec *"polysubstance"*. On retrouve aussi l'hyperonyme *"abuse"* et les co-hyponymes *"tobacco abuse"* et *"alcohol abuse"*. Deux liens intéressants, celui de l'overdose suicidaire, est établi entre le mot cible *"drug abuse"* et son voisin *"suicide attempt"*, et celui qui indique la portée récréative des drogues avec *"recreational"* qui apparaît à quelques reprises dans les voisins de notre mot cible.

Avec FastText, les termes sont plus souvent polylexicaux, notamment avec des termes ayant un fort lien morphologique avec notre mot cible, notamment des synonymes, *"substance abuse"* et *"polysubstance abuse"*, des co-hyponymes *"cocaine abuse"*, *"tobacco abuse"* et *"alcohol abuse"*.

Deux liens intéressants ici également, un qui touche à un aspect légal de la consommation de drogue *"illicit"* et ce lien qui lie parfois toxicomanie et SIDA avec *"hiv"*.

Pour l'ensemble des modèles et d'un point de vue purement qualitatif, le voisinage post-inclusion lexical permet principalement l'acquisition de co-hyponymes et de relations propres au domaine.

3.5 Classification des concepts médicaux

L'extraction terminologique terminée, nos vecteurs de mots prêts, nous disposons alors de notre ensemble pour la classification de nos concepts. 3 labels sont donnés par la campagne i2b2/VA 2010 pour cette tâche : *problem, test, treatment*. L'algorithme utilisé est le classifieur SVM. Les SVM sont efficaces dans des espaces multidimensionnels ainsi que sur des corpus de petites tailles. Nous pouvons observer en table 7 que les résultats de la classification pré et post inclusion lexicale avec l'approche en sacs de mots, pondérée avec une PPMI et avec et sans réduction de dimensionnalité. Les résultats nous montrent des résultats quasi égaux, avant ou après inclusion lexicale.

Pour les méthodes neuronales, la table 8 montre les résultats de cette classification avant inclusion lexicale. FastText dépasse à peine Word2Vec avec l'approche Skip-Gram. Les performances de l'approche CBOW sont assez clairement inférieures.

La table 9 montre que les meilleurs résultats après inclusion lexicale améliorent à peine les résultats que ce soit pour l'approche CBOW ou Skip-Gram (3 *runs* ont été effectués avec des résultats variables en moyenne de l'ordre de ± 0.1).

Modèle	Plus proches voisins						
$W2V-50-sg$	abuse	0-pack	tobacco abuse	suicide attempt	crack	widowed	alcoholism
$W2V-200-sg$	0-pack	abuse	suicide attempt	alcoholism	reflux disease	tobacco history	cig
$W2V-300-sg$	abuse	suicide attempt	0-pack-year	non-smoker	crack	reflux disease	tobacco history
$W2V_{IL}-50-sg$	recreational	suicide attempt	socially	polysubstance	crack	user	reflux disease
$W2V_{IL}-200-sg$	suicide attempt	crack	ex	user	bronchitis	recreational	tobacco history
$W2V_{IL}-300-sg$	recreational	suicide attempt	socially	ingestion	0-pack-year	alcohol history	tob
$W2V-50-cbow$	cigarette	depression	etoh	alcohol abuse	0-pack	migraine	alcohol
$W2V-200-cbow$	radiation exposure	colonic	depression	tb	abuse	sexually	social
$W2V-300-cbow$	tobacco	cigarette	alcoholism	social	alcohol	smoking	0-pack
$W2V_{IL}-50-cbow$	drug	polysubstance	etoh	seizure	abuse	suicide	illicit
$W2V_{IL}-200-cbow$	drug	etoh	illicit	contrast allergy	seizure	recreational	tobacco history
$W2V_{IL}-300-cbow$	contrast allergy	reflux disease	counselor	anxiety	alcohol history	ketoacidosis	alcohol abuse
$FT-50-sg$	tobacco abuse	abuse	alcohol abuse	illicit	tobacco user	ethanol abuse	cocaine abuse
$FT-200-sg$	abuse	tobacco	drug	illicit	etoh	alcohol	alcoholism
$FT-300-sg$	abuse	tobacco	drug	alcohol	illicit	etoh	smoker
$FT_{IL}-50-sg$	abuse	ethanol abuse	tobacco abuse	cocaine abuse	alcohol abuse	illicit	tobacco user
$FT_{IL}-200-sg$	abuse	ethanol abuse	tobacco abuse	cocaine abuse	alcohol abuse	polysubstance abuse	illicit
$FT_{IL}-300-sg$	abuse	tobacco abuse	ethanol abuse	cocaine abuse	illicit	alcohol abuse	tobacco user
$FT-50-cbow$	substance abuse	cocaine abuse	tobacco abuse	ethanol abuse	passage	quincy	hhnk
$FT-200-cbow$	cocaine abuse	tobacco abuse	ethanol abuse	pan	dyspesia	message	substance abuse
$FT-300-cbow$	substance abuse	page	tobacco abuse	onset dyspnea	passage	cocaine abuse	dyspepsia
$FT_{IL}-50-cbow$	ethanol abuse	tobacco abuse	cocaine abuse	abuse	alcohol abuse	tobacco	hiv
$FT_{IL}-200-cbow$	ethanol abuse	tobacco abuse	cocaine abuse	tobacco	etoh	dlco	tobacco user
$FT_{IL}-300-cbow$	ethanol abuse	tobacco abuse	cocaine abuse	abuse	hiv	hillpa	pamela

TABLE 6: Mots les plus proches de "*drug abuse*", classés par ordre de proximité, avec différentes configurations d'embeddings neuronaux

Modèle	$PPMI_\alpha$							
Spécificité	Sans IL				Avec IL			
Dimensions	Sans SVD	50	200	300	Sans SVD	50	200	300
Précision	0,75	0,69	0,64	0,66	**0,76**	0,70	0,66	0,68
Recall	**0,80**	0,72	0,66	0,67	**0,80**	0,72	0,67	0,68
F-mesure	0,76	0,69	0,62	0,66	**0,77**	0,71	0,66	0,68

TABLE 7: Résultat de classification des modèles obtenus en appliquant la $PPMI_\alpha$ sans SVD et avec SVD avec une dimension de 50, 200, 300, avec et sans abstraction des contextes et avec macro mesure

Modèle	Word2Vec						FastText					
Architecture	Skip-Gram			CBOW			Skip-Gram			CBOW		
Dimensions	50	200	300	50	200	300	50	200	300	50	200	300
Précision	0,79	0,77	0,79	0,73	0,73	0,76	**0,8**	0,79	0,79	0,76	0,77	0,77
Recall	0,78	0,78	0,79	0,72	0,73	0,77	**0,8**	0,78	0,78	0,75	0,76	0,76
F-mesure	0,79	0,78	0,79	0,72	0,73	0,76	**0,8**	0,79	0,78	0,76	0,76	0,77

TABLE 8: Résultat de classification pour les modèles Word2Vec et FastText sans inclusion lexicale, mais avec dédoublement du corpus avec fenêtre de 5 et avec macro mesure

Modèle	Word2Vec_IL						FastText_IL					
Architecture	Skip-Gram			CBOW			Skip-Gram			CBOW		
Dimensions	50	200	300	50	200	300	50	200	300	50	200	300
Précision	**0,79**	**0,79**	**0,79**	**0,79**	0,78	0,78	**0,79**	**0,79**	**0,79**	0,78	0,77	0,77
Recall	0,79	0,78	0,79	0,79	0,78	0,78	**0,8**	0,78	0,78	0,79	0,78	0,77
F-mesure	**0,79**	**0,79**	**0,79**	**0,79**	0,78	0,78	**0,79**	**0,79**	0,78	0,78	0,77	0,77

TABLE 9: Résultat de classification pour les modèles Word2Vec et FastText avec inclusion lexicale, avec fenêtre de 5 et avec macro mesure

4 Conclusion et perspectives

Les méthodes distributionnelles se sont largement généralisées avec la multiplication des données textuelles disponibles. Construire des bases de relations manuellement est très coûteux et les méthodes distributionnelles permettent de faire reposer le calcul du sens sur l'analyse statistique des contextes des mots dans un corpus sans avoir besoin de recourir à des sources extérieures de connaissance. Les approches neuronales sont venues renouveler les performances de ces méthodes.

Dans cet article, nous avons présenté les méthodes classiques et neuronales les plus communes pour la mise en oeuvre des modèles de sémantique distributionnelle. Nous avons listé les principales raisons pour lesquelles ses méthodes n'étaient pas adaptées corpus de spécialité et qu'il fallait les adapter pour qu'ils prennent en compte les particularités structurelles et linguistiques de ces derniers. Nous avons énoncé le problème de dispersion des données ainsi que d'importance des termes, notamment complexes, dans l'étude des textes en langue de spécialité. Nous avons ensuite présenté des ébauches de travaux sur lesquels nous avons commencé à nous pencher à savoir l'étude de l'impact de la généralisation des contextes, notamment ici de l'inclusion lexicale, sur la sélection de contextes significatifs pour nos mots cibles.

Nous avons montré que d'un point de vue qualitatif, les relations obtenus entre les unités terminologiques propre au domaine était plus homogènes et que cela augmentait très légèrement les résultats de classification. Globalement, les résultats obtenus par les modèles à base de sac de mots et ceux obtenus par approche neuronale avec l'architecture Skip-Gram ne diffèrent pas de façon significative. Cependant, l'inclusion harmonise et hisse légèrement les résultats des méthodes neuronales avec l'architecture CBOW au niveau des autres. Notre tâche consistant en l'amélioration endogène des contextes distributionnels en domaine de spécialité, nous avons pour le moment écarté tout recours vers l'utilisation de corpus, de ressources ou de modèles pré-entraînés (e.g. *ELMo* et *BERT*) venant enrichir nos modèles distributionnels.

Ces expériences préliminaires ouvrent des nombreuses perspectives. Premièrement, prendre en compte un plus grand nombre de relations (synonymie, etc.) et ne pas se contenter uniquement de remplacer un terme par sa tête (inclusion lexicale, hypéronymie). Deuxièmement, déterminer une manière de sélectionner les contextes les plus significatifs en amont de l'inclusion lexicale (avec par, exemple, la mesure du $Cf \times Itf$ proposée par Périnet (2015)). Troisièmement, nous avons vu qu'appliquer l'abstraction des contextes directement sur les outils Word2Vec et FastText n'étaient pas trivial et nous avons contourné le problème en gardant à la fois la phrase originale et celle sur laquelle a été effectuée une inclusion lexicale. Quatrièmement, il serait intéressant d'explorer d'autres façons d'acquérir les unités terminologiques propres au domaine sur lesquelles nous basons notre travail. Finalement, l'évaluation des représentations étant une tâche très ardue, il serait intéressant d'évaluer nos représentations sur une autre tâche comme l'alignement lexical bilingue, par exemple.

Remerciements

Ce travail a été réalisé dans le cadre du projet ANR ADDICTE (Analyse Distributionnelle en Domaine de Spécialité) soutenu par l'Agence Nationale de Recherche sous la référence ANR-17-CE23-0001.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In T. SALAKOSKI, F. GINTER, S. PYYSALO & T. PAHIKKALA, Eds., *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, p. 380–387 : Springer.

BENGIO Y., DUCHARME R., VINCENT P. & JANVIN C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, **3**, 1137–1155.

BOJANOWSKI P., GRAVE E., JOULIN A. & MIKOLOV T. (2016). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, **abs/1607.04606**.

CHURCH K. W. & HANKS P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, **16**(1), 22–29.

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, p. 160–167, New York, NY, USA : ACM.

DEERWESTER S., T. DUMAIS S., FURNAS G., LANDAUER T. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, 391–407.

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

FABRE C., HATHOUT N., SAJOUS F. & TANGUY L. (2014). Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille. In *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, p. 266–279, Marseille, France.

GRABAR N. & ZWEIGENBAUM P. (2004). Lexically-based terminology structuring. *Terminology*, **10**(1), 23–54.

HAZEM A. & DAILLE B. (2018). Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan : European Language Resource Association.

JURAFSKY D. & MARTIN J. H. (2018). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA : Prentice Hall PTR, 1st edition.

LEVY O., GOLDBERG Y. & DAGAN I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, **3**, 211–225.

MANNING C. D., RAGHAVAN P. & SCHÜTZE H. (2008). *Introduction to Information Retrieval*. Cambridge, UK : Cambridge University Press.

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

MITCHELL J. & LAPATA M. (2010). Composition in distributional models of semantics. *Cognitive Science*, **34**(8), 1388–1429.

MORIN E. (2007). Apport des termes complexes à l'acquisition lexicale multilingue à partir de corpus comparables spécialisés : entre intuition et réalité. In *7ème rencontres Terminologies et Intelligence Artificielle (TIA'07)*, p. 11–20, Sophia Antipolis, France.

PÉRINET A. (2015). *Distributional analysis applied to specialized corpora : reduction of data sparsity through context abstraction*. Theses, Université Paris 13 ; Laboratoire d'Informatique Médicale et d'Ingénieurie des Connaissances en e-Santé.

PETERS M. E., NEUMANN M., IYYER M., GARDNER M., CLARK C., LEE K. & ZETTLEMOYER L. (2018). Deep contextualized word representations. In M. A. WALKER, H. JI & A. STENT, Eds., *NAACL-HLT*, p. 2227–2237 : Association for Computational Linguistics.

UZUNER O., SOUTH B., SHEN S. & DUVALL S. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, **18**, 552–6.

ZWEIGENBAUM P. & HABERT B. (2006). Faire se rencontrer les parallèles : regards croisés sur l'acquisition lexicale monolingue et multilingue. In *Glottopol : Traitements automatisés des corpus spécialisés : contextes et sens*.