Représentation de termes complexes dans un espace vectoriel relié à une ontologie pour une tâche de catégorisation.

A. Ferré[1,2]

1 MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
2 LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France

INRA de Jouy en Josas, Domaine de Vilvert, 78352 Jouy-en-Josas
arnaud.ferre@universite-paris-saclay.fr

Résumé

Nous proposons dans cet article une méthode semi-supervisée pour étiqueter des termes de textes par les concepts d'une ontologie. La méthode génère des représentations vectorielles continues des termes complexes dans un espace sémantiquement structuré par l'ontologie du domaine. La méthode proposée s'appuie sur une approche de sémantique distributionnelle, qui génère des vecteurs initiaux pour chacun des termes extraits. Ces vecteurs sont alors plongés dans l'espace vectoriel construit à partir de la structure de l'ontologie. Ce plongement s'effectue par entraînement d'un modèle linéaire. Un calcul de distance est enfin effectué pour déterminer la proximité entre vecteurs de termes et vecteurs de concepts pour déterminer l'étiquetage ontologique des termes. Nous avons évalué la qualité de ces représentations pour une tâche de catégorisation en utilisant les concepts d'une ontologie comme étiquettes sémantiques. La performance de cette méthode atteint l'état de l'art pour cette tâche de catégorisation ouvrant d'encourageantes perspectives.

Mots Clef

TAL, extraction d'information, étiquetage de texte par une ontologie, espace vectoriel, sémantique distributionnelle, modèle linéaire

Abstract

We propose in this paper a semi-supervised method for labeling terms of texts by concepts of an ontology. The method generates continuous vector representations of complex terms in a semantically space structured by the ontology of the domain. The proposed method relies on a distributional semantics approach, which generates initial vectors for each of the extracted terms. Then these vectors are embedded in the vector space constructed from the structure of the ontology. This embedding is carried out by training a linear model. Finally, we apply a distance calculation to determine the proximity between vector of terms and vector of concepts and thus to determine the ontological tags of terms. We have evaluated the quality of these representations for a categorization task by using the concepts of an ontology as semantic labels. The performance of this method for the state of the art for this task of standardization opening up encouraging prospects.

Keywords

NLP, information extraction, ontology-based text tagging, vector space, distributional semantics, multivariate linear regression

1 Introduction

Beaucoup des connaissances du domaine biomédical ou biologique sont sous une forme non-structurée, comme celle exprimée dans les articles scientifiques [1]. Pour les experts de ces domaines, l'augmentation conséquente de la littérature spécialisée a créé un besoin important en méthodes automatiques d'extraction d'information [2]. La tâche de catégorisation est une des tâches principales pour répondre à ce besoin.

La catégorisation consiste à annoter des termes (mono- ou multi-mots) des textes avec une ou plusieurs catégories sémantiques (e.g. un terme extrait d'un corpus tel que « *children greater than 9 years of age who had lower respiratory illness* » pourrait être catégorisé par une catégorie sémantique ayant pour label « *pediatric patient* » et/ou « *patient with disease* »). Les concepts d'une ontologie peuvent être utilisés pour représenter ces catégories sémantiques de façon formelle et structurée. La catégorisation rencontre plusieurs difficultés, comme la variabilité importante de la morphologie des termes, qu'ils soient représentés par un mot ou par plusieurs [3]. Les termes multi-mots qui présentent des structures morphosyntaxiques variées et des imbrications complexes, tels que les groupes nominaux complexes (*complex noun phrases*) sont particulièrement difficiles à étiqueter par des catégories. Or, dans les textes de la littérature spécialisée, tels que les articles scientifiques en science du vivant, les groupes nominaux complexes sont abondant [4]. Une approche basée sur la similarité morphologique entre terme et étiquette sémantique apparaît limitée pour effectuer cette tâche [5], parce que la morphologie des labels des concepts n'est pas nécessairement proche de la morphologie des termes à annoter. Une autre difficulté vient du nombre important de catégories sémantiques utilisées, rendant une approche par classification supervisée coûteuse en annotation manuelle (plus de 2000 catégories par exemple dans l'ontologie des habitats bactériens OntoBiotope [6]).

Une alternative consiste à calculer la proximité sémantique entre des termes par sémantique distributionnelle. C'est une approche fondée sur la corrélation entre la similarité de

sens et la similarité de distribution des unités sémantiques (mot, combinaison de mots, phrase, documents, ...) [7], [8]. Une unité sémantique peut être représentée par un vecteur construit à partir de la distribution des informations de contexte dans lesquels elle est trouvée. La proximité des vecteurs dans cet espace est alors transposable à une proximité sémantique [9]. Il existe aujourd'hui de nombreuses méthodes de génération de tels espaces vectoriels, tel que Word2Vec [10], mais celles-ci se concentrent habituellement sur les jeux de données massifs [11] dans lesquels l'information est relativement répétée.

La question qui nous intéresse ici est : comment utiliser la sémantique distributionnelle pour catégoriser les termes par une ontologie, autrement dit comment relier l'information distributionnelle aux catégories d'une ontologie. Dans le cadre de la littérature spécialisée qui nous intéresse ici, la question se focalise sur des relativement petits corpus et un grand nombre de catégories sémantiques.

Nous proposons une méthode originale dans laquelle nous représentons des termes complexes basés sur un « plongement de mots » (*word embedding*), en représentant une ontologie sous forme d'espace vectoriel et en entraînant une transformation de vecteurs de termes en vecteurs de concepts. Ensuite, cette transformation est utilisée pour déterminer le concept le plus approprié pour chaque terme extrait.

2 Matériel

Les données utilisées sont celles de la tâche de catégorisation Bacteria Biotope (tâche 3) de BioNLP Shared Task en 2016 [12]. Les documents sont des références de la MEDLINE [13], composées de titres et de résumés d'articles scientifiques dans le domaine de la biologie. La tâche consiste, étant donné les entités du corpus dénotant les habitats bactériens, à leur assigner une catégorie de l'ontologie OntoBiotope. Le corpus (noté BB dans la suite) est divisé en trois : le corpus d'entraînement, le corpus de développement et le corpus de test. Dans les corpus d'entraînement et de développement les catégories des termes sont données : elles nous ont servi à entraîner notre méthode. Le corpus de test est celui pour lequel les catégories sont à prédire : il nous sert à évaluer notre méthode pour la tâche de catégorisation. Les entités de chacun de ces corpus ont été annotées manuellement. Voici un résumé de leurs caractéristiques :

	Entraîn.	Dév.	Test	**Total**
Documents	71	36	54	**161**
Mots	16 295	8 890	13 797	**38 982**
Entités	747	454	720	**1 921**
Entités distinctes	476	267	478	**1 125**
Cat. sémantiques	825	535	861	**2 221**
Cat. distinctes	210	122	177	**329**

TABLE 1 : Statistiques descriptives du corpus BB

En plus de ce corpus, nous avons utilisé un corpus élargi du même domaine pour générer des représentations vectorielles de chaque mot. Il est composé de 100 000 phrases venant de titres et de résumés d'articles scientifiques dans le domaine de la biologie disponibles sur PubMed. Cela représente un corpus de taille relativement petit, qui contient une majorité de mots non-outils avec une faible fréquence d'apparition (cf. *TABLE 2*).

Répétés >2	72 412	35%
Répétés 2 fois	31 569	15%
Non répétés	105 364	50%
Total mots non-outils	209 345	100%

TABLE 2 : Statistiques descriptives du corpus élargi

3 Méthode

3.1 Génération de vecteurs de mots

L'espace vectoriel des termes (EVT) est obtenu en générant un vecteur pour chacun des mots du corpus élargi qui comprend également les corpus d'entraînement et de développement, mais pas le corpus test. Pour cela, nous avons utilisé l'outil Word2Vec [10] en prenant pour contexte d'un mot, tous les mots contenus dans la phrase. Pour avoir suffisamment de données d'entraînement pour la génération de vecteurs de mots, et aussi pour éviter de prendre en compte des fautes de frappes ou des erreurs, il est habituellement conseillé d'utiliser Word2Vec sans les mots peu fréquents, n'apparaissant qu'une ou deux fois dans tout le corpus. Notre corpus contenant beaucoup de mots d'intérêt à faible fréquence, nous avons fait le choix de ne pas appliquer de seuil de fréquence. Après quelques tests de performance, la dimension 200 a été choisie pour les vecteurs de sorties (cf. FIGURE 1A), ce qui est du même ordre de grandeur que ce qui est conseillé habituellement [10].

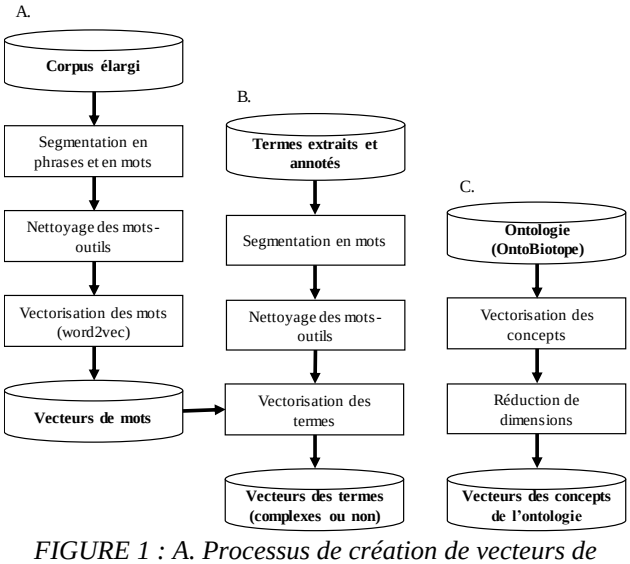

FIGURE 1 : A. Processus de création de vecteurs de mots. B. Processus de création des vecteurs de termes C. Processus de création de vecteurs de concepts

3.2 Génération de vecteurs de termes complexes

Pour calculer les représentations vectorielles des termes composés extraits des corpus (cf. FIGURE 1B), on

commence par les segmenter en mots. Pour chaque mot non-outil, on utilise le vecteur calculé par Word2Vec. Le vecteur du terme composé est obtenu par la moyenne des vecteurs des mots qui le composent :

$$v_{t_k} = \sum_{i=1}^{n_k} v_{m_i^k}/n_k \qquad (1)$$

Où v_{t_k} est le vecteur associé au terme t_k, n_k est le nombre de mots non-outils du terme t_k, $v_{m_i^k}$ est le vecteur du mot m_i^k issu de Word2Vec, et le terme t_k est tel que :

$$\forall i \in [1, n_k], m_i^k \in t_k \qquad (2)$$

3.3 Génération de vecteurs de concepts

Pour construire les vecteurs de concepts et donc un espace vectoriel ontologique (EVO), on initialise des vecteurs nuls possédant autant de dimension que de concepts dans l'ontologie. Chaque valeur du vecteur correspond donc à un des concepts de l'ontologie. La valeur est à 1 si le concept est un ancêtre du concept considéré, à 0 sinon :

$$v_{c_k} = (w_{c_k}^0, \ldots, w_{c_k}^i, \ldots, w_{c_k}^n) \qquad (3)$$

Où v_{c_k} est le vecteur associé au concept c_k, c_k est relié à la k^{ème} dimension des vecteurs (i.e. w_c^k), n est le nombre de concepts dans l'ontologie et $w_{c_k}^i$ est la valeur du vecteur v_{c_k} pour la dimension i, tel que :

$$w_{c_k}^i = \begin{cases} 1 & si\ i = k \\ 1 & si\ c_i\ parent\ (direct\ ou\ non)\ de\ c_k \\ 0 & sinon \end{cases} \qquad (4)$$

Cette représentation a comme intérêt de conserver les distances (distance cosinus) attendues entre les concepts (cf. FIGURE 2 et TABLE 3) : un concept est plus proche de ses fils, puis de ses parents.

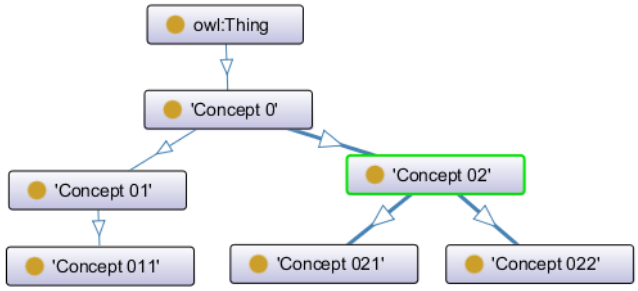

FIGURE 2 : Représentation conceptuelle d'une ontologie

Concept 02	Similarité
Concept 02	1,00
Concept 021	0,82
Concept 022	0,82
Concept 0	0,71
Concept 01	0,50
Concept 011	0,41

TABLE 3: Similarité cosinus du concept 02 avec ses concepts voisins

On remarque que la dimension de l'EVO généré a alors pour taille le nombre de concepts de l'ontologie (soit plus de 2000 pour l'ontologie OntoBiotope). Pour décomposer la transformation EVT->EVO recherchée, on peut commencer par réduire le nombre de dimension de l'EVO pour atteindre le même nombre de dimensions que l'EVT généré précédemment (cf. FIGURE 1C). Une analyse en composantes principales (ACP) et un positionnement multidimensionnel (MDS) ont été testés.

3.4 Entraînement (modèle linéaire)

L'objectif de la phase d'entraînement est de déterminer une transformation de l'EVT vers l'EVO qui minimise la distance entre les vecteurs de termes issus de cette transformation et les vecteurs des concepts associés. Nous avons choisi de nous limiter à une transformation linéaire et avons entraîné un algorithme de type modèle linéaire avec les corpus annotés d'entraînement et de développement (cf. FIGURE 3).

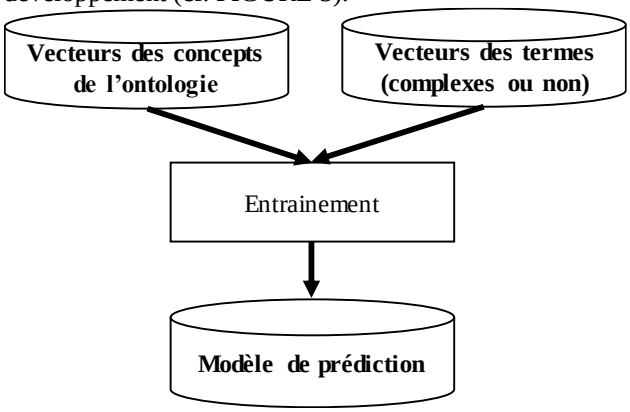

FIGURE 3 : Processus d'entraînement pour déterminer une transformation de l'EVT vers l'EVO

Les matrices de transformations obtenues permettent alors de prédire de nouveaux vecteurs associés aux termes du corpus de test exprimés dans l'EVO. Pour répondre à la tâche d'évaluation, on recherche le vecteur de concept le plus proche en terme de distance cosinus. Le concept ainsi trouvé est celui qui est assigné au terme (cf. FIGURE 4).

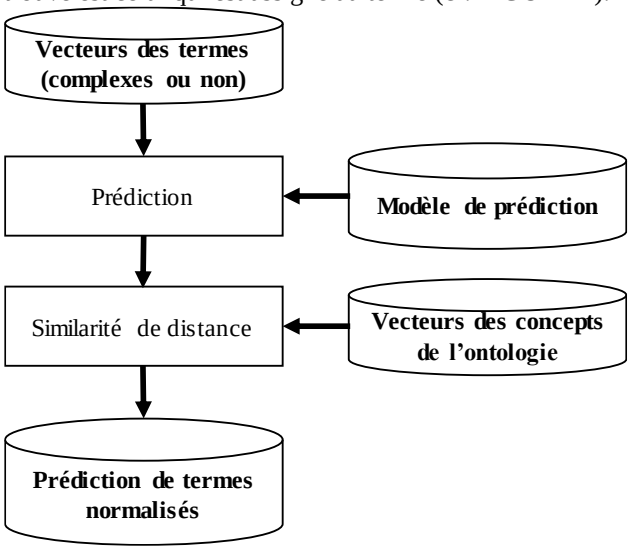

FIGURE 4 : Processus de prédiction des catégories sémantiques associées aux termes extraits

L'objectif de la phase d'entraînement est de déterminer une transformation de l'EVT vers l'EVO qui minimise toutes les distances entre les vecteurs des termes résultants dans

l'EVO et les vecteurs des concepts associés. Dans cet article, une transformation linéaire est étudiée car nous faisons l'hypothèse qu'il y a une certaine similitude de répartition entre les vecteurs de termes dans l'EVT et les vecteurs de concepts associés dans l'EVO. Autrement dit, une transformation non-linéaire pourrait fortement déformer la répartition des vecteurs de termes dans l'EVO pour s'adapter aux données d'entraînement peu nombreuses et ne recouvrant qu'une faible partie des annotations pouvant être détectées dans les textes ciblés.

Cet entraînement vise à obtenir les meilleurs paramètres pour approximer l'équation matricielle suivante :

$$Y = X.B + U \quad (5)$$

Où Y est une matrice formée d'une série de vecteurs de concepts, X est une matrice formée d'une série de vecteurs de termes (où la i^{ème} ligne de X représente le vecteur d'un terme qui est annoté par un concept qui a pour vecteur la i^{ème} ligne de Y), B est la matrice contenant les paramètres qui sont à estimer, et U est une matrice contenant une distribution gaussienne multivariée. Cet entraînement est réalisé sur les corpus d'entraînement et de développement (cf. FIGURE 3).

La matrice obtenue nous permet de concevoir une fonction de transformation linéaire, afin de permettre de prédire de nouveaux vecteurs associés aux termes du corpus d'essai exprimé dans l'EVO :

$$f: \begin{pmatrix} EVT \rightarrow EVO \\ v_{\text{term}} \rightarrow v'_{\text{term}} = f(v_{\text{term}}) \end{pmatrix} \quad (6)$$

Où v_{term} est un vecteur de terme dans l'EVT et v'_{term} est le vecteur résultant du même terme projeté dans l'EVO. Pour satisfaire aux exigences de la tâche d'évaluation, le vecteur de concept le plus proche (en terme de distance cosinus) de v'_{term} est choisi pour le terme annoté (cf. la FIGURE 4).

4 Résultats

4.1 Catégorisation

Pour évaluer la performance des systèmes participants à la tâche 3 de BB, une mesure de similarité sémantique est implémentée sur le site du challenge BioNLP-ST 2016. La mesure utilisée est celle définie par Wang et al. en 2007 [14], avec le paramètre de poids à 0.65. Avec cette mesure, nous pouvons calculer une *baseline* en attribuant à tous les termes le concept « bacteria habitat », qui est la racine de la hiérarchie de l'ontologie OntoBiotope. La mesure trouvée est alors de 32.17%.

Catégorisation	**Score final de similarité**
BOUN	0.62
CONTES	**0.60**
LIMSI	0.44

TABLE 4 : Résultats de la tâche de catégorisation de BioNLP-ST 2016

Deux équipes avaient participé à cette tâche de BioNLP-ST 2016 et avaient obtenu les résultats rapportés dans la *TABLE 4*. Notre méthode (CONTES - *CONcept-TErm System*) a obtenu un résultat de 60%, tout à fait comparable à celui de la première équipe et significativement au-dessus de la méthode du LIMSI qui s'appuyait sur une approche morphologique.

4.1 Vecteurs de termes extraits

En dépit de la faible fréquence d'apparition des mots du corpus élargi (cf. TABLE 2), les vecteurs de mots obtenus présentent des proximités relativement satisfaisantes du point de vue de la similarité sémantique des termes associés. De plus, la méthode utilisée pour former des vecteurs pour les termes complexes semblent elle aussi satisfaisante comme le montre l'exemple suivant :

cell	**Similarité**
HCE cell	0,99
13C-labeled cell	0,99
parietal cell	0,99
Schwann cell	0,99
CD8+ T cell	0,98
PMN cell	0,97
macrophage cell	0,95

TABLE 5 : Termes à proximité du terme "cell"

Il semble également que des différences morphologiques n'empêchent pas l'agglomération de vecteurs de termes proches (cf. TABLE 6 et TABLE 7), ce qui était une des propriétés recherchées.

younger ones	**Similarité**
children less than five years of age	0,81
children less than 2 years of age	0,81
children less than two years of age	0,80

TABLE 6 : Termes à proximité du terme "younger ones"

seawater	**Similarité**
sediments	0,77
sediment sample from a disease-free fish farm	0,75
fish farm sediments	0,73
subterranean brine	0,73
lagoon on the outskirts of the city of Cagliari	0,71
petroleum reservoir	0,71
marine environments	0,71
marine bivalves	0,69
sediment samples from diseased farms	0,69
urine sediments	0,68
petroleum	0,66
subterranean environment	0,65
fresh water	0,65
fresh water supply	0,64
Seafood	0,64
marine	0,64

TABLE 7 : Termes à proximité du terme "seawater"

Néanmoins, la cooccurrence de certains mots semble agglomérer certains termes de catégorie différente. 2 mots apparaissant fréquemment dans des contextes communs se

retrouvent alors avec des vecteurs similaires. Cette similarité persiste alors également lors du calcul des vecteurs de termes. C'est par exemple le cas pour les termes relatifs au poisson et ceux relatifs aux fermes d'élevage de poissons (cf. TABLE 8). Ces représentations sont moins satisfaisantes car elles ne permettent pas de différencier les catégories sémantiques sous-jacentes.

fish	**Similarité**
fish farming	*0,98*
fish farm	*0,92*
disease-free fish farm	*0,91*
fish farm sediments	*0,87*
healthy fish	*0,81*

TABLE 8 : Termes les plus proches du terme "fish"

4.2 Vecteurs de concepts de l'ontologie

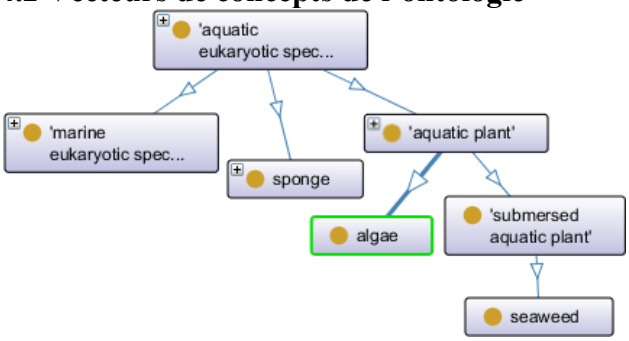

FIGURE 5 : Taxonomie des concepts autour du concept "algae" (visualisé avec le logiciel Protégé)

On peut estimer la qualité des vecteurs de concepts créés en observant la cohérence entre la proximité vecteur/vecteur et leur sens (cf. FIGURE 5 et TABLE 9). Une réduction de dimension semble détériorer progressivement l'EVO et impacte parfois ces distances de manière suffisante pour empêcher la distinction entre concepts parents, concepts fils et concepts frères (cf. FIGURE 5 et TABLE 10). Comme conséquence, à partir de l'EVO réduit, il semble difficile de restituer parfaitement la structure de l'ontologie. Prenons l'exemple du concept '*algae*' :

<OBT:001922: *algae*> sans ACP	**Similarité**
<OBT:001777: *aquatic plant*>	0,93
<OBT:001895: *submersed aquatic plant*>	0,86
<OBT:001967: seaweed>	0,80

TABLE 9 : Similarité pour le concept algue de l'ontologie (sans ACP)

<OBT:001922: *algae*> avec ACP	**Similarité**
<OBT:001777: *aquatic plant*>	0,99
<OBT:001895: *submersed aquatic plant*>	0,99
<OBT:001967: *seaweed*>	0,99
<OBT:000372: *sponge*>	0,93
<OBT:000269: *marine eukaryotic species*>	0,93

TABLE 10 : Similarité pour le concept algue de l'ontologie (avec ACP de dimension finale 300)

En comparant plusieurs exemples, il semble que l'ACP ne modifie pas l'ordre de proximité des concepts, mais on peut observer des augmentations de densité de ces vecteurs (cf. comparaison entre le TABLE 9 et TABLE 10). Cela semble cohérent du fait de la diminution de l'espace concerné. En effet, l'ordre des plus proches voisins d'un vecteur de concept ne semble pas modifié.

4.3 Influence de la dimension de l'EVT

Word2Vec permet l'utilisation de 2 architectures différentes pour générer des vecteurs associés à des mots d'un corpus : Continuous Bag Of Words (CBOW) et Skip-Gram. L'architecture CBOW peut être défini par l'objectif de prédire un mot en fonction de son contexte, alors que le Skip-Gram est de prédire le contexte en fonction d'un mot en entrée. Quelque soit l'architecture utilisée, Word2Vec permet de générer des vecteurs associés à chaque mot du corpus. Nous avons testé les 2 architectures sur des dimensions de vecteurs de sortie différentes (cf. FIGURE 6). Pour des espaces vectoriels générés avec une dimension entre 100 et 250, les scores finaux semblent relativement stables, particulièrement pour le CBOW. De même, l'écart de score entre les 2 architectures restent en dessous des 3%. Au dessus d'une dimension de 250, on assiste à une diminution du score pour les 2 architectures, avec une pente plus importante avec le CBOW.

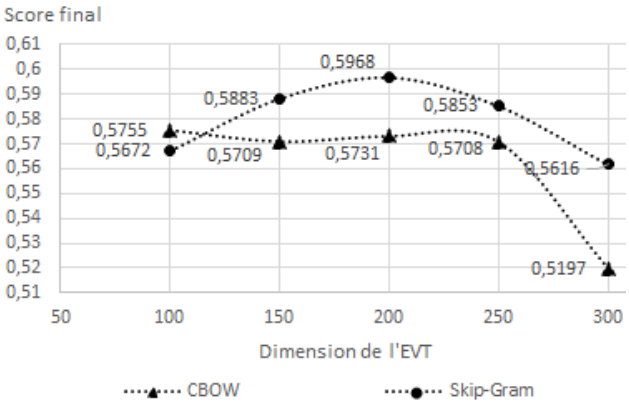

FIGURE 6 : Comparaison des architectures CBOW et Skip-Gram sur différentes dimensions d'EVT

4.4 Influence d'une réduction de dimension de l'EVO sur la catégorisation

L'EVO possède une dimension importante par rapport à l'information spécifique qui y est codée (i.e. la structure de l'ontologie). Cela peut poser des difficultés d'ordre combinatoire mais aussi théorique : une projection linéaire de l'EVT sur l'EVO (de dimension plus grande que l'EVT) ne devrait alors se faire que sur un sous-espace de l'EVO, limitant ainsi les résultats. Il était donc intéressant d'étudier l'impact d'une réduction de dimension de l'EVO sur le score final. On peut alors observer qu'une réduction par analyse en composantes principales - ACP (avec des résultats similaires avec un positionnement multidimensionnel - MDS) diminue systématiquement le score obtenu par rapport à l'utilisation de l'EVO non-réduit (cf. FIGURE 7). Néanmoins, on observe également un

palier d'une certaine performance (moins de 3% en dessous du score sans réduction) jusqu'à un certain point où le score diminue drastiquement.

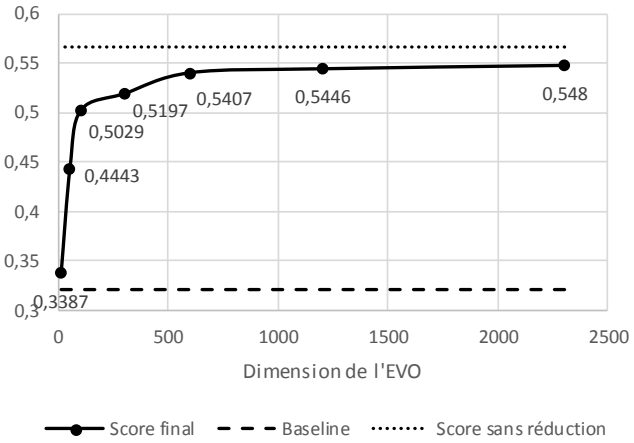

FIGURE 7 : Impact de la réduction de l'EVO (pour un EVT généré de dimension 100 avec Skip-Gram)

5 Discussions

Pour étendre les interprétations issues d'exemples, il serait intéressant d'évaluer la qualité globale des espaces vectoriels générés : espace vectoriel des mots, des termes, des concepts ainsi que l'espace finale contenant les transformations des vecteurs des termes.

Un des plafonds de la méthode présentée dans cet article est dû au fait que, pour cette tâche de catégorisation, un terme peut être catégorisé par plusieurs concepts de l'ontologie (ex : le terme « school age children with wheezing illness » devrait être catégorisé par le concept <OBT:002307: pediatric patient> ainsi que le concept <OBT:002187: patient with disease>), mais c'est également le cas des autres systèmes participants.

6 Perspectives

Pour de futurs travaux, il serait pertinent d'appliquer des méthodes d'évaluation globale de la qualité des espaces vectoriels générés. En particulier, cela permettrait d'évaluer plus exhaustivement les processus intermédiaires et d'observer avec plus de précision l'impact des modifications sur leurs paramètres internes. De nouvelles méthodes plus élaborées pourraient alors être envisagées pour améliorer les résultats. Par exemple, il serait certainement positif d'utiliser une méthode de représentation vectorielle d'une ontologie qui génèrerait un espace possédant une dimension plus faible tout en conservant la possibilité de discerner la structure initiale de l'ontologie. De même, la méthode utilisée ici pour générer les vecteurs de l'EVT pourrait être améliorée pour prendre en compte le contexte syntaxique des termes. Cela pourrait résoudre les problèmes de similarité sémantique entre '*fish'* et '*fish farm'* (cf. TABLE 8).

Il arrive fréquemment que des termes doivent être annotés par plusieurs concepts de l'ontologie cible (par exemple : '*children greater than 9 years of age who had lower respiratory illness*' devait ici être annoté par le concept <OBT:002307: *pediatric patient*> et par le concept <OBT:002187: *patient with disease*>). Avoir à disposition une ontologie complètement définie possédant explicitement tous les concepts résultant de l'intersection possible de ses autres concepts (par exemple : un concept '*pediatric patient with disease*' qui est un sous-ensemble de <OBT:002307: *pediatric patient*> et de <OBT:002187: *patient with disease*>) devrait améliorer les résultats. Si de telles ontologies semblent relativement rares dans le domaine biologique, il pourrait être intéressant de commencer par générer automatiquement tous les concepts équivalents à l'intersection des concepts non-disjoints pour répondre à ce problème. Néanmoins, si les concepts partagent de nombreuses intersections entre eux ou que le caractère disjoint n'a pas été formalisé, la taille de l'ontologie générée risque de poser des difficultés combinatoires.

Il est relativement peu commun d'avoir pour données initiales les termes extraits (ainsi que leur appartenance à des entités nommées d'intérêt). Les méthodes d'extraction terminologique possédant des performances relativement acceptables, il serait intéressant d'en utiliser en amont de la tâche actuelle.

Enfin, malgré la limitation inhérente des méthodes de catégorisation basées sur la morphologie des mots, celles-ci pourraient néanmoins être utilisées pour effectuer une pré-catégorisation du corpus. En conséquence, on pourrait envisager d'utiliser ces annotations pour entraîner la méthode au lieu d'utiliser une annotation manuelle. Ainsi, cela transformerait cette méthode en une méthode non-supervisée.

7 Conclusion

L'objectif de cet article était de proposer une approche pour la création de représentations vectorielles pour des termes (complexes ou non) dans un espace sémantique. De plus, il visait à proposer une méthode capable de s'adapter à un corpus spécialisé de petite taille où les termes d'intérêts apparaissent avec une fréquence relativement faible. Les méthodes les plus utilisées actuellement génèrent des espaces vectoriels dont il est difficile d'interpréter le sens autrement qu'en terme de proximité spatiale/similarité sémantique. Notre méthode semble montrer qu'en combinant des approches relativement classiques, il est possible d'utiliser une ontologie pour générer des vecteurs dans un espace vectoriel plus interprétable. Les résultats comparables à ceux de l'état de l'art, semble ouvrir d'encourageantes perspectives. Au-delà de la tâche de catégorisation, de nouvelles méthodes performantes de génération d'espace vectoriel interprétables pourraient également répondre à de nombreuses problématiques.

Remerciements

This work is supported by the "IDI 2015" project funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Bibliographie

[1] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, et J. A. Kors, « Using rule-based natural language processing to improve disease normalization in biomedical text », *J. Am. Med. Inform. Assoc.*, vol. 20, nº 5, p. 876‑881, sept. 2013.

[2] S. Ananiadou et J. McNaught, Éd., *Text mining for biology and biomedicine*. Boston: Artech House, 2006.

[3] A. Nazarenko, C. Nédellec, E. Alphonse, S. Aubin, T. Hamon, et A.-P. Manine, « Semantic annotation in the alvis project », in *International Workshop on Intelligent Information Access (IIIA)*, 2006, p. 5–pages.

[4] F. Maniez, « Prémodification et coordination : quelques problèmes de traduction des groupes nominaux complexes en anglais médical », *ASp*, nº 51‑52, p. 71‑94, déc. 2007.

[5] W. Golik, P. Warnier, et C. Nédellec, « Corpus-based extension of termino-ontology by linguistic analysis: a use case in biomedical event extraction », in *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, 2011, p. 37–39.

[6] R. Bossy, W. Golik, Z. Ratkovic, D. Valsamou, P. Bessières, et C. Nédellec, « Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task », *BMC Bioinformatics*, vol. 16, nº 10, p. S1, 2015.

[7] J. R. Firth, « The technique of semantics », 1957.

[8] Z. S. Harris, « Distributional Structure », WORD, vol. 10, nº 2‑3, p. 146‑162, août 1954.

[9] C. Fabre et A. Lenci, « Distributional Semantics Today Introduction to the special issue », *Trait. Autom. Lang.*, vol. 56, nº 2, p. 7–20, 2015.

[10] T. Mikolov, K. Chen, G. Corrado, et J. Dean, « Efficient estimation of word representations in vector space », *ArXiv Prepr. ArXiv13013781*, 2013.

[11] C. Fabre, N. Hathout, F. Sajous, et L. Tanguy, « Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille », in *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, 2014, p. 266–279.

[12] L. Deléger, E. Chaix, M. Ba, A. Ferré, P. Bessières, et C. Nédellec, « Overview of the Bacteria Biotope Task at BioNLP Shared Task 2016 », 2016. [En ligne]. Disponible sur: https://aclweb.org/anthology/W/W16/W16-3002.pdf. [Consulté le: 30-mars-2017].

[13] MEDLINE, « Base de données bibliographique MEDLINE : https://www.ncbi.nlm.nih.gov/pubmed ». .

[14] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, et C.-F. Chen, « A new method to measure the semantic similarity of GO terms », *Bioinformatics*, vol. 23, nº 10, p. 1274‑1281, mai 2007.