Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Caractérisation des relations sémantiques entre termes multi-mots fondée sur l'analogie

Yizhe Wang[1] Béatrice Daille[2] Nabil Hathout[1]
(1) CLLE, CNRS & Université Toulouse Jean Jaurès, France
(2) LN2S, CNRS & Université de Nantes, France
yizhe.wang@univ-tlse2.fr, beatrice.daille@ln2s.fr,
nabil.hathout@univ-tlse2.fr

RÉSUMÉ

La terminologie d'un domaine rend compte de la structure du domaine grâce aux relations entre ses termes. Dans cet article, nous nous intéressons à la caractérisation des relations terminologiques qui existent entre termes multi-mots (MWT) dans les espaces vectoriels distributionnels. Nous avons constitué un jeu de données composé de MWT en français du domaine de l'environnement, reliés par des relations sémantiques lexicales. Nous présentons une expérience dans laquelle ces relations sémantiques entre MWT sont caractérisées au moyen de l'analogie. Les résultats obtenus permettent d'envisager un processus automatique pour aider à la structuration des terminologies.

ABSTRACT

Semantic relations recognition between multi-word terms by means of analogy

Terminologies reflect the structure of specialised domains with the help of internally labelled relations between terms. In this article, we are interested in the recognition of conceptual relations between multi-word terms (MWTs) in vector space models. We created a dataset of semantically related MWTs of the environmental field. We present an experiment where we characterize lexical semantic relations between MWTs by means of analogy. The results show that our method can be used as a first step towards an automatic process for structuring terminology.

MOTS-CLÉS : relations terminologiques, termes multi-mots, analogie, projection sémantique.

KEYWORDS: terminological relations, multi-word terms, analogy, semantic projection.

1 Introduction

Le terme en tant qu'étiquette d'un concept d'un domaine s'inscrit dans un système linguistique spécialisé où il est mis en relation avec d'autres termes. La notion de relation joue un rôle important pour la construction des ressources terminologiques. Les dictionnaires spécialisés récents, les banques et les bases de données terminologiques, recensent les termes d'un domaine et rendre compte de l'organisation des concepts et des relations conceptuelles qui s'établissent entre eux. Les recherches actuelles sur l'identification des relations entre les termes se sont concentrées sur les termes simples (SWT) (Grabar & Hamon, 2006; Zhu *et al.*, 2016; Zhang *et al.*, 2017). Peu de travaux ont porté sur l'acquisition des relations entre MWT. Les travaux sur les relations entre MWT concernent principalement l'exploitation de la structure interne des MWT en utilisant différents types d'informations linguistiques, syntaxiques (Verspoor *et al.*, 2003) et sémantiques (Hazem & Daille, 2018).

Volume 1 : conférence principale, pages 115–124.

L'analogie est une relation proportionnelle entre deux couples de d'objets qui sont dans la même relation (Lepage & Shin-ichi, 1996; Lepage, 1998; Skousen, 2002; Claveau & L'Homme, 2005; Turney, 2008; Langlais *et al.*, 2009). Elle connaît un regain d'intérêt à la suite des travaux de Mikolov *et al.* (2013b) qui ont montré sa capacité à capter certaines relations linguistiques dans les espaces vectoriels distributionnels. Dans ce même cadre, Gladkova *et al.* (2016) ont étudié la capacité de l'analogie à capter les relations sémantiques lexicales. Plus récemment, Allen & Hospedales (2019) ont proposé une explication mathématique de l'existence des analogies dans les plongements statiques.

Dans cet article, nous nous intéressons à la caractérisation des relations entre MWT dans les espaces vectoriels distributionnels au moyen de l'analogie. Nous considérons uniquement dans ce travail les termes qui contiennent deux mots lexicaux, mais notre méthode pourrait être adaptée à des MWT de plus de deux mots. Trois groupes de relation nous intéressent : (1) ANTI qui regroupe la relation contraire et la relation contrastive ; (2) HYP qui réunit l'hyperonymie et hyponymie ; (3) QSYN, composé de la synonymie, de la quasi-synonymie et de la co-hyponymie. Notre étude s'appuie sur un jeu de données composés de couples de MWT du domaine de l'environnement, sémantiquement reliés. Ce jeu de données est construit par projection sémantique, une méthode fondée sur l'hypothèse que les MWT ont un sens compositionnel dont l'une des conséquences est que les relations sémantiques entre les SWT sont préservées dans les MWT qui les contiennent (Hamon & Nazarenko, 2001). Une annotation manuelle la préservation de ces relations a été effectuée pour les couples de MWT du jeu de données. Notre étude se distingue des travaux existants sur l'analogie entre mots ou entre termes simples (Drozd *et al.*, 2016; Liu *et al.*, 2017; Koehl *et al.*, 2020) par le fait qu'elle porte sur des relations sémantiques lexicales (hyperonymie, antonymie, synonymie) entre MWT et entre SWT comme dans *froid*:*chaud*::*air froid*:*air chaud*. Les résultats obtenus montrent que l'analogie peut être utilisée pour prédire ces relations et qu'elle fonctionne mieux pour les relations symétriques comme l'antonymie et la synonymie que pour les relations asymétriques comme l'hyperonymie et l'hyperonymie. L'analogie dans l'espace vectoriel permet réciproquement de valider les projections sémantiques des SWT.

Dans la suite de cet article, la section 2 présente un bref état de l'art sur l'acquisition des relations sémantiques au moyen de l'analogie. La section 3 présente les ressources utilisées et la construction du jeu de données. La section 4 décrit le modèle utilisé et la méthode d'évaluation. Les résultats obtenus et l'analyse des résultats sont présentés en section 5. Nous concluons notre travail et présentons les perspectives envisagées pour des travaux futurs en section 6.

2 Caractérisation des relations sémantiques par l'analogie

Les recherches actuelles sur l'analogie avec les plongements de mots se concentrent sur l'« analogie proportionnelle » du type $a : b :: c : d$ (Drozd *et al.*, 2016). Le point de départ de notre étude est (Mikolov *et al.*, 2013a) qui montre que les relations entre les mots peuvent être captées dans une large mesure par soustraction entre vecteurs distributionnels : $a - b \approx c - d$. Ainsi, les solutions d'une équation analogique $a : b :: c :?$ se trouvent parmi les vecteurs d similaires au vecteur $b - a + c$. Plus précisément, la solution de l'équation serait alors : $\mathrm{argmax}_{d \in V}(sim(d, c - a + b))$ où sim est une mesure de similarité entre vecteurs, généralement cos. Cette méthode est habituellement appelée *3cosADD*. Une autre méthode, appelée *PairDirection*, et plus fidèle à la formule initiale peut également être utilisée (Mikolov *et al.*, 2013b). La solution est alors : $\mathrm{argmax}_{d \in V}(sim(d - c, b - a))$. Levy & Goldberg (2014) ont montré que *PairDirection* est supérieure à *3cosADD* pour résoudre les

analogies syntaxiques.

Suite à (Mikolov *et al.*, 2013a), plusieurs études ont été menées sur l'évaluation quantitative de l'analogie afin de déterminer la capacité des méthodes analogiques à capter différents types de relations linguistiques (Gladkova *et al.*, 2016; Köper *et al.*, 2015). Les résultats montrent que les relations lexicales comme la synonymie sont les plus difficiles à capter parmi toutes les relations évaluées. L'analogie proportionnelle a également été utilisée par Liu *et al.* (2017) pour apprendre des plongements multi-relationnels. Dans ce travail, l'analogie est comparée à différents types de modèles sur un jeu de données composé de mots connectés par différebtes relations lexicales. Les auteurs montrent que l'analogie surpasse les autres modèles avec un rang réciproque moyen (MRR) de $0,942$ (voir section 4).

L'analogie a aussi été utilisée pour l'acquisition de relations sémantiques entre termes. Chen *et al.* (2018) l'utilisent pour identifier les relations entre les termes médicaux et Nooralahzadeh *et al.* (2018) pour trouver les antonymes et les synonymes de termes du domaine du pétrole et du gaz. Si la plupart des travaux concernent l'analogie entre les mots ou les termes simples, certaines études portent sur les relations entre termes complexes. C'est le cas de Xu *et al.* (2018) dans le domaine de la biologie qui s'intéressent à des analogies comme *carbon 14 atom*:*radioactivity*::*C4 plant*:*C4 photosynthesis*. Les travaux précédents montrent que la capacité de l'analogie à capter les relation lexicales dépend de la qualité des représentations et de la prise en compte des variations lexicales et sémantiques dans les espaces vectoriels (Hamilton *et al.*, 2016; Vu Xuan *et al.*, 2019; Saha *et al.*, 2020), de la méthode utilisée pour résoudre les équations analogiques, des relations elles-mêmes et des caractéristiques des jeux de données comme SAT (Turney *et al.*, 2003), Google (Mikolov *et al.*, 2013a) ou BATS (Gladkova *et al.*, 2016).

Notre étude se distingue de celles évoquées ci-dessus poar le fait que nous travaillons sur des termes du français du domaine de l'environnement. Comme Xu *et al.* (2018), nous travaillons aussi avec des MWT. Cependant, nos données sont plus spécifiques : (*i*) dans les analogies que nous considérons, deux SWT sont des composants de deux MWT ; (*ii*) la relation entre les MWT est identifiée en faisant l'hypothèse que s'il existe une analogie $SWT_1 : SWT_2 :: MWT_1 : MWT_2$, alors la relation entre MWT_1 et MWT_2 est la même que celle qui existe entre SWT_1 et SWT_2. La projection sémantique est une méthode basée sur l'hypothèse que le sens des MWT est compositionnel. Une conséquence de cette hypothèse est que pour deux termes multi-mots MWT_1 et MWT_2 qui ne diffèrent que par un de leurs constituants, notons C_1 celui qui apparaît dans MWT_1 et C_2 celui qui apparaît dans MWT_2 (par exemple, lorsque $MWT_1 = C_0$ prep det C_1 et $MWT_2 = C_0$ prep det C_2), la relation sémantique entre MWT_1 et MWT_2 est la même que celle qui existe entre C_1 et C_2 dans la mesure où la contribution de la partie partagée (C_0 prep det) aux sens des deux MWT est identique.

3 Matériel expérimental

Corpus. Nous avons utilisé le corpus monolingue français PANACEA Environnement (ELRA-W0065) [1]. Il se compose de 35 453 documents (environ 50 millions de mots) de différents niveaux de spécialisation. Trois opérations ont été réalisées sur le corpus : extraction du texte à partir des documents XML ; conversion en UTF-8 ; lemmatisation.

1. http://catalog.elra.info/en-us/repository/browse/ELRA-W0065/

SWT amorces. Le jeu de données est construit par projection sémantique à partir d'une liste de référence de SWT reliés par des relations sémantiques lexicales. Cette liste que nous désignerons par *RefProj* contient 831 couples de SWT nominaux et adjectivaux (116 couples ANTI ; 191 couples HYP ; 524 couples QSYN) extraits de la ressource proposée par Bernier-Colborne & Drouin (2016). *RefProj* contient des couples de termes comme : *conservation*:*protection*, *combustible*:*oil*, *flore*:*faune*.

Jeu de données. Pour générer le jeu de données [2], nous avons tout d'abord extrait du corpus PANACEA les candidats-termes qui contiennent deux mots lexicaux en utilisant l'extracteur TermSuite (Cram & Daille, 2016). La méthode de projection sémantique a ensuite été appliquée sur les couples de SWT de *RefProj* et filtrée par les candidats-termes extraits permettant ainsi d'étendre à ces derniers les relations qui existent entre les termes simples. La relation contrastive entre *flore* et *faune* peut par exemple être étendue aux termes *protection de la flore* et *protection de la faune*. Le statut de terme des candidats a été validé au moyen de trois banques terminologiques en ligne : TERMIUM Plus [3] ; Le Grand Dictionnaire [4] ; IATE [5]. À l'issue de cette validation, le jeu de données se compose de 231 couples de MWT qui se répartissent comme suit : 80 couples ANTI, 51 couples HYP et 100 couples QSYN. Une annotation manuelle de ces données a été effectuée par trois annotateurs qui ont indiqué si la relation sémantique qui existe entre les SWT est ou n'est pas préservée entre les MWT sur la base de 5 contextes extraits aléatoirement du corpus pour chaque MWT. L'accord inter-annotateurs de 0,69, ce qui reste assez fort. Une phase d'adjudication a ensuite été réalisée pour créer le jeu de données.

Les 231 couples de MWT permettent de construire 231 quadruplets formé d'un couple de SWT et d'un couple de MWT. La relations sémantique existant entre les SWT est préservée entre les MWT dans 181 de ces quadruplets (classe positive) comme pour *sec*:*humide*::*climat sec*:*climat humide* ; la classe négative est composée de 50 quadruplets comme *autoroute*:*route*::*autoroute maritime*:*route maritime* où les SWT sont dans une relation d'hyperonymie tandis que les MWT sont synonymes. La taille réduite du jeu de données s'explique par le fait que de nombreux termes extraits, comme *conservation du papillon*, sont trop spécifiques pour être présents dans les banques terminologiques que nous avons utilisées.

Le tableau 3 présente un extrait du jeu de données. *SWT1*, *SWT2* sont les deux termes simples, *MWT1* et *MWT2* les MWT qui les contiennent, *Rel* la relation entre les SWT. *Anno* indique si la relation est préservée (1) ou si elle ne l'est pas (0).

SWT1	SWT2	MWT1	MWT2	Rel	Anno
froid	chaud	air froid	air chaud	ANTI	1
piscicole	agricole	domaine piscicole	domaine agricole	ANTI	1
terre	planète	climat de la terre	climat de la planète	HYP	0
culture	agriculture	culture biologique	agriculture biologique	HYP	1
neige	glace	cristaux de neige	cristaux de glace	QSYN	0
protection	conservation	protection de l'espace	conservation de l'espace	QSYN	1

2. Le jeu de données et les ressources utilisées sont disponibles à l'adresse suivante: https://github.com/YizWang/List-of-semantically-linked-MWTs
3. https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra
4. http://www.granddictionnaire.com/
5. https://iate.europa.eu/

4 Expériences

Modèle. Le modèle utilisé pour identifier les analogies entre les représentations vectorielles des SWT et des MWT est construit avec FastText (Joulin *et al.*, 2017). Il se distingue par la possibilité qu'il offre de construire des modèles de plongements statiques qui combinent à la fois les mots et leurs n-grammes de caractères (Vu *et al.*, 2019), ce que Word2Vec ou Glove ne permettent pas. Cette caractéristique peut être exploitée pour construire des modèles qui contiennent en même temps des représentations pour les SWT et les MWT et dans lesquels les représentations des MWT sont indépendantes de celles des SWT qu'ils contiennent. Cela est essentiel car nous voulons comparer dans le même espace les différences entre les représentations des SWT et entre celles des MWT. Les représentations de ces derniers ne doivent donc pas être calculées compositionnellement à partir de celles de leurs constituants car les deux différences seraient alors toujours égales.

Pour créer de tels modèles, nous avons annoté dans le corpus les occurrences des MWT afin que FastText calcule leurs représentations et celles des mots qu'ils contiennent. Par exemple, les occurrences du MWT *air froid* interviennent dans le calcul des représentations de *air*, de *froid* et de *air_froid*. En pratique, nous avons remplacé dans le corpus toutes les occurrences de `air froid` par `< air froid air_froid >` et nous avons forcé FastText à ne procéder à aucun autre découpage de mots en positionnant le paramètre `-maxn` à 0. Ainsi, les représentations des MWT et celles des SWT sont générées de manière séparées dans le même espace sémantique.

Par ailleurs, la réalisation d'un MWT dans le corpus peut prendre différentes formes. Par exemple, *changement du climat* a comme variante attestée *changement climatique*. Nous avons observé que le prise en compte des variantes des MWT n'a pas d'incidence notable sur les résultats de la tâche et avons décidé de ne pas les annoter comme des occurrences des MWT.

Si les performances des plongements de mots peuvent être considérablement affectées par les hyperparamètres (Levy *et al.*, 2015), l'optimisation de l'exactitude moyenne sur un ensemble de relations hétérogènes peut ne pas être significative (Gladkova *et al.*, 2016). Par conséquent, nous n'avons pas réalisé d'optimisation des hyperparamètres. Les paramètres utilisés sont : dim=100, min_count=3, maxn=0, window_size=5, model=skipgram, epoch=20, lr=0.05 et les valeurs par défaut pour tous les autres paramètres. Par ailleurs, notre travail ne portant que sur les termes simples et multi-mots, nous avons restreint le vocabulaire du modèle aux 3 254 termes qui sont présents dans le corpus.

Évaluation. Dans cette expérience, nous nous intéressons à la classe positive, car notre objectif est de caractériser les relations entre les MWT. De ce fait, nous avons évalué notre modèle en utilisant la mesure MRR (*Mean Reciprocal Rank*) (Radev *et al.*, 2002; Chowdhury, 2010), la précision, le rappel et la F-mesure aux rangs 1, 5 et 10. Rappelons que MMR est définie comme suit :

$$MRR = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rang_i}$$

où W représente la liste des réponses positives et $Rang_i$ correspond au rang du i-ième candidat qui appartient à W. Par ailleurs, chaque couple de MWT donne lieu à deux analogies, $SWT_1 : SWT_2 :: MWT_1 :?$ et $SWT_2 : SWT_1 :: MWT_2 :?$; les résultats présentés ci-dessous sont la moyenne des deux tests pour les relations symétriques (ANTI et QSYN). Pour les relations d'hyponymie (resp. d'hyperonymie), les résultats sont calculés séparément pour les MWT hyponymes et les MWT hyperonymes.

5 Résultats et analyses

L'expérience a été réalisée pour les deux méthodes : *3cosADD* et *PairDirection*. Les résultats obtenus étant meilleurs avec *3cosADD*, c'est cette méthode qui a été choisie pour résoudre les équations analogiques.

Résultats de la prédiction. Les résultats présentés dans la Table 1 montrent que l'analogie est assez efficace dans la prédiction des relations entre les MWT avec une MRR de 0,692. Les Tables 2 et 3 présentent les résultats obtenus pour chacune des relations. Elles montrent notamment que les relations hiérarchiques sont les plus difficiles à capter. Ces résultats sont conformes à ceux de Gladkova *et al.* (2016). Nous pouvons également observer que la MRR de l'hyperonymie est légèrement supérieure à celle de l'hyponymie, ou en d'autres termes qu'il est plus facile de prédire l'hyperonyme à partir des hyponymes que l'inverse. Par exemple, pour un quatruplet *combustible : pétrole :: gaz de combustible : gaz de pétrole*, il est plus facile de prédire *gaz de combustible* en prenant *gaz de pétrole* comme MWT inconnu au lieu de prédire *gaz de pétrole* en prenant *gaz de combustible* comme MWT inconnu.

MRR	P1	R1	F1	P5	R5	F5	P10	R10	F10
0,692	0,828	0,572	0,677	0,819	0,722	0,767	0,814	0,796	0,805

TABLE 1 – Évaluation de la capacité de l'analogie à la capture des relations sémantiques lexicales entre MWT

	ANTI	QSYN	Hyponyme	Hyperonyme
MRR	0,683	0,745	0,508	0,567

TABLE 2 – Résultats pour chaque type de relation estimés par la mesure MRR

	P1	R1	F1	P5	R5	F5	P10	R10	F10
ANTI	0,904	0,558	0,690	0,896	0,717	0,797	0,888	0,790	0,836
QSYN	0,875	0,653	0,747	0,868	0,771	0,816	0,863	0,828	0,845
Hypo	0,444	0,296	0,356	0,526	0,556	0,541	0,559	0,704	0,623
Hypero	0,526	0,370	0,435	0,552	0,593	0,571	0,559	0,704	0,623

TABLE 3 – Précision, rappel et F-mesure pour chaque type de relation sémantique lexicale

Pour la classe négative en revanche, les résultats sont insuffisants avec une F-mesure de 0,377 pour le candidat de rang 1. L'analogie dans les espaces vectoriels distributionnels n'est donc pas adaptée à l'identification des quadruplets dans lesquels la relation entre les SWT n'est pas préservée entre les MWT. Nous avons d'autre part estimé la généralité du modèle en élargissant le vocabulaire à l'ensemble de mots du corpus (120 606 mots). Le modèle prédit alors la classe positive avec un MRR de 0,6118. Ce score bien qu'inférieur à celui obtenu avec un vocabulaire limité aux termes démontre que la capacité de l'analogie à capter les relations sémantiques reste acceptable lorsque le vocabulaire est élargi.

Analyse d'erreurs. Une analyse manuelle des erreurs parmi les voisins du Top 5 a été réalisée pour identifier les quadruplets difficiles à prédire [6]. Globalement, les erreurs peuvent être due à la manière dont les plongements sont calculés ou à la méthode utilisée pour prédire les analogies. Nous avons notamment observé que l'une des causes d'erreurs principales est la polysémie des termes simples qui affecte toutes les relations sémantiques comme dans le cas de *route* qui peut désigner une infrastructure (*route terrestre*) ou une région de l'espace (*route maritime*). Le modèle s'avère également peu sensible aux variations sémantiques déterminées par le contexte comme dans le cas d'*agriculture durable* qui en discours peut être utilisé comme un hyperonyme d'*élevage durable* ou comme un antonyme de ce dernier lorsque la relation est induite par le contraste entre les plantes et les animaux.

Les difficultés de l'analogie à capter certaines relations lexicales comme l'hyperonymie explique les résultats obtenus peuvent encore être améliorés, comme l'observent Gladkova *et al.* (2016) sur un ensemble de relations plus varié. Une autre difficulté rencontrée par l'analogie est également due au fait que lorsque l'on soustrait la représentation vectorielle d'un mot M_1 à celle d'un autre mot M_2, la différence ne représente pas le sens d'un mot et qu'elle ne capte que de manière approximative la relation lexicale sémantique qui s'établie entre M_1 et M_2 (Vylomova *et al.*, 2016).

6 Conclusion et perspectives

Cet article présente une première étude de la caractérisation des relations sémantiques entre termes multi-mots dans des espaces sémantiques. Nous avons constitué un jeu de donnée composé de couples de MWT du domaine de l'environnement reliés sémantiquement. Nous avons étudié la capacité de l'analogie à identifier les relations sémantiques lexicales fondamentales entre les MWT dans les espaces vectoriels distributionnels. Ses performances sont globalement bonnes avec des différences marquées entre les relations lexicales considérées et un meilleur comportement pour les relations synmétriques. L'analyse d'erreurs fait ressortir la polysémie comme l'une des causes principales des mauvaises prédictions. La prochaine étape de ce travail est l'adaptation de cette méthode à des modèles plus sensibles au contexte comme BERT (Devlin *et al.*, 2019) pour prédire les relations sémantiques lexicales entre les MWT et la préservation des relations entre SWT qu'ils contiennent.

Références

ALLEN C. & HOSPEDALES T. (2019). Analogies explained: Towards understanding word embeddings.

BERNIER-COLBORNE G. & DROUIN P. (2016). Evaluation des modeles sémantiques distributionnels: le cas de la dérivation syntaxique. In *Proceedings the 23rd French Conference on Natural Language Processing (TALN)*, p. 125–138.

CHEN Z., HE Z., LIU X. & BIAN J. (2018). Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC medical informatics and decision making*, **18**(2), 65.

CHOWDHURY G. G. (2010). *Introduction to modern information retrieval*. Facet publishing.

6. Rappelons que l'annotation du jeu de donnée est basée sur 5 contextes extraits aléatoirement tandis que la représentation vectorielle des termes est calculée à partir de l'ensemble des contextes dans lesquels ils apparaissent.

CLAVEAU V. & L'HOMME M.-C. (2005). Apprentissage par analogie pour la structuration de terminologie. Utilisation comparée de ressources endogènes et exogènes. In *Actes de la conférence terminologie et intelligence artificielle (TIA-2005)*, Rouen.

CRAM D. & DAILLE B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, p. 13–18.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

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

DROZD A., GLADKOVA A. & MATSUOKA S. (2016). Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, p. 3519–3530.

GLADKOVA A., DROZD A. & MATSUOKA S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, p. 8–15.

GRABAR N. & HAMON T. (2006). Terminology structuring through the derivational morphology. In *International Conference on Natural Language Processing (in Finland)*, p. 652–663: Springer.

HAMILTON W. L., CLARK K., LESKOVEC J. & JURAFSKY D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, p. 595: NIH Public Access.

HAMON T. & NAZARENKO A. (2001). Detection of synonymy links between terms: experiment and results. *Recent advances in computational terminology*, **2**, 185–208.

HAZEM A. & DAILLE B. (2018). Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

KOEHL D., DAVIS C., NAIR U. & RAMACHANDRAN R. (2020). Analogy-based assessment of domain-specific word embeddings. In *2020 SoutheastCon*, p. 1–6: IEEE.

KÖPER M., SCHEIBLE C. & IM WALDE S. S. (2015). Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th international conference on computational semantics*, p. 40–45.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis: ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In *Actes de la 14e conférence sur le traitement automatique des langues naturelles (TALN 2007)*, p. 101–110, Toulouse.

LANGLAIS P., YVON F. & ZWEIGENBAUM P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conférence*

of the European Chapter of the ACL (EACL 2009), p. 487–495, Athens, Greece: Association for Computational Linguistics.

LEPAGE Y. (1998). Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*, volume 2, p. 728–735, Montréal.

LEPAGE Y. & SHIN-ICHI A. (1996). Saussurian analogy: A theoretical account and its application. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, volume 2, p. 717–722, Copenhagen.

LEVY O. & GOLDBERG Y. (2014). Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, p. 171–180.

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.

LIU H., WU Y. & YANG Y. (2017). Analogical inference for multi-relational embeddings.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013b). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, p. 746–751.

NOORALAHZADEH F., ØVRELID L. & LØNNING J. T. (2018). Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

RADEV D. R., QI H., WU H. & FAN W. (2002). Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, p. 1153–1156, Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA).

SAHA B., LISBOA S. & GHOSH S. (2020). Understanding patient complaint characteristics using contextual clinical bert embeddings.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conférence sur le traitement automatique des langues naturelles (TALN 2007)*, p. 401–410, Toulouse.

SKOUSEN R., Éd. (2002). *Analogical Modeling. An exemplar-based approach to language*. Volume 10 de Human Cognitive Processing. Amsterdam / Philadelphia: John Benjamins Publishing Company.

TURNEY P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of The 22nd International Conference on Computational Linguistics (COLING 2008)*, p. 905–912, Manchester.

TURNEY P. D., LITTMAN M. L., BIGHAM J. & SHNAYDER V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. *CoRR*, **cs.CL/0309035**, 482–489.

VERSPOOR C. M., JOSLYN C. & PAPCUN G. J. (2003). The gene ontology as a source of lexical semantic knowledge for a biological natural language processing application. In *SIGIR workshop on Text Analysis and Search for Bioinformatics*, p. 51–56.

VU X.-S., VU T., TRAN S. N. & JIANG L. (2019). Etnlp: a visual-aided systematic approach to select pre-trained embeddings for a downstream task.

VU XUAN S., VU T., TRAN S. & JIANG L. (2019). ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, p. 1285–1294, Varna, Bulgaria: INCOMA Ltd. DOI : 10.26615/978-954-452-056-4_147.

VYLOMOVA E., RIMELL L., COHN T. & BALDWIN T. (2016). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1671–1682, Berlin, Germany: Association for Computational Linguistics. DOI : 10.18653/v1/P16-1158.

XU J., AUNG H. L. & WEERAWARDHENA S. (2018). Solving biology analogies with deep learning.

ZHANG L., LI J. & WANG C. (2017). Automatic synonym extraction using word2vec and spectral clustering. In *2017 36th Chinese Control Conference (CCC)*, p. 5629–5632: IEEE.

ZHU W., ZHANG W., LI G.-Z., HE C. & ZHANG L. (2016). A study of damp-heat syndrome classification using word2vec and tf-idf. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, p. 1415–1420: IEEE.