Exploring synonymy relation between multi-word terms in distributional semantic models

Yizhe Wang[1], Béatrice Daille[2], Nabil Hathout[3]

[1] University of Science and Technology of China
[2]LN2S, CNRS & University of Nantes
[3]CLLE, CNRS & University of Toulouse2

Abstract

Terminology describes the knowledge structure of a domain through the relationships between its terms. However, relations between multi-word terms (MWTs) are often underrepresented in terminology resources. Moreover, most of the work on this issue concerns the relations between simple terms (STs). In this paper, we explore the ability of distributional semantic models (DSMs) to capture synonymy between MWTs by lexical substitution based and analogy based methods. We evaluated our methods on the English and French MWTs of the environmental domain. Our experiments show that the results obtained using analogy in static word embeddings are globally better than the ones obtained using lexical substitution in pre-trained contextual models.

Keywords:

1. Introduction

The demand for structured terminological resources is strong, especially for extracting and acquiring information from texts. Terminology resources gather the terms of a domain and describe the relations that exist between them, such as synonymy. While multi-word terms (MWTs) are widely represented in terminology, the relationships between them are often missing. Synonymy is an important relation in terminology. It has been the subject of several studies in which a variety of methods have been proposed, including methods based on syntactic patterns, multilingual methods and distributional methods. However, most of these studies concern single terms (STs) and very few focus on the acquisition of synonymy between MWTs. Works on synonymy between MWTs in the literature often explore the internal structure of MWTs using different types of linguistic information, especially semantic information. In this paper, we explore the ability of distributional semantic models (DSMs) to capture synonymy between MWTs in the environmental domain in English and French. Our study focuses on nominal MWTs composed of two lexical words (i.e., biterms). Two methods are proposed. The first is based on lexical substitution using a masked language model (MLM). The second captures synonymy through analogy between STs and MWTs representations in a FastText (Bojanowski et al., 2017) model. Both methods are tested on two datasets of English and French MWT synonyms of the environment domain extracted from the IATE translation dictionary.

Section 2. presents related works on the acquisition of relations between words and terms. Our methods are introduced in Section 3.. Section 4. presents the resources we used to create our data set and to perform the experiments. Section 5. describes the implementation of the methods. We present and discuss the results obtained in Sections 6.. Section 7. concludes the article and presents future avenues of research.

2. Relates work

We propose two methods for identifying synonymy between MWTs: lexical substitution and analogy. Lexical substitution is a task that aims at predicting candidate words that can replace a target word in a given context. In recent studies, lexical substitution has often been used to acquire semantic relations and is usually performed using masked language models (MLMs). For example, Schick and Schütze (2020) and Arefyev et al. (2020) use Transformer models to test the ability of these models to capture lexical relations between words from the general domain without any task-specific optimization. Their results show that BERT is able to capture relational semantic properties and that most of the returned substitutes are synonyms and co-hyponyms when the masked word is a noun. This observation confirms that of Ferret (2021). The way we use lexical substitution is close to the ones presented in these works. However, our method differs in several respects. We seek to identify lexical semantic relations between MWTs in French (in addition to English) in the environmental domain, whereas the work presented focuses on relations between single words in English in the general domain. Moreover, we use contexts extracted from corpora and not patterns that express these relations as Schick and Schütze (2020) do. In addition, we use a conditioning strategy that allows us to provide the model with additional information about the masked word, but in a different way than Arefyev et al. (2020) (cf. Section 3.).

Analogy is a method we use for detecting whether two pairs of words are in the same relation. The study of Mikolov et al. (2013) shows that analogy is able to capture linguistic relations in vector space models and that the identification of these relations can be estimated by the offset between their distributional representations ($V_a - V_b \approx V_c - V_d$ for an analogical quadruplet $a : b :: c : d$). In line with Mikolov et al. (2013), many studies have focused on the ability of analogy to capture various lexical, encyclopedic, or specialized domain relations (Gladkova et al., 2016; Chen et al., 2018; Wohlgenannt et al., 2019). While most studies focus on word analogy between single terms, Chaudhri et al. (2022) focus on solving analogous equations between single and multi-word terms in the biology domain in English in order to capture domain-specific relationships like *a type of*. The study of Paullada et al. (2020) also focuses on analogy between STs and MWTs in the biomedical domain. Their objective is to acquire domain-specific relations, such as gene-disease relations. The authors created a DSM from a corpus of sentences extracted from the biomedical literature and annotated with syntactic dependencies. They show that embeddings that incorporate syntactic information do improve the resolution of biomedical analogy equations.

Our study differs from the ones we have just presented in several respects. As we already pointed out, we are working on the identification of terminological relations between MWTs in English and French in the environmental domain. We are interested in classical lexical semantic relations between MWTs and not in domain-specific ones. Like Paullada et al. (2020), we use vector offset instead of seq2seq and seq2vec models as do Chaudhri et al. (2022) to solve analogy equations. However, the model we use is different from that of Paullada et al. (2020) because we use a FastText model where the MWTs and their components are represented in the same vector space.

3. Methods

In this section, we describe in detail our methods for acquiring synonymy between MWTs in the environmental domain.

3.1. Lexical substitution

Our first method is lexical substitution using MLMs. MLMs are models trained to predict which tokens are likely to replace a special token `<mask>`. They can thus easily be used to acquire synonymy between MWTs. Let MWT_1 and MWT_2 be two MWTs with the same syntactic structure, such that MWT_1 contains the lexical words W_1 and W_3 and MWT_2 contains W_2 and W_3. We assume that MWT_1 and MWT_2 have a compositional meaning. Therefore, W_3 contributes identically to the meaning of MWT_1 and MWT_2 and as a consequence, the relationship between W_1 and W_2 is preserved between MWT_1 and MWT_2. Let S_1 be a context of MWT_1 and S_2 a context of MWT_2. Let k_1 be the rank of W_1 among the MLM predictions for the query obtained by masking W_2 in S_2. Let k_2 be the rank of W_2 among the MLM predictions for the query obtained by masking W_1 in S_1. Let N be the number of neighbors that we consider to be close enough. If $k_1 < N$ or if $k_2 < N$, we predict (*i*) that W_1 and W_2 are synonymous and (*ii*) that MWT_1 and MWT_2 are probably synonymous.

The method can be illustrated with the following example. The context S_1 is used to create the query Q_1 whose $N = 10$ first answers contain the other word (*protection*). This allows the method to predict that the relation (synonymy) between the two TS also exists between the two MWTs.

MWT pair: forest preservation ; forest protection

$\mathbf{M_1}$: preservation

$\mathbf{M_2}$: protection

Target relation: synonymy

Masked context $\mathbf{S_1}$: financial support for the $< mask >$ of forests will be a major topic at the conference

N = 10

Observation: *protection* appears at rank 2 in the list of predictions for query Q_1

Conclusion: *forest preservation* and *forest protection* are synonyms

In our study, we compare “basic” MLM queries and conditioned MLM queries. Zhou et al. (2019) observe that MLMs produce candidates that can be semantically very different from the masked word while being perfectly suited to the context. To solve this problem, we adopt the conditioning method proposed by Qiang et al. (2019). The method uses queries composed of the concatenation of the original context (where the target word is not masked) and the masked context (where the target word is masked).

3.2. Analogy

The second method is based on analogy in static word embeddings. The detection of relations between MWTs by analogy follows from the observation that if $W_1 : MW_2 :: MWT_1 : MWT_2$ is a proportional analogy then the relation between W_1 and W_2 is the same as the one between MWT_1 and MWT_2. The analogy function *3CosADD* (Mikolov et al., 2013) can be used to solve analogy equations in DSMs. For example, if we choose MWT_2 as the unknown, then we seek to estimate the distance between the representation of MWT_2 and the expected vector $V_{\text{expected}} = V_{MWT_1} - V_{W_1} + V_{W_2}$. Each quadruplet produces two analogy equations taking respectively MWT_1 or MWT_2 as the unknown. The final result is the average of the rank of the unknown MWT in the predictions for both equations. The following example illustrates the method:

Quadruplet: *dry : wet : dry climate : humid climate*

Known relationship: antonymy between *dry* and *humid*

Analogy equations:

equation_1: *dry : wet :: dry climate : ?*;

equation_2: *dry : wet :: ? : humid climate*

Number of neighbors considered close: 5

Observation: the expected MWT for both equations are found in the first 5 predictions

Conclusion: *dry climate* and *humid climate* are antonyms

4. Data

Corpus. We used the English and French monolingual PANACEA Environment corpora (ELRA-W00653 and ELRA-W0065) which were built in the framework of the PANACEA project[1]. These corpora are more heterogeneous than typical specialized ones which normally contain specialized texts only because the environmental domain is heterogeneous in nature Bernier-Colborne (2016).

[1] `http://www.panacea-lr.eu/en/info-for-researchers/data-sets/monolingual-corpora`

DicoEnviro. DiCoEnviro[2] is a multilingual dictionary of environmental terms developed by the Observatoire de linguistique Sens-Texte (OLST)[3]. It describes the meaning and linguistic properties (especially the lexical-semantic ones) of terms belonging to various sub-domains of the environment domain.

IATE. IATE[4] (Interactive Terminology for Europe) is an EU translation terminology resource that contains synonymy relations between terms. It is a rich resource from which datasets can easily be extracted.

Data sets. Our datasets are created from IATE. We extracted 786 pairs of synonymous English biterms (we will call this set Data_en in the following) like *climate conference* : *climate summit* and 928 pairs in French (we will call this second set Data_fr in what follows) like *analyse du risque*: *risk study*. We manually validated the synonymy relation between the biterms in the extracted pairs. In order to further select our data, we performed an analysis of the pairs extracted from IATE. We observed that more than 85% of the pairs of synonymous biterms share one lexical word. We used the following subsets of Data_en and Data_fr in our experiments: Data_MLM_en (510 pairs) and Data_MLM_fr (563 pairs) are made up of MWT pairs that have the same pattern and share one lexical element while Data_FastText_en (431 pairs) and Data_FastText_fr (599 pairs) contain MWT pairs of frequency higher than 5 in PANACEA.

5. Experiments

We use the MRR score and the precision at Top1, Top5, and Top10 to evaluate the quality of methods for all our experiments.

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

where $|W|$ is the number of queries and $Rank_i$ is the rank of the first correct answer for the i-th query. The closer the MRR score is to 1, the better the model performs.

$$\text{Precision} = \frac{n}{|W|}$$

where n is the number of queries that produce a correct result among the answers at Top1, Top5, or Top10 and where $|W|$ is the total number of queries.

5.1. Lexical substitution

Data sets. The lexical substitution experiments were performed using Data_MLM_en and Data_MLM_fr. We removed the pairs of biterms in which one of the lexical items is not included in the vocabulary of the model because out-of-vocabulary words are divided into several wordpieces, which makes the identification of the possible substitutes for the target word difficult. Moreover, we need contexts to build the queries submitted to the MLM model. For each MWT in the dataset, we extracted 100 contexts from the corresponding PANACEA corpus. Only contexts that meet some of the quality criteria of good contexts proposed by Kilgarriff et al. (2008) were selected. Note that some MWTs appear less than 100 times in PANACEA. Overall, the data used for the lexical substitution experiment in English consists of 317 term pairs and 24,265 contexts (referred to as Test_MLM_en). For French, the data used consists of 385 term pairs and 24 404 contexts (referred to as Test_MLM_fr).

Models. We conducted the experiments of synonymy acquisition by lexical substitution using the BERT-large-uncased model for English and CamemBERT-large for French.

Vocabulary. In these experiments, we only count as candidates the substitutes that are simple terms. As the MWTs were extracted from IATE, we considered using as a reference the simple terms that appear in this same bank. However, the number of single terms in IATE proved to be too small. For that reason, we used a larger vocabulary consisting of the 818 English and 784 French lexical units that appear in the MWTs of Data_en and Data_fr and that are part of the vocabulary of the models.

5.2. Analogy

Data sets. The data used for the analogy experiments are quadruplets $W_1 : W_2 :: MWT_1 : MWT_2$ such that MWT_1 contains W_1; MWT_2 contains W_2; MWT_1 and MWT_2 share a word W_3; W_1 and W_2 are synonyms. For each of the two languages, two datasets were created using the biterms pairs in Data_Fasttext_en and Data_Fasttext_fr and the synonymy relations in DiCoEnviro and IATE. These datasets will be referred to as Quad_IATE_en (9 quadruplets) and Quad_Envi_en (33 quadruplets) for English and Quad_IATE_fr (20 quadruplets) and Quad_Envi_fr for French (63 quadruplets).

Models. The representations of the MWT should not be computed by composition from the representations of their constituents because the analogy equation would then always be trivially true. Therefore, we use FastText models for the acquisition of synonyms by analogy because can include independent representations for MWTs and their constituents within the same vector space. To compute these representations, we first annotated the corpus so that MWTs and their constituents are indexed separately. For example, a MWT such as *cold air* produces the three tokens: `air`, `cold`, and `air_cold`. We have also forced the model not to split the words into character n-grams by setting the `maxn` parameter to 0.

Vocabulary. The task being the acquisition of synonymy between MWT, the rank of the candidate solutions of the analogical equation is computed with respect to a vocabulary composed of all the nominal biterms in IATE which appear at least 5 times in the PANACEA corpus (5 465

[2] http://olst.ling.umontreal.ca/cgi-bin/DiCoEnviro/search_enviro.cgi

[3] http://olst.ling.umontreal.ca

[4] https://iate.europa.eu/

biterms for English and 5 002 biterms for French).

6. Results and discussions

The results of the lexical substitution and analogy experiments are presented in Tables 1 and 2. Overall, the two methods perform similarly on the English and French datasets.

Method	MRR	P1	P5	P10
Data_MLM_en without conditionning	0.304	0.186	0.433	0.551
Data_MLM_en with conditionning	**0.443**	**0.315**	**0.589**	**0.689**
Data_MLM_fr without conditionning	0.302	0.189	0.416	0.532
Data_MLM_fr with conditionning	0.374	0.253	0.502	0.613

Table 1: MRR score and precision at Top1, Top5 and Top10 of the lexical substitution methods using MLM queries without and with conditioning

Method	MRR	P1	P5	P10
Quad_IATE_en	0.733	0.612	**0.889**	0.889
Quad_Envi_en	0.698	**0.727**	0.83	**0.909**
Quad_IATE_fr	**0.744**	0.650	0.875	0.900
Quad_Envi_fr	0.624	0.548	0.723	0.746

Table 2: MRR score and precision at Top1, Top5 and Top10 of the analogy method using FastText models

Table 1 shows that lexical substitution results are improved by query conditioning. The MRR scores increase from 0.304 to 0.443 for English biterms and from 0.302 to 0.374 for French biterms. Precision is also improved. Query conditioning improves synonymy acquisition in English more than in French. This could be due to the fact that we used different MLMs for the two languages. A second possible reason could be that MWT contexts in English are less informative than those in French, which could be roughly estimated by the length of the contexts: on average, English queries contain 30 words while French ones contain 35 words. Moreover, we also checked in both languages that short queries benefit more from the conditional strategy than long ones. A qualitative analysis of the first 10 predictions of 100 randomly selected queries with conditioning shows that most of the predictions are semantically similar to the masked word. Most of them are synonyms and variants, including derivational ones. These results are in line with the observations of Ferret (2021) and Arefyev et al. (2020). For most queries where the expected term does not appear in the Top10 predictions, some of its synonyms do. For example, when *habitation* 'house' is masked in a context of *habitation individuelle* 'individual house', the expected term *maison* 'house' only appears at rank 71, but its synonym *logement* 'housing' appears at rank 2.

Table 2 shows that analogy captures synonymy between MWTs effectively. The best MRR score of 0.744 is obtained for Quad_IATE. We can also see that the quadruplets constructed using relations between simple terms from IATE give the best results. These good numbers could be explained by the fact that the synonymy relations between MWTs and the simple terms that compose these quadruplets come from the same source. Remember that IATE is a translation dictionary while DiCoEnvio is a terminology database. We conducted a qualitative analysis similar to the one we did for lexical substitution. We examined the first 5 predictions of all queries whose unknown is MWT_2. We observed that MWT_2 is present in the first five candidates for more than 70% of the quadruplets. When MWT_2 does not appear among the first five candidates, we found in most cases one of its synonyms among the candidates. For example, for the quadruplet *effet:incidence::effet sur l'environnement:incidence sur l'environnement* 'effect:impact::environmental effect:environmental impact', the unknown term *incidence sur l'environnement* 'impact on the environment' is at rank 3 698, but the first prediction, *incidence environnementale* 'environmental impact', is a derivational variant of the target MWT. In addition, we also noticed that the frequency of MWTs in the corpus also has an impact on the results. For queries where the unknown term and its synonym do not appear in the first five predictions, the frequency of MWT_1 or/and MWT_2 is often less than 10.

The main difference between the lexical substitution and analogy methods is that MLM predictions are highly context dependent unlike the predictions based on FastText models. Moreover, FastText representations are built on a small specialized corpus, while BERT models are pre-trained on a large variety of corpora. The differences also arise from the fact that BERT is queried at the occurrence level, whereas FastText representations are based on a set of occurrences. Our analogy-based method is better suited for synonymy identification. This observation is consistent with that of Peters et al. (2018) who show that contextual language models underperform compared to static models on analogy-based semantic relation identification tasks.

It should also be noted that the analogy method gets more information than the lexical substitution method because it is provided with the two related simple terms, which is not the case for the latter. The better results obtained with analogy also suggest that semantic composition in MWTs is better captured by FastText models than it is by MLMs. These observations are consistent with the conclusion of Hupkes et al. (2020) that Transformer models have a low level of compositional generalization.

7. Conclusion

In this paper, we explored the capacity of DSMs to capture synonymy between biterms of the environment domain in English and French. We performed experiments using MLMs for the lexical substitution method and using static FastText models for the analogy method. The results of the experiments show that overall, both methods perform well in both languages. However, the analogy methods outperform the lexical substitution methods. The analogy method obtains an MRR score of 0.744 on the French dataset extracted from IATE. Our results also suggest that semantic composition is better grasped by static dives; conversely, the level of compositional generalization of Transformer models seems to be lower. Overall, this study is one of the first attempts to identify synonymy between MWTs in a specialized domain, the environment, by exploring DSMs. This work also provides a roadmap for the application of DSMs to the terminology structuring task. The future step of this work is to increase the performance of the lexical substitution method by improving the quality of the contexts. We also plan to use generative models like GPT-3 (Brown et al., 2020) instead of MLMs to perform the lexical substitution task.

8. Acknowledgements

This work has been funded by the French National Research Agency (ANR) under the ADDICTE (ANR-17-CE23-0001) and the DELICES (ANR-19-CE38-0005) projects

References

Arefyev, Nikolay, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko, 2020. A comparative study of lexical substitution approaches based on neural language models. *arXiv preprint arXiv:2006.00031*.

Bernier-Colborne, Gabriel, 2016. *Aide à l'identification de relations lexicales au moyen de la sémantique distributionnelle et son application à un corpus bilingue du domaine de l'environnement*. Ph.D. thesis, Université de Montréal, Montréal, Canada.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, 2020. Language Models are Few-Shot Learners. In H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc.

Chaudhri, Vinay K, Justin Xu, Han Lin Aung, and Sajana Weerawardhena, 2022. *A Corpus of Biology Analogy Questions as a Challenge for Explainable AI*. Cham: Springer International Publishing, pages 327–337.

Chen, Zhiwei, Zhe He, Xiuwen Liu, and Jiang Bian, 2018. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases. *BMC medical informatics and decision making*, 18(2):53–68.

Ferret, Olivier, 2021. Exploration des relations sémantiques sous-jacentes aux plongements contextuels de mots (exploring semantic relations underlying contextual word embeddings). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*. Lille, France: ATALA.

Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka, 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*. San Diego, California.

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni, 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý, 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*. Barcelona, Spain: Documenta Universitaria.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig, 2013. 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*. Atlanta, Georgia: Association for Computational Linguistics.

Paullada, Amandalynne, Bethany Percha, and Trevor Cohen, 2020. Improving biomedical analogical retrieval with embedding of structural dependencies. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics.

Peters, Matthew E., Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih, 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.

Qiang, Jipeng, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu, 2019. A simple bert-based approach for lexical simplification. *ArXiv*, abs/1907.06226.

Schick, Timo and Hinrich Schütze, 2020. Rare words: A major problem for contextualized embeddings and how to fix it by attentive mimicking. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 34. New York, USA: AAAI Press.

Wohlgenannt, Gerhard, Ekaterina Chernyak, Dmitry Ilvovsky, Ariadna Barinova, and Dmitry Mouromtsev, 2019. Relation extraction datasets in the digital humanities domain and their evaluation with word embeddings. *arXiv preprint arXiv:1903.01284*.

Zhou, Wangchunshu, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou, 2019. Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.