Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020), pages 50–54
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A study of semantic projection from single word terms to multi-word terms in the environment domain

Yizhe Wang[1], Béatrice Daille[2], Nabil Hathout[1]
CLLE, CNRS & University of Toulouse[1], LN2S, CNRS & University of Nantes [2]
yizhe.wang@univ-tlse2.fr, nabil.hathout@univ-tlse2.fr, Beatrice.Daille@univ-nantes.fr

Abstract

The semantic projection method is often used in terminology structuring to infer semantic relations between terms. Semantic projection relies upon the assumption of semantic compositionality: the relation that links simple term pairs remains valid in pairs of complex terms built from these simple terms. This paper proposes to investigate whether this assumption commonly adopted in natural language processing is actually valid. First, we describe the process of constructing a list of semantically linked multi-word terms (MWTs) related to the environmental field through the extraction of semantic variants. Second, we present our analysis of the results from the semantic projection. We find that contexts play an essential role in defining the relations between MWTs.

Keywords: semantic projection, multiwords terms, terminological relations

1. Introduction

Terminology is structured by semantic relations between terms. The relations may be identified by experts, obtained from existing resources or extracted from corpora. They include synonymy, quasi-synonymy, hypernymy, hyponymy, etc. In this study, we focus on the identification of terminological relations between multi-word terms (MWTs) using the semantic projection method. We built a set of French MWT candidates related to the environment domain and containing two lexical words such as *réchauffement climatique* 'global warming'. Three relation categories (antonymy, quasi-synonym, and hypernymy) between single word terms (SWTs) are extended to these candidates. A subset of these MWT pairs has been validated by three judges to assess the preservation and the validity of the inferred relations between MWTs. The main finding of the evaluation is that the context is crucial for the assessment because they determine the actual meaning of the MWTs.

The remaining of the paper is organized as follows. Section 2. presents the related work. Section 3. outlines the projection of the semantic relations on the MWT pairs. Section 4. describes the data and resources used for the generation of semantically linked MWTs (Section 5.). The manual evaluation and an analysis of the projection results are presented in Section 6.. A short conclusion is then given in Section 7..

2. Related work

Several approaches to semantic relation recognition have been proposed in the literature. They may be classified into three types: lexicon-based approaches (Senellart and Blondel, 2008); pattern-based approaches (Wang et al., 2010; Ravichandran and Hovy, 2002); distributional approaches (Rajana et al., 2017; Shwartz and Dagan, 2019). Since MWTs are compositional or at least weakly compositional (L'homme, 2004), the semantic projection method, also known as semantic variation and often referred to as a compositional method, is widely used to generate MWTs and predict relations between them from semantically related SWTs.

Synonymy is an important relation in terminology and is addressed in several studies, like Hamon et al. (1998) who identify synonymous term candidates through inference rules. The authors extract MWT candidates from a corpus on electric power plants and analyze the candidate terms (*ligne d'alimentation* 'supply line') as being made up of a head (*ligne* 'line') and an expansion (*alimentation* 'supply'). They then replace the head or the expansion (or both) by their synonyms obtained from a general dictionary. They assume that the substitution preserves the synonymy relation. In their study, 396 MWT pairs have been validated by an expert; 37% are real synonyms. The same method is used by Hamon and Nazarenko (2001) in order to detect synonymous MWTs in specialized corpora on nuclear power plants and coronary diseases. Their results show that general language dictionaries complement specialized hand-built lexical resources for the detection of semantic variants.

In a similar study, Morin (1999) uses inference rules to identify hierarchical relations (hypernymy) between MWTs. Instead of using relations from the general dictionary, they take as reference semantically linked SWTs extracted from the AGROVOC terminology. They not only add syntactic and semantic constraints on the reference rules but also use the semantic relations with morphological relations to detect the semantic variants. They then compare the relations generated from AGROVOC with relations generated from a general language dictionary and show that the latter has a significantly lower precision. More recently, Daille and Hazem (2014) have generalized the projection method to all types of lexical relations while Hazem and Daille (2018) use it to extract synonymous MWTs with variable lengths.

The main difference between our study and the ones presented above is that we use the context to validate the inferred relations. In our experiment, we have extracted from the corpus 5 contexts for each candidate in the validation dataset. We consider that the projection is valid if the meaning of two MTWs in at least two of their contexts is in the relation stated between the two SWTs that yielded them. The above studies do not use the context except (Hamon and Nazarenko, 2001) who checks whether the two MWT candidates can be substituted one for the other in one context. In other words, one contribution of our study is to take into account the possible ambiguity of the MWTs, and the way contexts determine their meanings.

3. Composition method

Our method is based on the assumption that MWT meaning is compositional. One consequence of this hypothesis is that when two MWTs t_1 and t_2 only differ by one of their components c_1 and c_2, the semantic relation between c_1 and c_2 is identical to the one between t_1 and t_2 because c_1 and c_2 contribute in the same way to the meanings of t_1 and t_2. For instance, the relation between the MWTs *croissance de la population* ‘population growth’ and *diminution de la population* ‘population decline’ is the same as the one between the SWTs *croissance* ‘growth’ and *diminution* ‘decline’, that is antonymy. Our hypothesis is actually a bit stronger because we consider that the equivalence holds even when t_1 and t_2 do not have the same (syntactic) structure. More formally, let t_1 and t_2 be two MWTs such as $\mathbf{voc}(t_1) = \{u_1, v_1\}$ and $\mathbf{voc}(t_2) = \{u_2, v_2\}$ where $\mathbf{voc}(x)$ is the set of the content words of x. If u_1 and u_2 are SWTs, if $v_1 = v_2$ and if there is a semantic relation R between u_1 and u_2, then R also holds between t_1 and t_2. In other words, if $\mathcal{M}$ is a set of MWTs of a domain and $\mathcal{S}$ is a set of SWTs, the hypothesis can be stated as follows:

$$\begin{aligned} \forall t_1 \in \mathcal{M}, \forall t_2 \in \mathcal{M} \text{ such as } \exists u_1, v_1, u_2, v_2 / \\ \mathbf{voc}(t_1) = \{u_1, v_1\} \wedge \mathbf{voc}(t_2) = \{u_2, v_2\} \\ \wedge u_1 \in \mathcal{S} \wedge u_2 \in \mathcal{S}, \\ [v_1 = v2 \wedge \exists R, R(u_1, u_2) \Rightarrow R(t_1, t_2)] \end{aligned}$$

4. Data and resources

4.1. Corpus

The corpus used for extracting MWT candidates is a specialized monolingual French corpus in the environment domain (ELRA-W0065) created in the framework of the PANACEA project[1]. The corpus contains 35453 documents (about 50 million words) with different levels of specialization. The corpus has been preprocessed: extraction of the text, normalization of the characters, lemmatization with TreeTagger (Schmid, 1994).

4.2. TermSuite

The MWT candidates were extracted from the PANACEA corpus through TermSuit, a terminology extraction tool developed at LS2N[2] (Cram and Daille, 2016). TermSuit only extracts noun phrases; the candidates are provided with their part of speech, specificity, and frequency. Table 1 illustrates the extracted candidates. For this study, we only consider the candidates composed of two lexical words (e.g. *milieu naturel* ‘natural environment’).

#	type	pattern	pilot	freq	spec
3	T	N A	parc national	10198	4.17
3	V[s]	N A A	parc naturel national	59	1.94
4	T	A	communautaire	8864	4.11
13	T	T	biomasse	6239	3.96
17	T	N A	diversité biologique	5412	3.90
21	T	N A	milieu naturel	4328	3.80
21	V[s]	N A A	milieu naturel aquatique	23	1.54

Table 1: Excerpt of the TermSuite output

4.3. Reference list of linked terms

The semantic relations between MWT candidates are predicted from relations between SWTs. These semantically linked SWTs are taken from a dataset made available by Bernier-Colborne and Drouin (2016). This reference list (RefCD) is extracted from DiCoEnviro (L’Homme and Lanneville, 2014), a specialized dictionary of the environment field which describes the meaning of 1382 entry terms of various sub-fields: energy, climate change, transportation, etc. RefCD is composed of 1314 term pairs, mainly SWTs, connected by four relation categories:

1. **Quasi-synonyms (QSYN):** synonyms (*diesel* ‘diesel’ ↔ *gazole* ‘diesel’); quasi-synonyms (*conserver* ‘preserve’ ↔ *protéger* ‘protect’); close meanings (*électricité* ‘electricity’ ↔ *énergie* ‘energy’); variants (*autopartage* ‘car sharing’ ↔ *auto-partage* ‘car sharing’).

2. **Hierarchical relations (HYP):** hyponyms (*autoroute* ‘highway’ → *route* ‘road’); hypernyms (*combustible* ‘fuel’ → *pétrole* ‘oil’). Because HYP mixes hyponyms and hypernyms, the pairs it connects are not in order.

3. **Opposites (ANTI):** antonyms (*accélérer* ‘accelerate’ ↔ *ralentir* ‘slow down’); contrastives (*flore* ‘flora’ ↔ *faune* ‘fona’).

4. **Derivatives (DRV):** terms with the same meaning but different parts of speech (*sensibilité* ‘sensitivity’ ↔ *sensible* ‘sensible’).

Because we are focusing on the projection of lexical-semantic relations, we did not use the 259 DRV pairs and excluded them from RefCD. We also excluded the 225 pairs of verbs because TermSuite only extracts noun phrases. Since RefCD does not contain information between simple terms describing other relations, like co-hyponyms, our study on semantic relations between MWTs concentrates on QSYN, HYP, and ANTI. The distribution of the three relation categories is imbalanced, as shown in table 2.

	ANTI	HYP	QSYN	total
Pairs	116	191	523	830
Terms	107	122	415	429

Table 2: Number of terms and semantic relations in RefCD

5. Generation of semantically-linked MWTs

5.1. Raw projection

We extracted all the MWT candidates which contain two content words and formed all the MWTs pairs that share a

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

[2] https://www.ls2n.fr

common word and where the two other words are a pair of SWTs connected in RefCD. We did not impose any other restriction on PoS, the order of the constituents, nor the patterns of the MWT candidates. 18,382 pairs of MWT candidates have been created. Table 3 presents their distribution over the three relation categories.

ANTI	HYP	QSYN	total
3414	3696	11,272	18,382

Table 3: MWTs yielded by the semantic projection

5.2. Data filtering

The raw projection yields symmetrical pairs of MWT candidates because some of the SWT pairs in RefCD are in random order. For instance, the projection produced the couple *climat régional* : *climat local* 'regional climate : local climate' and the couple *climat local* : *climat régional*. Therefore, we deleted the symmetries of hierarchical relationships. Table 4 shows the number of pairs that remained after the data filtering.

ANTI	HYP	QSYN	total
2065	2403	6777	11,245

Table 4: Number of unordered pairs of MWT candidates

5.3. Selection of a validation subset

In order to assess the hypothesis that MWT meaning is compositional and that semantic relations between SWTs are preserved when they are projected on MWTs, we performed a manual validation on a subset of the MWT candidate pairs we have extracted. Since our study focuses on the preservation and the validity of the semantic relations, we do not want to include the quality of candidates in the validation (are they terms of the environmental field?). For instance, a candidate like *lutte contre le changement* 'fight against the change' is not a term because it is syntactically incomplete, and the actual term is *lutte contre le changement climatique* 'fight against climate change'. Additionally, a candidate like *cadre régional* 'regional framework' does not belong to the environment domain.

Therefore, we choose to check the term status of the MWT candidates through three online terminological dictionaries, namely TERMIUM Plus[3], Le Grand Dictionnaire[4] and IATE[5] (Interactive Terminology for Europe). We consider any candidate present in any of these resources is a term of the environmental field since it was extracted from a specialized corpus of this domain. Since many of the extracted terms are specific, such as *conservation du papillon* 'butterfly conservation', only a fraction of the pairs have both of their MWT candidates present in one of the resources. As shown in Table 5, the validation subset is rather small.

In general, all selected candidates are noun phrases because all MWT candidates extracted by TermSuite are noun phrases. In addition, most of the valid pairs are composed of two candidates having the same patterns, NA or NPN.

ANTI	HYP	QSYN	total
80	51	100	231

Table 5: Validation subset

NA-NA	NA-NPN	NN-NN	NN-NPN	NPN-NPN
123	1	1	2	104

Table 6: Distribution of pattern pairs of the validation subset

6. Evaluation of semantic projection

6.1. Contexts

The meaning of a word strongly depends on the contexts where it is used. In this study, we show that the context also determines the meaning of MWTS and the relations that connect them. The annotation of the MWT pairs is based on the relation between the two SWTs they contain and five contexts (i.e., sentences) extracted from the corpus for each MWT. The validity of the projected relation is decided based on the meanings of the MWT occurrences in the extracted contexts. The relation is valid if it holds between the meanings of at least one occurrence of each of the MWTs.

The context may help the judges understand the meaning of a MWT like *zone de recharge* 'recharge zone' which refers to a free aquifer where water collects. It can be used to disambiguate a term like *air frais* 'fresh air' which does not mean cool air but air from the outside (1). Contexts may also highlight the polysemy of MWTs like *changement du climat* 'climate change' which has two meanings: 'global warming' in (2a) and 'climate variability' in (2b).

(1) *la ventilation est à double flux (l'air vicié intérieur réchauffe l'**air frais** entrant)*

'the ventilation is double flow (the inside stale air heats the incoming **fresh air**)'

(2) a. *il a établi que le **changement du climat** était « sans équivoque » et que les émissions de gaz à effet de serre provenant des activités humaines étaient responsables (avec 90% de certitude) de l'augmentation des températures depuis cent ans*

'it established that the **climate change** was "unequivocal" and that greenhouse gas emissions from human activities were responsible (with 90% certainty) for the increase in temperatures over the past hundred years'

b. *à quelle vitesse la réduction des concentrations atmosphériques de GES de courte durée entraînerait un **changement du climat***

'how quickly reducing short-lived atmospheric GHG concentrations would cause **climate change**'

[3] https://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-fra.html?lang=fra

[4] http://www.granddictionnaire.com/

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

6.2. Criteria

The selected pairs have been annotated according to two criteria: the preservation of semantic relations and their validity in the environment domain. Both criteria are based on the expert knowledge of judges on semantic relations and the contexts in which the MWT candidates appear.

1. We consider that a relation is preserved when the relation that holds between two SWTs also holds between two MWT candidates generated from these two SWTs, regardless of its validity as an instance of its category. In other words, the relation is preserved when SWT_1:SWT_2::MWT_1:MWT_2 form an analogy.
2. We consider that a relation between two MWT candidates is valid in the domain when it actually belongs to the category to which it is assigned.

We assessed the preservation of the relation and its validity separately because we have slightly changed the scope of the relation categories. We consider that co-hyponyms are not quasi-synonyms and cannot belong to QSYN. Furthermore, we consider the relationship between a pair of contrastive co-hyponym terms as an instance of ANTI.

6.3. Preservation

The preservation of the relation only depends on the relations between the two SWTs and the two MWTs. If the relations are identical, the relation is considered as being preserved as in the case of *temps froid* : *temps chaud* 'cold weather : warm weather' (3) with respect to *froid* : *chaud* 'cold' : 'warm'.

(3) a. *par* ***temps froid****, cette technique consiste à ne pas laisser tourner son moteur au ralenti plus de 30 secondes*

'by **cold weather**, this technique consists in not leaving the engine idling for more than 30 seconds'

b. *par* ***temps chaud****, le compromis entre confort et pratique est difficile à trouver*

'by **warm weather**, the compromise between comfort and practicality is difficult to find'

On the other hand, *diversité* is a hypernym *biodiversité* in RefCD, but the contexts in (4) show that the relation between the MWTs *gestion de la diversité* 'management of diversity' and *gestion de la biodiversité*
'management of biodiversity' is different since they are used with the same meaning.

(4) a. *les variétés paysannes, issues de millénaires de* ***gestion de la diversité*** *par les agriculteurs sont trop vivantes pour se plier aux critères d'inscription*

'peasant varieties, coming from millennia of **diversity management** by farmers are too alive to comply with the criteria for registration'

b. *elle même distincte de l'utilisation (par les agriculteurs) des semences, la* ***gestion de la biodiversité*** *cultivée réunit dans un processus continu*

'itself distinct from the use (by farmers) of seeds, the cultivated **management of biodiversity** unites in a continuous process'

6.4. Domain validity

Relations that are not preserved are considered as invalid. However, not all preserved relations are valid in the domain. For instance, *agriculture* 'agriculture' is a hypernym of *élevage* 'lifestock farming' in RefCD, and the relation holding between these SWTs is preserved in the MWTs *agriculture biologique* 'organic agriculture' and *élevage biologique* 'organic lifestock farming'. However, a context like (5) shows that these MWTs are actually co-hyponyms because hypernyms cannot be coordinated in this way. The reason is that agriculture is polysemous and may also mean cultivation. In this context, *agriculture* and *élevage* are co-hyponyms, and the inferred relation is not valid because it is not a relation of hypernymy.

(5) *... expérience avec une matrice agricole "sans pesticides ni intrants chimiques" (agriculture ou élevage biologique ou de prairies ...*

'... experience with an agricultural matrix "without pesticides or chemical inputs" (agriculture or organic farming or meadows ...'

6.5. Analysis of the inferred relations

Three judges have annotated the pairs of the validation subset. Table 7 shows that the inter-annotator agreement measured by Fleiss' kappa is substantial. The cases where the judges disagreed were then resolved.

ANTI	HYP	QSYN
0.77	0.68	0.61

Table 7: Fleiss' kappa

The results (Table 8) show that most of MWTs have compositional meaning, which confirms the claim of (L'homme, 2004). They also show that the preservation and the validity of the projected relations vary with their category.

	Preservation			Validity		
	ANTI	HYP	QSYN	ANTI	HYP	QSYN
Yes	68	27	85	68	27	74
No	12	24	15	12	24	26

Table 8: Results of the validation

Even if no restriction on the patterns was used for the generation of the MWT pairs, we observed that in all of the valid pairs, the MWTs have the same patterns and the SWTs that they contain appear in the same positions.

51 out of 231 pairs of MWTs are not preserved. They fall into three groups. (*i*) The MWTs do not have the same structure like *eau de surface* 'surface water' and *surface de la terre* 'Earth's surface'. *eau* 'water' and *terre* 'land' are linked by ANTI relation but the MWTs are not because

eau and *terre* do not appear in the same position. (*ii*) The meaning of the SWTs is not preserved in the MWTs as in *route maritime* : *autoroute maritime* 'shipping route : marine highway'. *route* 'road' is a hypernym of *autoroute* 'highway', but *route maritime* and *autoroute maritime* are synonyms in the contexts extracted for these two MWTs. (*iii*) The change in meaning may also come from the content word shared by two MWTs as in *air libre* : *eau libre* 'outdoor : open water'. The 62 pairs where the relation has been considered invalid are mainly co-hyponyms formed by SWTs linked by a QSYN relation like *trafic ferroviaire* : *trafic routier* 'rail traffic : road traffic'.

7. Conclusion and Future Works

In this study, we have created a dataset of MWT pairs linked by semantic relations. These relations are projected from a reference list of SWTs connected by the same relations. The annotation of a subset of the data highlighted the importance of the contexts because they determine the real meaning of MWTs and subsequently, the semantic relation that holds between them. The following step in this research is to design a method to automate the annotation on the basis of the semantic relations between SWTs and contextual semantic model like BERT (Devlin et al., 2019).

8. References

Bernier-Colborne, G. and 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)*, pages 125–138.

Cram, D. and Daille, B. (2016). Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18.

Daille, B. and Hazem, A. (2014). Semi-compositional method for synonym extraction of multi-word terms. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1202–1207, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota.

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

Hamon, T., Nazarenko, A., and Gros, C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 498–504. Association for Computational Linguistics.

Hazem, A. and 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).*

L'homme, M.-C. (2004). *La terminologie: principes et techniques*. Pum.

L'Homme, M.-C. and Lanneville, M. (2014). Dicoenviro. dictionnaire fondamental de l'environnement. *Consulté à l'adresse http://olst. ling. umontreal. ca/cgibin/dicoenviro/search. cgi.*

Morin, E. (1999). Projecting corpus-based semantic links on a thesaurus. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 389–396.

Rajana, S., Callison-Burch, C., Apidianaki, M., and Shwartz, V. (2017). Learning antonyms with paraphrases and a morphology-aware neural network. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 12–21.

Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 41–47. Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Senellart, P. and Blondel, V. D. (2008). Automatic discovery of similarwords. In *Survey of Text Mining II*, pages 25–44. Springer.

Shwartz, V. and Dagan, I. (2019). Still a pain in the neck: Evaluating text representations on lexical composition. *arXiv preprint arXiv:1902.10618.*

Wang, W., Thomas, C., Sheth, A., and Chan, V. (2010). Pattern-based synonym and antonym extraction. In *Proceedings of the 48th annual southeast regional conference*, page 64. ACM.