Word Embedding Approach for Synonym Extraction of Multi-Word Terms

Amir Hazem and Béatrice Daille
Laboratoire des Sciences du Numérique de Nantes (LS2N)
Université de Nantes, 44322 Nantes Cedex 3, France
Amir.Hazem@univ-nantes.fr,Beatrice.Daille@univ-nantes.fr

Abstract

The acquisition of synonyms and quasi-synonyms of multi-word terms (MWTs) is a relatively new and under represented topic of research. However, dealing with MWT synonyms and semantically related terms is a challenging task, especially when MWT synonyms are single word terms (SWTs) or MWTs of different lengths. While several researches addressed synonym extraction of SWTs, few of them dealt with MWTs and fewer or none while MWTs synonyms are of variable lengths. The present research aims at introducing a new word-embedding-based approach for the automatic acquisition of synonyms of MWTs that manage length variability. We evaluate our approach on two specialized domain corpora, a French/English corpus of the wind energy domain and a French/English corpus of the breast cancer domain and show superior results compared to baseline approaches.

Keywords: Synonym extraction, Multi-word terms, Compositionality, Word embeddings

1. Introduction

Synonyms acquisition has mainly concerned single word terms (SWTs) using a variety of approaches such as: lexicon-based approaches (Blondel and Senellart, 2002), multilingual approaches (Wu and Zhou, 2003; van der Plas and Tiedemann, 2006; Andrade et al., 2013), distributional approaches (Lin, 1998; Hagiwara, 2008), etc. However, exploring multi-word terms (MWTs) and their synonyms or semantically related terms can be useful in many applications such as: word sense disambiguation, machine translation, information retrieval, text simplification, etc. MWTs are motivated combinations that clearly convey the concept they designate. The requirement of term transparency argues in favor of compositional semantics for complex terms. Compositionality means that the whole meaning can be deduced from the meaning of its components and the syntactic rule by which they are combined (Partee et al., 1990). Pirrelli et al. (2010) claim that the most productive compounds are compositional (at least weak compositional) constructions. Synonymic variants of multi-words exhibit multiple phenomena ranging from compositional multi-word terms synonyms of the same length such as: *wind turbine*/*wind machine*[1]; MWT synonyms of variable length such as: *wind farm*/*wind power plant*; to non compositional MWT synonyms such as: *pole tower*/*mast*.

Few works addressed the acquisition of MWT synonyms. The main approaches that have been proposed in the experimental literature deal with the acquisition of synonyms of MWTs that are compositional and often of the same length. Synonym extraction approaches implement the principle of compositionality by substituting parts of the MWT by synonyms provided by a dictionary (Hamon and Nazarenko, 2001), or by distributional analysis (Hazem and Daille, 2014).

It has been recently shown that words, phrases, sentences, paragraphs and more generally, pieces of texts of any length can be efficiently represented by word embeddings using operations on vectors and matrices like addition or multiplication (Mitchell and Lapata, 2010; Mikolov et al., 2013b; Socher et al., 2011; Mikolov et al., 2013b; Le and Mikolov, 2014; Kalchbrenner et al., 2014; Kiros et al., 2015; Wieting et al., 2016; Arora et al., 2017; Hazem et al., 2017). For phrase representation, Mikolov et al. (2013b) have shown for instance that the embedding vector of the phrase *Volga river* is similar to the addition of the embedding vector of *Volga* and the embedding vector of *river*. The addition property that word embbeding models exhibit offers key information for representing phrases and by extension MWTs and there synonyms or quasi-synonyms. Drawing inspiration from these findings and based on the principle of compositionality and distributed approaches, we propose several techniques based on word embedding models to deal with synonyms acquisition of MWTs. More specifically, we extend the work of Hazem and Daille (2014) and explore synonym extraction of single word terms and multi-word terms of variable lengths. Our first proposition is an extension of the Semi-compositional approach using word embeddings to extract synonyms of parts of MWT. Our second proposition is a Full-compositional approach based on the additive property of word embeddings to extract synonyms of the entire MWT. We conduct several experiments on two specialized datasets that is: a French/English wind energy corpus and a French/English breast cancer corpus. The obtained results of the proposed approaches outperform the state of art baseline approaches.

The remainder of this paper is organized as follows. Section 2. describes the state of art approaches as well as our proposed techniques. Section 3. describes the different linguistic resources used in our experiments. The experimental setup and the obtained results on the wind energy and the breast cancer corpora are respectively presented in Sections 4. and 5. Section 6. initiates a discussion regarding the obtained results and finally, we conclude our work in Section 7.

2. Approaches

In this section we first describe the two main baseline approaches that deal with MWTs synonyms acquisition that

[1] In the renewable energy domain.

is, the compositional approach and the semi-compositional approach. Then, we develop our proposed techniques that is: semi-compositional word embeddings and full-compositional word embeddings approaches. Except the last approach, all these methods hypothesise that MWT semantics is compositional and thus that a synonym of a MWT could be obtained by substituting one of the component parts by a synonymic expression at a given syntactic position. They differ according to how they provide synonym components.

2.1. Compositional Approach

The compositional approach substitutes one of the component of the MWT by one of its synonyms provided by a synonym dictionary. The synonym MWT is considered as valid if and only if it can be found in the corpus. For instance, given the MWT *collecteur général* 'general collector' extracted from the wind energy corpus (cf. Section 3.), several synonyms of *général* are proposed by dictionary of synonyms: *habituel*, *ordinaire*, *commun*, ... The MWT *collecteur commun* 'common collector' which is the correct synonym of *collecteur général* is validated as it occcurs in the wind energy corpus.

Hamon and Nazarenko (2001) defined three rules to extract synonymy relations by assuming a compositional semantics. Given the multi-word candidate terms $CCT_1 = (T_1, E_1)$ and $CCT_2 = (T_2, E_2)$ and $syn(CT_1, CT_2)$ a synonym relation between the candidate terms CT_1 and CT_2, the following inference rules are used:

- R_1: $T_1 = T_2 \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$
- R_2: $E_1 = E_2 \wedge syn(T_1, T_2) \supset syn(CCT_1, CCT_2)$
- R_3: $syn(T_1, T_2) \wedge syn(E_1, E_2) \supset syn(CCT_1, CCT_2)$

In rule R_1, the heads are identical and the expansions are synonymous, while in rule R_2 heads are synonymous and expansions are identical. Finally, R_3 is a generalization of rules R_1 and R_2. If the compositional approach of Hamon and Nazarenko (2001) is based on a dictionary of synonyms, it can be generalized using external resources that provide semantically related words such as Wordnet for instance. Nonetheless, this approach remains resource dependent. To alleviate this drawback, Hazem and Daille (2014) proposed an approach based on distributional analysis that does not need a dictionary of synonyms or external thesauri and extracts synonyms and semantically related words automatically from the corpus. We present their approach in the next section.

2.2. Semi-Compositional Approach

Like the compositional approach, the semi-compositional variant is based on the principle of compositionality of MWTs. The main difference lies on the nature of the substituted elements of the MWT. It is no longer constrained by the sole relation of synonymy like in Hamon and Nazarenko (2001). Hazem and Daille (2014) generalized the substitution on MWT elements to semantically related terms of any type. They extended the compositional rules R_1 and R_2 by replacing $syn(CCT_1, CCT_2)$ which means synonym relation between CCT_1 and CCT_2 by $sem(CCT_1, CCT_2)$, which means semantic relation between CCT_1 and CCT_2. R_1^G corresponds to the generalized rule R_1 (respectively, R_2^G corresponds to the generalized rule R_2) and T_1, T_2, E_1, E_2 can be MWTs. In addition, they remove the rule R_3 relying on the results of Hamon and Nazarenko (2001) where they have shown that R_3 is the less productive and reliable rule. They obtained the two following rules:

- $R_1^G : T_1 = T_2 \wedge sem(E_1, E_2) \supset sem(CCT_1, CCT_2)$
- $R_2^G : E_1 = E_2 \wedge sem(T_1, T_2) \supset sem(CCT_1, CCT_2)$

For example, the synonym of *énergie renouvelable* 'renewable energy' can be obtained by first extracting each part of the MWT; then, finding the semantically related words of *énergie* 'energy' and/or *renouvelable* 'renewable' with distributional methods; finally, filtering all expressions using monolingual specialized corpora. In the next paragraph we introduce the distributional approach that is used to extract semantically related terms.

Distributional Approach Instead of using a dictionary that will provide synonyms of each lexical element of the MWT, another way to do it is by exploiting distributional relationships. The distributional approach is based on the assumption that words with similar meanings are more likely to share similar contexts. Hence, each word is represented by its context which corresponds to all its surrounding words in the corpus. The surrounding words are often delimited by a window of size n (n is often small 3, 5 or 7 words). Hereafter the main steps of the distributional approach:

- The context vector $v_{w_i^s}$ of a given source word w_i^s is first built. The vector $v_{w_i^s}$ contains all the words that co-occur with w_i^s within a window of n words that surround w_i^s. Let us denote by $occ(w_i^s, w_j^s)$ the co-occurrence count of w_i^s and a given word of its context w_j^s.
- The process of building context vectors is repeated for all words of the specialized corpus.
- Words of the context vectors are weighted using association measures such as the point-wise mutual information (noted MI) (Fano, 1961), the log-likelihood (noted LLR) (Dunning, 1993) or the discounted odds-ratio (noted LO) (Laroche and Langlais, 2010). These measures aim at strengthening the correlation between a word and all the words of its context vector.
- To extract the semantically related words of a given source word w_i^s, a similarity measure such as the cosine similarity (Salton and Lesk, 1968) (noted COS) or the weighted Jaccard index (noted JAC) (Grefenstette, 1994) is applied between $v_{w_i^s}$ and all the target vectors of the corpus $v_{w_i^s}^t$.

- The semantically related candidates of the word w_i^s are the target words ranked according to their similarity scores.

2.3. Word Embeddings Approaches

We introduce two new techniques for synonyms extraction of multi-word terms. The first technique called *Semi-compositional word embeddings*, follows the principle of the semi-compositional approach based on distributional analysis (Hazem and Daille, 2014). It mainly differs in the procedure of extracting SWTs synonyms or semantically related terms which are parts of MWTs. The second technique called *Full-compositional word embeddings*, is inspired by the idea that phrases can be represented by an element-wise sum of the word embeddings of semantically related words of its parts (Mikolov et al., 2013b). It also follows the principle of sentence representation performed by an element wise addition of word embeddings of its parts (Wieting et al., 2016; Arora et al., 2017; Hazem et al., 2017). We adapt this idea and apply it to MWTs. We also experiment the word-embedding state of art approach of Mikolov et al. (2013b) to extract MWTs. We refer to this baseline as *Distributed representation of phrases* and denote it by *Mikolov* approach.

2.3.1. Semi-Compositional Word Embeddings

The Semi-compositional word embeddings approach is also based on the composition of the elements of MWTs. It can be considered as a variant of the semi-compositional approach introduced in (Hazem and Daille, 2014). The difference resides in the manner of extracting semantically related terms of the SWTs ($sem(E_1, E_2)$ and $sem(CCT_1, CCT_2)$). If to do so, Hazem and Daille (2014) use distributional approach as introduced in Subsection 2.2., here we use distributed models (Mikolov et al., 2013b). We explore the two well-known word embedding representation: the Skip-gram model and the continuous bag-of-word model (CBOW).

2.3.2. Full-Compositional Word Embeddings

The Full-compositional word embeddings approach aims at extracting MWTs synonyms of any length. It provides a joint representation for all the MWTs which facilitates MWTs comparison. If the compositional property is applied after-hand in the previous approaches which is problematic when MWTs are of lengths higher than two, the Full-compositional word embeddings approach integrates it beforehand thanks to the additive property of embedding models. All the MWTs are represented by a single embedding vector. Each MWT is first characterized by an element wise sum of its word embedding elements. Then, the cosine similarity measure is applied to extract MWTs synonyms. The implementation of the Semi-compositional and the Full-compositional approaches can be found here `https://github.com/hazemAmir/FullComp.git`

2.3.3. Distributed Representation of Phrases

We apply Mikolov et al. (2013b) approach originally introduced for phrases to MWTs synonyms extraction. Basically, the approach is two-fold. First, (i) we detect and extract all the MWTs of the corpus, then (ii) we consider them as single tokens and build embedding vectors based on their contexts as it is usually done for words by the skip-gram and the CBOW models. Hence, each MWT is characterized by a single embedding vector. Finally, we use the cosine similarity to extract MWTs synonyms. In this approach, the compositionality property is not taken into account. Also, due to the relatively smaller number of MWTs comparing to SWTs, especially in specialized domains, it might be difficult to build efficient embedding models of MWTs. Nonetheless and for a matter of comparison, it is interesting to report the results of this approach.

CBOW and Skip-gram are two distributed representations introduced by Mikolov et al. (2013b) that capture linguistic regularities, namely the Continuous Bag-of-Words (CBOW) model and the Skip-gram model. The principle of the CBOW model is to combine the representations of surrounding words to predict the word in the middle, while the training objective of the Skip-gram model is to learn how to predict the surrounding words based on the representations of the middle word. If CBOW and Skip-gram exhibit similar architectures, CBOW is faster and is more suitable for large datasets while Skip-gram gives better word representations when monolingual data is small (Mikolov et al., 2013a).

3. Data and Resources

In this section, we describe the data and the different resources used in our experiments.

3.1. Corpora

The experiments have been carried out on the French/English specialized corpus from the domain of wind energy of 400,000 words[2] and the French/English specialized corpus from the domain of breast cancer of 500,000 words.

Wind energy corpus is part of the TTC project[3] and has been crawled from the web using *Babouk* (Groc, 2011) crawler. As search engine requests, several technical words have been used such as *wind*, *energy* and *renewable* for English and *vent*, *énergie* and *renouvelable* for French.

Breast cancer corpus has been extracted from *Istex* portal[4] using as keywords *breast cancer* for English and *cancer du sein* for French. The gathered documents concern the period ranging from 2001 to 2015.

The wind energy and breast cancer corpora have both been pre-processed using tokenization, part-of-speech tagging, and lemmatization.

[2] `http://www.lina.univ-nantes.fr/?Ressources-linguistiques-du-projet.html`

[3] `www.ttc-project.eu/index.php/releases-publications`

[4] `https://api.istex.fr/documentation/`

3.2. Reference Lists

Reference lists have been built from various terminological resources. Only the resources that list synonymic terms in their terminological records have been examined. In such lists or databases, synonyms are not systematically present, and for records including them, synonymic variants are various. Many of them are terms related by other types of semantic relations, such as near-synonymy or hypernymy.

For the French part of the wind energy corpus, we selected the French MWT pairs from the *Terminalf*[5] linguistic resource. From 84 MWTs of the wind energy domain, we obtained 34 French MWT synonyms as a result of filtering out SWT synonyms and after checking that the MWT synonyms occur in the specialized corpora. For English, we selected the MWT pairs from the glossary of wind energy from the online book (Gipe, 2004) and from the linguistic resource *Termium* [6]. As a result of filtering and of corpus projection, we obtained 20 English MWT pairs.

This method has been reiterated in order to build the lists of synonyms of multi-word terms in the breast cancer domain. *Termium* has been used. Here again, by discarding the same types of variants for wind energy. After filtering with breast cancer corpus in each language, the lists of reference of the breast cancer domain contain 20 French terms and 16 English terms associated with their synonyms.

The small size of the reference lists can be explained by the small size of the specialized corpora which contain few specialized terms and few synonymic variants. But a more plausible explanation is that the majority of these synonymic variants are contextual. It is difficult for a terminologist to predict and to detect all synonymic variants that can be produced. Contextual synonymic variants do not generally appear in a dictionary resource (Kremer et al., 2014). To evaluate the Full-compositional approach we built a reference list that contains only pairs of synonyms of variable lengths. Following the same procedure for the above described lists, we built a reference list of 10 French and 9 English pairs of synonyms on the wind energy corpus. Here again the small size of the reference lists is due to the lack of synonyms of variable lengths however it is interesting to use these list as a preliminary result.

4. Experimental Setup

For all the experiments the mean average precision *MAP* (Manning et al., 2008) is used to evaluate the quality of the different approaches.

$$MAP = \frac{1}{|W|} \sum_{i=1}^{|W|} \frac{1}{Rank_i} \quad (1)$$

where $|W|$ corresponds to the size of the evaluation list, and $Rank_i$ corresponds to the ranking of a correct synonym candidate i.

[5] http://terminalf.scicog.fr

[6] http://www.btb.termiumplus.gc.ca/tpv2alpha/alpha-eng.html?lang=eng

English term synonyms

aerogenerator	wind turbine generator
windmill	wind turbine
mast	pole tower
rotor-swept area	reference area
wind farm	wind power plant
vertical axis wind turbine	darrieus rotor
wind turbine	wind machine
power supply	energy supply
power plant	electricity plant
savonius model	savonius type
energy output	energy production
wind farm	wind power station
sea wind farm	offshore wind farm
wind turbine	aeroturbine

French term synonyms

éolienne	moulin à vent
rotor de Savonius	anémomètre
générateur synchrone	alternateur
éolienne à axe horizontal	moulin à hélice
parc éolien	implantation
éolienne à axe vertical	rotor de Darrieus
aérogénérateur	turbine éolienne
aéromoteur	moteur éolien
énergie renouvelable	énergie durable
centrale électrique	centrale éolienne
unité de stockage	dispositif de stoskage
arbre primaire	arbre lent
force du vent	vitesse du vent
aérogénérateur	générateur éolien

Table 1: Examples of English/French synonyms and quasi-synonyms of MWTs recorded in terminology banks of the wind energy domain.

4.1. Dictionary-based Method

We used as first baseline the method proposed in Hamon and Nazarenko (2001). To extract French synonyms of single-word terms we used the on-line dictionary *DES* [7]. *DES* contains 49,168 entries and 201,511 synonym relations. The initial database has been constructed from seven dictionaries. The extraction of English synonyms has been conducted using the lexical database *WordNet* [8]. *WordNet* contains approximately 117,000 synsets. The main relation among words in *WordNet* is synonymy.

4.2. Distributional Method Settings

Using the *distributional method*, three main parameters need to be set: the size of the window used to build the context vectors (Morin et al., 2007; Gamallo, 2008), the association measure (the log-likelihood (Dunning, 1993), the point-wise mutual information (Fano, 1961), the discounted odds-ratio (Laroche and Langlais, 2010),...) and the similarity measure (the weighted Jaccard index

[7] http://www.crisco.unicaen.fr/des/synonyms

[8] http://wordnetweb.princeton.edu/perl/webwn/

(Grefenstette, 1994), the cosine similarity (Salton and Lesk, 1968),...). To build the context vectors we chose a 7-window size. We used MI, LLR and LO as association measures and COS and JAC as similarity measures. We refer to the distributional-based approaches by: Semi-Comp (MI-COS), Semi-Comp (LO-COS) and Semi-Comp (LLR-JAC). Other combinations of parameters were assessed, but on average the chosen parameters turned out to give the best performance.

4.3. Word Embeddings Settings

The second baseline is the distributed representation-based approach that we denote by $Mikolov$. For word embeddings, we used as settings a window size ranging from 1 to 20 words[9], negative sampling of 5, sampling of 1e-3 and training over 15 iterations. We applied both Skip-gram and CBOW models[10] to create vectors of dimension ranging from 50 to 800 dimensions. We used hierarchical softmax for training the Skip-gram model. In the proposed approaches, SG100 stands for using skip-gram (100 dimensions) and CBOW300 stands for CBOW (300 dimensions).

5. Results

The experimental results conducted on the French/English wind energy and breast cancer corpora are presented in Tables 2 and 3.

Method	French	English
Hamon&Nazarenko	0.25	3.63
Mikolov	4.56	6.78
Semi-Comp (MI-COS)	27.4	32.6
Semi-Comp (LO-COS)	26.8	27.2
Semi-Comp (LLR-JAC)	31.4	36.1
Semi-Comp (SG50)	30.9	50.3
Semi-Comp (SG100)	**34.9**	55.9
Semi-Comp (SG200)	34.8	52.7
Semi-Comp (CBOW50)	23.0	49.0
Semi-Comp (CBOW100)	23.7	49.4
Semi-Comp (CBOW200)	23.8	49.4
Full-Comp (SG100)	27.3	57.8
Full-Comp (SG200)	28.9	**58.4**
Full-Comp (SG300)	28.5	55.3
Full-Comp (CBOW50)	22.6	47.0
Full-Comp (CBOW100)	20.1	45.1
Full-Comp (CBOW200)	21.6	44.5

Table 2: Results (MAP%) on the wind energy corpus.

First, we observe the very low results of Hamon&Nazarenko approach. This can be explained by the lack of synonymy relations for SWTs part of MWTS. Second, we observe the slightly better results but still low of Mikolov approach. Here, the small size of the datasets is certainly one of the main reasons that can explain the results. Indeed, embedding models of MWTs can't be efficient with small data size. Concerning the Semi-Comp approach, we notice higher results for both distributional-based and embeddings-based approaches

Method	French	English
Hamon&Nazarenko	4.92	7.03
Mikolov	8.37	9.12
Semi-Comp (MI-COS)	19.9	12.6
Semi-Comp (LO-COS)	27.1	11.0
Semi-Comp (LLR-JAC)	13.9	13.3
Semi-Comp (SG50)	32.1	15.0
Semi-Comp (SG100)	**32.2**	15.2
Semi-Comp (SG300)	27.9	9.60
Semi-Comp (CBOW50)	29.1	15.1
Semi-Comp (CBOW100)	29.2	15.3
Semi-Comp (CBOW300)	29.4	15.8
Full-Comp (SG100)	25.6	17.4
Full-Comp (SG200)	28.0	**18.9**
Full-Comp (SG300)	30.5	16.0
Full-Comp (CBOW100)	24.9	10.6
Full-Comp (CBOW200)	24.9	11.6
Full-Comp (CBOW300)	25.0	10.5

Table 3: Results (MAP%) on the breast cancer corpus.

with a better performance for our word-embeddings adaptation (Semi-Comp(SG))[11]. Overall, the best results are mainly obtained by the Full-Comp approach for English datasets and by the Semi-Comp(SG) approach for French datasets.

Figures 1 shows the performance of the semi-compositional approach (noted SemiCBOW and SemiSG[12]) and the full-compositional approach (noted FullCBOW and FullSG) using CBOW and Skip-gram models while varying the context window size (from 1 to 20) and the dimension size (from 50 to 800). We observe that the best results are obtained using small window size and small dimension size. Overall, the FullSG approach obtains the best performance followed by the SemiSG. The SemiCBOW and FullCBOW obtain lower results in general. The best combination is w=5 and dim=100 for FullSG, w=1 and dim=50 for SemiSG, w=3 and dim=400 for SemiCBOW and w=3 and dim=50 for FullCBOW.

To evaluate the Full-compositional approach we did an extra experiment only on synonyms of variable lengths using the wind energy corpus for French and English. We obtained a MAP score of 10.2% and a recall of 66.6% for English and 4.46% of MAP score and a recall of 40% for French. The state of art and Semi-comp proposed approaches can't be applied for this experiment because they don't deal with MWTs length variability. If the results of Full-Comp approach are still low, this approach offers an alternative to pairs of MWTs synonyms that have different lengths.

6. Discussion

Synonyms extraction of MWTs can be addressed using different strategies. When using compositionality property, dictionary-based approach is beneficial when the dictionary of SWT's synonyms is available as shown in (Hamon and Nazarenko, 2001). However, in many cases this resource is difficult to obtain, one interesting alternative

[9]Figures 1 shows the best window size for each approach.

[10]To train word embedding models we used the gensim toolkit (Rehurek and Sojka, 2010).

[11]Except for the Fr breast cancer dataset.

[12]SemiSg with SG that stands for Skip-gram.

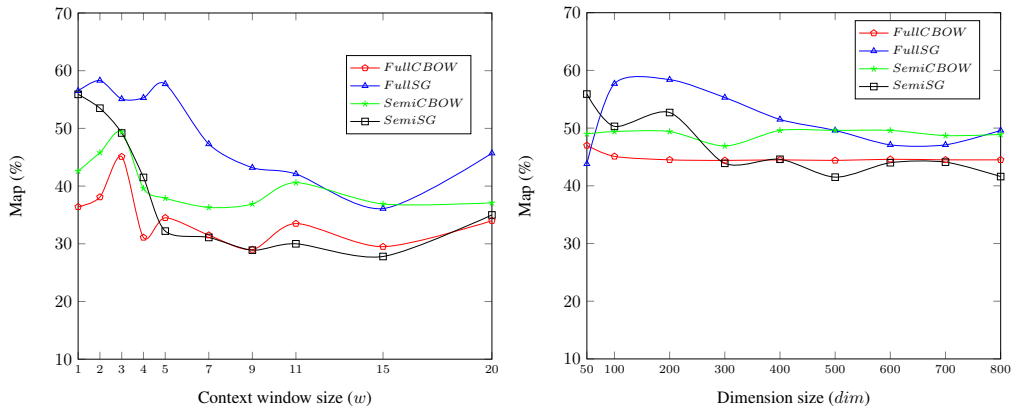

Figure 1: Semi-Comp and Full-Comp comparison while varying the window and dimension size of CBOW and Skip-gram models on the wind energy corpus.

is the distributional-based semi-compositional approach as shown in (Hazem and Daille, 2014). Hence, extracting automatically synonyms of parts of MWT turned out to give better results than looking for them in a dictionary as reported in Tables 2 and 3. With the boom of word embeddings, a straightforward extension of the distributional-based semi-compositional approach is the use of word embeddings to extract synonyms of SWTs. This is the first contribution of this paper. Here again we notice over the results, better performance in most cases using Skip-gram and CBOW models. If the above mentioned approaches are suitable for synonyms extraction of MWTs, they can hardly deal with MWTs synonyms of variable lengths. For instance to extract the synonym of *vertical axis wind turbine* which is *darrieus rotor*, it is not obvious to know that the four-gram length synonym is a bigram in this example. The above cited approaches should know this information or experience all the n-grams possibilities to extract this type of synonyms which is clearly laborious. One alternative which is the second contribution of this paper is the Full-compositional approach. Taking advantage of the additive property of word embeddigs, a MWT can be represented by a single embedding vector which is the result of adding the embedding vectors of its parts. If the Full-compositional approach achieved promising results on the variable length reference lists, the main problem remains its productivity. The question is how to deal with duplicates in the candidates. Filtering is necessary to alleviate repetitive n-grams in different positions. We believe that using sophisticated filtering process[13] based on linguistic patterns for instance, should improve the performance of the Full-Compositional approach. We will pursue this direction in the near future.

7. Conclusion

In this paper, we have proposed different word embeddings approaches for synonyms extraction of MWTs. We have shown that using word embeddings with compositionality and additive composition improve the results comparing to baseline approaches.The full compositional approach which is length independent for MWT representation, has shown the best results in almost all the experiments. If the results on the variable length experiment are still low due to the productivity of this approach, the preliminary results are encouraging since no specific filtering process has been applied. For the future we will pursue this direction by giving more attention to relations between synonyms of variable lengths and their linguistic patterns.

[13]In the variable length FullComp evaluation, we only applied n-grams frequency filtering.

8. Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-17-CE38-0008 HORAE (Hours - Recognition, Analysis, Editions) project and the CominLabs excellence laboratory financed by the National Research Agency under reference ANR-10-LABX-07-01 (project LIMAH `limah.irisa.fr`).

9. Bibliographical References

Andrade, D., Tsuchida, M., Onishi, T., and Ishikawa, K. (2013). Synonym acquisition using bilingual comparable corpora. In *International Joint Conference on Natural Language Processing (IJCNLP'13)*, Nagoya, Japan.

Arora, S., Yingyu, L., and Tengyu, M. (2017). A simple but tough to beat baseline for sentence embeddings. In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

Blondel, V. D. and Senellart, P. (2002). Automatic extraction of synonyms in a dictionary.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74.

Fano, R. M. (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA, USA.

Gamallo, O. (2008). Evaluating two different methods for the task of extracting bilingual lexicons from comparable corpora. In *Proceedings of LREC 2008 Workshop on Comparable Corpora (LREC'08)*, pages 19–26, Marrakech, Marroco.

Gipe, P. (2004). *Wind power: renewable energy for home, farm, and business*. Chelsea Green Pub. Co.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA, USA.

Groc, C. D. (2011). Babouk : Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction. In *Proceedings of 10th International Conferences on Web Intelligence (WIC'11)*, pages 497–498, Lyon, France.

Hagiwara, M. (2008). A supervised learning approach to automatic synonym identification based on distributional features. In *Proceedings of the ACL-08: HLT Student Research Workshop*, pages 1–6, Columbus, Ohio, June. Association for Computational Linguistics.

Hamon, T. and Nazarenko, A. (2001). Detection of synonymy links between terms: experiment and results. In *Recent Advances in Computational Terminology*, pages 185–208. John Benjamins.

Hazem, A. and Daille, B. (2014). Semi-compositional method for synonym extraction of multi-word terms. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Hazem, A., el amel Boussaha, B., and Hernandez, N. (2017). Mappsent: a textual mapping approach for question-to-question similarity. Recent Advances in Natural Language Processing, RANLP 2017, 2-8 September, 2017, Varna, Bulgaria.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *CoRR*, abs/1404.2188.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 540–549.

Laroche, A. and Langlais, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *CoRR*, abs/1405.4053.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manning, D. C., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013a). Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Partee, B., Meulen, A., and Wall, R. (1990). *Mathematical Methods in Linguistics*. Studies in Linguistics and Philosophy. Springer Netherlands.

Pirrelli, V., Guevara, E., and Baroni, M. (2010). Computational issues in compound processing. In Sergio Scalise et al., editors, *Cross-disciplinary issues in compounding*, volume 311 of *Current issues in linguistic theory*, pages 271–285. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Salton, G. and Lesk, M. E. (1968). Computer evaluation of indexing and text processing. *Journal of the Association for Computational Machinery*, 15(1):8–36.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *Advances in Neural Information Processing Systems 24*.

van der Plas, L. and Tiedemann, J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics ACL'06*, Sydney, Australia.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016). Towards universal paraphrastic sentence embeddings. *International Conference on Learning Representations, CoRR*, abs/1511.08198.

Wu, H. and Zhou, M. (2003). Optimizing synonym extraction using monolingual and bilingual resources. In *In Proceedings of the second international workshop on Paraphrasing*, page 72.