Unsupervised Compositional Translation of Multiword Expressions

Pablo Gamallo
Centro de Investigación en
Tecnoloxías Intelixentes (CiTIUS)
University of Santiago de Compostela
Galiza
pablo.gamallo@usc.es

Marcos Garcia
LyS Group
University of A Corunha
Galiza
marcos.garcia.gonzalez@udc.gal

Abstract

This article describes a dependency-based strategy that uses compositional distributional semantics and cross-lingual word embeddings to translate multiword expressions (MWEs). Our unsupervised approach performs translation as a process of word contextualization by taking into account lexico-syntactic contexts and selectional preferences. This strategy is suited to translate phraseological combinations and phrases whose constituent words are lexically restricted by each other. Several experiments in adjective-noun and verb-object compounds show that mutual contextualization (co-compositionality) clearly outperforms other compositional methods. The paper also contributes with a new freely available dataset of English-Spanish MWEs used to validate the proposed compositional strategy.

1 Introduction

In the field of compositional distributional semantics there have been some interesting research, though not too much, making use of a syntax-sensitive vector space to compose the meaning of phrases and sentences (Erk and Padó, 2008; Thater et al., 2010; Erk et al., 2010; Weir et al., 2016). In those approaches, dependency-based combination of vectors enables words to be disambiguated as a process of contextualization. More precisely, given two words, a and b, related by a syntactic dependency (r), the meaning of the corresponding composite expression is actually two contextualized senses: a', which is the contextualized sense of a resulting from combing this word with the selectional restrictions imposed by b in relation r; and b', which stands for the contextualized sense of b as a result of combining this word with the restrictions imposed by a in r.

Moving towards a multilingual scenario, the objective of this paper is to apply this unsupervised method to a bilingual vector space so as to model translation as a process of compositional contextualization. In this regard, we first create contextualized vectors using selectional preferences, and then we generate possible translations by taking advantage of cross-lingual word embeddings learned from monolingual corpora. The results of several experiments in English-Spanish adjective-noun and verb-object compounds show that mutual contextualization (or co-compositionality) clearly outperforms other compositional methods.

Additionally, this paper also contributes with a new freely available dataset of 273 English-Spanish compound equivalents. This new resource contains multiword expressions (MWEs) with different degrees of semantic compositionality (free combinations such as *use a computer*, collocations –for instance, *hard drug*–, light-verb construction –e.g., *take a cab*–, or idioms such as *lone wolf*), which are useful to evaluate translation strategies using compositional approaches. It is worth noting that MWEs can fall into a wide spectrum of compositionality, from compositional compounds to idiomatic expressions (Cordeiro et al., 2019). To restrict the object of study, in this article, we focus on a specific subset of MWEs: adjective-noun and verb-noun compounds.

The rest of this article is organized as follows. Section 2 describes the compositional translation method. In Section 3 we describe the English-Spanish dataset and use it to evaluate the proposed strategy. Then, some related work is presented in Section 4. Finally, Section 5 addresses conclusions, drawbacks of the strategy and future work.

2 Compositional Translation with Cross-Lingual Embeddings

The proposed method consists of two main tasks: i) the construction of contextualized word mean-

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 40–48
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

ing by means of a syntax-sensitive compositional distributional strategy (see 2.1); ii) word contextualization in a bilingual vector space allowing the translation of compounds (See 2.2). We will focus on the translation two-word compounds encoded through a single syntactic dependency.

2.1 Compositional Distributional Meaning

We abandon the traditional choice of representing the meaning of a phrase or sentence as a single vector. In our approach, the meaning of a composite expression is represented by a contextualized vector for each constituent word rather than by a single vector standing for the entire expression (Erk and Padó, 2008; Weir et al., 2016; Gamallo, 2017). This is in accordance with the main postulates of Dependency Grammar which only defines linguistic categories for words and relations, but not for composite units such as phrases or sentences.

Let us take the dependency (r, h, d), where r is a binary relation between the head word, h, and the dependent one, d. This dependency can be used to yield two lexico-syntactic contexts:

$$(\downarrow r, h) \tag{1}$$

$$(\uparrow r, d) \tag{2}$$

where $\downarrow r$ and $\uparrow r$ are the head and dependent roles of relation r, respectively. The tuple in 1 represents a lexico-syntactic context of word d while tuple 2 is a context of h. Given these two contexts, the meaning of a binary dependency is represented by two contextualized vectors: $\mathbf{h}_{(\downarrow\mathbf{r},\mathbf{d})}$ and $\mathbf{d}_{(\uparrow\mathbf{r},\mathbf{h})}$, which are defined as follows:

$$\mathbf{h}_{(\downarrow\mathbf{r},\mathbf{d})} = \mathbf{h} + \mathbf{d}^{\uparrow r} \tag{3}$$

$$\mathbf{d}_{(\uparrow\mathbf{r},\mathbf{h})} = \mathbf{d} + \mathbf{h}^{\downarrow r} \tag{4}$$

where $\mathbf{h}^{\downarrow r}$ and $\mathbf{d}^{\uparrow r}$ are vectors representing selectional preferences, more precisely, $\mathbf{h}^{\downarrow r}$ stands for the selectional preferences imposed by the head, h, to the dependent word, d, and $\mathbf{d}^{\uparrow r}$ represents those imposed by the dependent one to the head. So, the contextualized sense of a word is the result of adding (by component-wise vector sum) its direct vector with another one representing the selectional preferences imposed by the word linked to it in the syntactic dependency. Head and dependent selectional preferences are defined as follows:

$$\mathbf{h}^{\downarrow r} = \frac{1}{N} \sum_{d:(\downarrow r,d)\in Sal_{\downarrow r}(h)} \mathbf{d} \tag{5}$$

$$\mathbf{d}^{\uparrow r} = \frac{1}{N} \sum_{h:(\uparrow r,h)\in Sal_{\uparrow r}(d)} \mathbf{h} \tag{6}$$

where $Sal_{\downarrow r}(h)$ and $Sal_{\uparrow r}(d)$ are two sets of salient contexts: the most salient contexts of the head, h, with the role $\downarrow r$ and the salient contexts of the dependent d with the role $\uparrow r$, N being the cardinality of each set. The set of salient contexts of a word consists of its top-N contexts ranked using a lexical association measure (e.g., PPMI, *log-likelihood*, etc). The top-N contexts are considered to be the most *salient* and informative for the given word. The summation runs through the lemmas that make up the salient contexts in equations 5 and 6. Equation 5 defines the *head preferences* and Equation 6 the *dependent preferences*.

Let us take an example. The dependency $(amod, drug, hard)$, from the compound *"hard drugs"*, gives rise to two contextualized senses:

$$\mathbf{drug}_{(\downarrow\mathbf{amod},\mathbf{hard})} = \mathbf{drug} + \mathbf{hard}^{\uparrow amod} \tag{7}$$

$$\mathbf{hard}_{(\uparrow\mathbf{amod},\mathbf{drug})} = \mathbf{hard} + \mathbf{drug}^{\downarrow amod} \tag{8}$$

The resulting vector in Equation 7 is the contextualized sense of *drug* as being modified by the adjective *hard*, while the vector in 8 represents the contextualized sense of *hard* when it modifies the noun *drug*. The selectional preferences imposed by the noun (head preferences), noted $\mathbf{drug}^{\downarrow amod}$, are actually the result of adding the vectors of the most representative (salient) adjectives modifying that noun, divided by the number of representative adjectives. Intuitively, it represents the main properties of drugs, for instance, *psychoactive*, *hallucinogenic* and *illicit* are the three more salient adjectives modifying the noun *drug* in our experiments. On the other hand, the selectional preferences imposed by the adjective (dependent preferences), and noted $\mathbf{hard}_{(\uparrow\mathbf{amod},\mathbf{drug})}$, are the result of adding the vectors of the most representative nouns modified by the adjective, divided by the number of representative nouns. So, it represents the set of most salient *hard things*; for example, *bop*, *disc* and *rock* are the three most salient nouns modified by the adjective *hard* in our corpus.

English dependency	**Spanish candidates**
(*amod, drug, hard*)	(*amod, medicamento, duro*) , (*amod, medicamento, difícil*) (*amod, medicamento, fácil*) , (*amod, medicamento, imposible*) (*amod, medicamento, arduo*) , **(*amod, droga, duro*)** (*amod, droga, difícil*) , (*amod, droga, fácil*) (*amod, droga, imposible*) , (*amod, droga, arduo*) (*amod, estupefaciente, duro*) , (*amod, estupefaciente, difícil*) (*amod, estupefaciente, fácil*) , (*amod, estupefaciente, imposible*) (*amod, estupefaciente, arduo*) , (*amod, cocaína, duro*) (*amod, cocaína, difícil*) , (*amod, cocaína, fácil*) (*amod, cocaína, imposible*) , (*amod, cocaína, arduo*) (*amod, fármaco, duro*) , (*amod, fármaco, difícil*) (*amod, fármaco, fácil*) , (*amod, fármaco, imposible*) (*amod, fármaco, arduo*)

Table 1: 25 Spanish candidate translations of the English collocation *"hard drug"*. Only the one in bold is an acceptable translation. The English *drug* was translated into Spanish by: *medicamento* (*medicine*), *droga* (*narcotic*), *estupefaciente* (*narcotic*), *cocaína* (*cocaine*), and *fármaco* (*medicine*). And the adjective *hard* was translated by: *duro* (*hard*), *difícil* (*difficult*), *fácil* (*easy*), *imposible* (*impossible*), and *arduo* (*arduous*). We added the most common English translation of each Spanish word so that readers who do not know Spanish will understand the ambiguity issue.

2.2 Compositional Translation of Dependencies

The compositional translation of an expression syntactically codified in a binary dependency consists of three steps: i) generation of translation candidates in the target language, ii) construction of the compositional meaning of the source dependency and the candidates in the target language, and iii) selection of the most similar candidate to the source dependency.

The input of the system is a dependency in the source language which is expanded into a set of candidate translations in the target language by making use of a translation lexicon automatically built with cross-lingual embeddings and Cosine similarity. For instance, let us take an English-Spanish translation lexicon and select the five most similar nouns to *drug* and the five most similar adjectives to *hard*. Taking into account these translations, the English dependency $(amod, drug, hard)$ is expanded in the 5x5 Spanish candidates shown in Table 1.

Once the candidates have been generated, the next step is to build the compositional vectors (contextualized senses) of both the input dependency and translation candidates, by making use of the algorithm used in the previous sub-section (2.1) and the cross-lingual embeddings of the previous step.

Finally, the compositional vectors of the candidates are compared pairwise with the source compositional vectors by means of cosine similarity and the most similar is selected. For the binary dependency in the source language, a translation candidate is selected by computing the contextualized translation measure, CT, which selects the most similar dependency in the target language by comparing the degree of similarity between heads and dependents in both languages. More precisely, given a dependency (r, h, d) in the source language, its translation into the target language is computed as follows:

$$CT(r,h,d) = \operatorname*{arg\,max}_{(r',h',d')\in\phi} \frac{S(\mathbf{h}_{(\downarrow r,d)}, \mathbf{h'}_{(\downarrow r',d')}) + S(\mathbf{d}_{(\uparrow r,h)}, \mathbf{d'}_{(\uparrow r',h')})}{2} \tag{9}$$

where (r', h', d') is any target dependency belonging to the set of translation candidates, ϕ. The first S computes the similarity between the two compositional vectors derived from the contextualized heads in the two languages. The second one computes the similarity between the vectors derived from the contextualized dependent words. So, CT is nothing more than the overall similarity between two composite expressions, which is the addition mean of the similarity scores obtained by comparing their head-based and dependent-based compositional vectors. The resulting translation is, thus, the composite expression belonging to ϕ with the highest overall similarity score.

3 Experiments

To have an idea about the quality of compositional vectors, most of the research done so far has made use of monolingual datasets prepared to measure the correlation between individual human similarity scores and the system's predictions (Mitchell and Lapata, 2008; Grefenstette and Sadrzadeh, 2011). Nonetheless, we consider that translation of composite expressions and MWEs is a more reliable way of evaluating the quality of compositional strategies. For instance, it is not clear whether *blue car* is semantically closer to *red car* than to *yellow car*, however, no one doubts that the Spanish translation of *red car* is *coche rojo*. In order to allow an evaluation based on compositional translation, we have created two bilingual datasets with MWEs syntactically coded by means of two dependencies: adjective-noun ($amod$) and verb-noun ($vobj$).

3.1 Test Datasets

To evaluate our compositional translation algorithm, it is required a bilingual resource containing a set of phrases with a simple syntactic structure in the source language with their possible translations into the target language. As there is no such resource, we decided to generate it by taking advantage of a free list of multilingual MWEs which was obtained using parallel corpora (Garcia, 2018).

The method presented in the referred paper extracts candidates of syntactic collocations using PPMI and frequency thresholds, and then identifies multilingual equivalents using bilingual word embeddings. From this resource, we selected 200 English-Spanish examples: 100 bilingual equivalents of adj-noun (*amod*) collocations (e.g., *facial hair*), and 100 verb-object (*vobj*) examples (e.g., *take [a] cab*). These lists were manually reviewed and enlarged with more possible translations, obtaining a final resource of 273 English-Spanish pairs (92 *amod* expressions with 143 translations, and 83 *vobj* English examples with 130 Spanish equivalents).

It is worth mentioning that as these lists were built using statistical association measures they contain not only phraseological combinations, but also other expressions with different degrees of semantic compositionality: free combinations (*use [a] computer*), true collocations (e.g., *deep condolence*, and also light-verb constructions such as *take [a] cab*), terms (*sulfuric acid*), quasi-idioms (*buy [the] silence*), or idioms (*lone wolf*) (Mel'čuk, 1998).[1] Thus, this variety of expressions converts the lists into a valuable resource for evaluating the translation of adj-noun and verb-object instances. [2]

3.2 Corpora and Distributional Models

In order to build bilingual compositional vectors, we made use of English and Spanish wikipedias (dumps files of December 2018), with 21 and 5 billion words, respectively. The two wikipedias were PoS tagged and syntactically analyzed with LinguaKit (Gamallo et al., 2018). The syntactically analyzed corpus was the basis for the elaboration of the salient lexico-syntactic contexts with which we constructed selectional preferences and contextualized vectors. Preliminary experiments were performed to find the best configuration, which was set to 50 salient contexts per lemma/PoS tag pair.

Bilingual embeddings were created with VecMap (Artetxe et al., 2018a) by using the supervised configuration and an open available English-Spanish dictionary, Apertium, containing 6,249 nouns, verbs, and adjectives.[3]. To make the evaluation fairer, we have removed from the dictionary all English words belonging to the test datasets. The original embeddings mapped by VecMap were created with Word2Vec, configured with CBOW algorithm, window 5, and 300 dimensions (Mikolov et al., 2013b). Word2Vec was applied on PoS tagged wikipedias and each token was coded as a lemma/tag pair. The bilingual mapped models with lemma/tag embeddings are made freely available.[4]

3.3 Translation Candidates

Using the bilingual vectors built from Wikipedia, each English word appearing in the test datasets was associated with the 10 most similar Spanish words and, so, each English binary dependency of the dataset was expanded with 10x10 candidate

[1] Note, however, that in ambiguous cases, the compositional translation was preferred (e.g., *cut [a] cable*).

[2] Both datasets have been added as suplementary material to the submission

[3] https://github.com/apertium/apertium-trunk

[4] https://ufile.io/lrze1 (anonymous account)

Spanish dependencies. It means that each English expression was compared with 100 Spanish translation candidates. It is worth pointing out that the correct translation is not always present in the 100 candidates. Yet, previous experiments allowed us to verify that increasing the number of translation candidates did not improve the final results.

3.4 Evaluation

To evaluate our compositional strategy, *CT(head+dep)*, which combines both head and dependent contextualized words (see equation 9), we compared its performance to five other approaches: *CT(head)*, which only considers the contextualized head; *CT(dep)*, which only takes into account the contextualized dependent word; *mult*, which combines the vectors of the two related words by pairwise multiplication; *add*, which combines vectors by pairwise addition; and *corpus*, which implements the corpus-based strategy described in (Grefenstette, 1999) by just selecting the most frequent translation candidates in the Spanish corpus. All strategies but *corpus* use the same bilingual word embeddings and the same similarity measure (cosine) between compositional vectors.

Additionally, we also included UNdreaMT in the evaluation. UNdreaMT is a recent neural machine translation system which uses monolingual corpora and cross-lingual word embeddings to learn translation models in an unsupervised way (Artetxe et al., 2018b). In the learning process, UNdreaMT applies backtranslation and uses a single shared encoder for both languages. To compare our compositional strategy with UNdreaMT, this system was trained with exactly the same monolingual corpora and word embeddings used by the other models. As UNdreaMT works with surface structures (and not dependency pairs), we adapted the input to not harm the system (e.g., *package,bring* → *bring the package*). Also, we manually modified the output to adapt it to the gold-standard format (e.g., *básico instinto* → *instinto,básico*).

Table 2 shows the results of all these methods on the two datasets (*amod* and *vobj*) described above. The table shows the accuracy, which is the number of correct translations divided by the number of different English expressions (source language) in each dataset. It is worth noting the significant difference between the proposed strategy, *CT(head+dep)*, and the rest of methods. The two methods based on just one contextualized word, *CT(head)* and *CT(dep)*, obtain similar scores to the well-known baselines, *mult* and *add*, as well as to the unsupervised MT strategy implemented with UNdreaMT. However, all these systems reached values far below those obtained by *CT(head+dep)* combining the two contextualizations within the dependency. Going into more detail, vector addition (*add*) outperforms vector multiplication (*mult*) in the two datasets, and also the contextualized dependent word performs better than the contextualized head in the two datasets. Finally, *corpus* gets the lowest values of all the compared methods.

System	*amod*	*vobj*
CT(head+dep)	**0.847**	**0.843**
UNdreaMT	0.543	0.571
CT(dep)	0.510	0.564
CT(head)	0.462	0.400
add	0.543	0.564
mult	0.354	0.505
corpus	0.326	0.297

Table 2: Accuracy of our system, *CT(head+dep)*, on English-Spanish *amod* and *vobj* expressions, compared to UNdreaMT and to five baseline methods: contextualized dependent (*CT(dep)*), contextualized head (*CT(head)*), vector addition (*add*), vector multiplication (*mult*), and corpus-based strategy (*corpus*).

3.5 Error Analysis

We carried out an error analysis of the *CT(head+dep)* model to know in detail in what types of expressions our strategy fails. So every wrong translation of the system was analyzed and classified into the following five error types (see Table 3 for quantitative results):

DistSimil: the most frequent errors arose from the distributional strategy (they are common in other vector-based approaches), since words belonging to different semantic relations (e.g., antonyms) may have very similar vectors. In our experiments, *CT(head+dep)* translated *male victim* by *víctima femenina* (*female victim*), or *take a cab* as *tomar un furgón* (*take a van*).

Conventions: another frequent source of errors was the generation of expressions which do not collocate, e.g., they do not follow the conventions

Type	*amod*	*vobj*	*Total*
DistSimil	42.86	66.67	53.85
Convention	21.43	25	23.08
Translation	14.29	8.33	11.54
Idiomacity	14.29	0	7.69
DataProcess	7.14	0	3.85

Table 3: Error classification (type and percentage) of the *CT(head+dep)* system. *Total* values are the micro-average.

of the target language, even if the meaning is transparent. In this regard, *fill a report* was translated by *llenar un informe* instead of *rellenar un informe* (both verbs in Spanish mean *to fill*, but *llenar* is most used for physical objects, e.g., *llenar el vaso*, *fill the glass*). Similarly, the system generated *evidencia verdadera* (instead of *evidencia real*) from *real evidence*.

Translation: 11% of the errors were approximate translations which do not appear in the dataset. This includes some combinations which may have slightly different meaning (depending on the context), such as *próxima década* and *siguiente década* (from *next decade*), and cases of polysemy: *share a cell*, where *cell* may refer to a biological cell (*célula* in Spanish), and a room in a prison or a part of a spreadsheet (both translated as *celda*).

Idiomacity: some non-compositional expressions were not correctly translated, such as *lone wolf* (which usually refers to a person and not to an animal), which was translated as *lobo indefenso* (*vulnerable* or *defenseless wolf*).

Data processing: finally, few errors emerged from problems in the data (or in its preprocessing: tokenization, lemmatization, etc.). As an example, the noun in *industrial area* was translated by *area* (which does not exist in Spanish) instead of *área*.

3.6 Discussion on Co-Compositionality

The high accuracy reached by the strategy based on the two contextualizations seems to verify the co-compositionality hypothesis (Pustejovsky, 1995), which states that the head word imposes selectional restrictions on the dependent one, while this one also imposes its restrictions on the former. It follows that a syntactic dependency between two words carries two complementary selective functions, each one imposing its own selectional preferences. These two functions allow the two related words to mutually disambiguate or discriminate the sense of each other by co-composition

However, co-compositionality has not been considered by many formal semantic approaches. In most approaches to formal semantics, inspired by Categorial Grammar, the interpretation of composite expressions such as *"hard drug"* relies on a rigid function-argument structure. In an adjective-noun construction, the adjective denotes an unary function applied to the noun denotation. Any syntactic dependency between two lexical words is generally represented in the semantic space as the assignment of an argument to a lexical function which impose its selectional preferences. There is just one direction in the process of contextualization: the word representing the lexical function contextualizes (imposes its preferences to) the word representing the passive argument. This one-way compositional procedure is also present in some work on distributional compositional semantics (Baroni et al., 2014; Grefenstette and Sadrzadeh, 2011). Unfortunately, a comparison with these one-way strategies has not been possible because they have not yet been applied to compositional translation.

4 Related Work

The proposed compositional method integrates three different tasks: to build compositional vectors representing the contextualized sense of composite expressions; to build cross-lingual word embeddings from monolingual corpora; to propose contextualized translations with compositional and cross-lingual vectors.

The basic approach to distributional composition is to combine vectors of two syntactically related words with arithmetic operations: addition and component-wise multiplication (Mitchell and Lapata, 2008, 2009, 2010). This approach is not strictly compositional since it does not take into account the syntactic structure underlying the expression. It does not consider the function-argument relationship underlying compositionality in Categorial Grammar approaches (Montague, 1970).

Other approaches propose compositional models inspired by Categorial Grammar. Some induce the compositional meaning of functional words from examples adopting regression techniques commonly used in machine learning (Baroni and Zamparelli, 2010; Krishnamurthy and Mitchell, 2013; Baroni, 2013; Baroni et al., 2014), and others use tensor products for composition (Coecke et al., 2010; Grefenstette et al., 2011). Although compositional, none of them is based on co-compositional strategy, like ours.

There are also studies making use of neural-based approaches, namely bidirectional long short-term memory networks, to deal with word contextualization (Melamud et al., 2016; McCann et al., 2017; Peters et al., 2018). However, word contextualization is not defined by means of syntax-based compositional functions, as they do not consider the syntactic functions of the constituent words.

As has been said, our compositional approach is inspired by the work described in Erk and Padó (2008) and Erk et al. (2010), in which second order vectors represent selectional preferences and each word combination gives rise to two contextualized word senses. More recently, Weir et al. (2016) describe a similar approach where the meaning of a sentence is represented by the contextualized senses of its constituent words. Each word occurrence is modeled by what they call *anchored packed dependency tree*, which is a dependency-based graph that captures the full sentential context of the word. The main drawback of this context-based approach is its critical tendency to build very sparse word representations. Our approach is an attempt to join the main ideas of these syntax-sensitive models (namely, the use of selectional preferences and two returning word senses per combination) in order to apply them to contextualized translation.

The method proposed in this paper also relies on count-based techniques to build bilingual vectors from monolingual corpora (Fung and McKeown, 1997; Rapp, 1999; Saralegi et al., 2008; Ansari et al., 2014). Neural-based strategies also have been used to learn translation equivalents from word embeddings (Mikolov et al., 2013a; Artetxe et al., 2016, 2018a). They learn a linear mapping between embeddings in two languages that minimizes the distances between equivalences listed in a bilingual dictionary.

Finally, many approaches to compositional translation of phrases and composite terms consist in decomposing the source term into atomic components, translating these components into the target language and recomposing the translated components into target terms (Delpech et al., 2012; Morin and Daille, 2012; Tanaka and Baldwin, 2003; Grefenstette, 1999). Selection of the best translation candidate is performed by means of corpus-based searching. However, this strategy has not yielded good results in the experiments described in the previous section. Our translation approach also follows the decomposing strategy but, unlike the works cited above, we use compositional/contextualized vectors to select the best candidate instead of basic corpus-based frequencies.

5 Conclusions

In this article, we tried to show that it is possible to apply compositional distributional semantics on a bilingual vector space to propose contextualized translations.

However, the proposed contextualization method has several drawbacks that need to be addressed in future work. First, it will be necessary to deal with *fertile translations*, i.e. translations in which the target term has a different number of words (and so a different syntactic structure) than the source one. For this purpose, we will expand the set of translation candidates by making use of a great variety of extraction strategies as, for instance, a Mel'čuk-based strategy consisting of identifying similar words to the *base* of a collocation (Mel'čuk, 1998). Second, our method does not distinguish between compositional and non-compositional expressions. It will probably be necessary to first identify the degree of compositionality of the source MWE before choosing the compositional translation strategy that best suits that expression (Cordeiro et al., 2019). And third, increasingly complex expressions consisting of more than one dependency will have to be dealt with. For this purpose, the method will have to be generalized to any input sentence with any syntactic structure, giving rise to an unsupervised machine translation approach.

Acknowledgments

This work has received financial support from the DOMINO project (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08) and the European Regional Development Fund (ERDF), 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation), Juan de la Cierva-incorporación grant (IJCI-2016-29598). Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan Xp GPUs used for this research.

References

Ebrahim Ansari, M. H. Sadreddini, Alireza Tabebordbar, and Mehdi Sheikhalishahi. 2014. Combining different seed dictionaries to extract lexicon from comparable corpus. *Indian Journal of Science and Technology*, 7(9):1279–1288.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Marco Baroni. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7:511–522.

Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *LiLT*, 9:241–346.

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP'10, pages 1183–1193, Stroudsburg, PA, USA.

B. Coecke, M. Sadrzadeh, and S. Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1-4):345–384.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*. Impact Factor: 1.319. http://www.mitpressjournals.org/doi/pdf/10.1162/coli_a_00341.

Estelle Delpech, Béatrice Daille, Emmanuel Morin, and Claire Lemaire. 2012. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In *COLING 2012, 24th International Conference on Computational Linguistics, Mumbai, India*, pages 745–762.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, Honolulu, HI.

Katrin Erk, Sebastian, Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translation from non-parallel corpora. In *5th Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

P. Gamallo, M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. Linguakit: A big data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Pablo Gamallo. 2017. The role of syntactic dependencies in compositional distributional semantics. *Corpus Linguistics and Linguistic Theory*, 13(2):261–289.

Marcos Garcia. 2018. Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents. In Stella Markantonatou, Carlos Ramisch, Agata Savary e Veronika Vincze, editor, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop. Phraseology and Multiword Expressions 3*, pages 319–342. Language Science Press.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Conference on Empirical Methods in Natural Language Processing*.

Edward Grefenstette, Mehrnoosh Sadrzadeh, Stephen Clark, Bob Coecke, and Stephen Pulman. 2011. Concrete sentence spaces for compositional distributional models of meaning. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pages 125–134.

Gregory Grefenstette. 1999. The World Wide Web as a resource for example-based machine translation tasks. In *Translating and the Computer 21: Proceedings of the 21st International Conference on Translating and the Computer*.

Jayant Krishnamurthy and Tom Mitchell. 2013. *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, chapter

Vector Space Semantic Parsing: A Framework for Compositional Vector Space Models. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *CoRR*, abs/1708.00107.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61. Association for Computational Linguistics.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony P. Cowie, editor, *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Claredon Press, Oxford.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 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*, pages 746–751, Atlanta, Georgia.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244.

Jeff Mitchell and Mirella Lapata. 2009. Language models based on semantic composition. In *Proceedings of EMNLP*, pages 430–439.

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

Richard Montague. 1970. Universal grammar. theoria. *Theoria*, 36:373–398.

Emmanuel Morin and Béatrice Daille. 2012. Revising the compositional method for terminology acquisition from comparable corpora. In *COLING 2012, 24th International Conference on Computational Linguistics, Mumbai, India*, pages 1797–1810.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *ACL'99*, pages 519–526.

X. Saralegi, I. San Vicente, and A. Gurrutxaga. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In *LREC 2008 Workshop on Building and Using Comparable Corpora*.

Takaaki Tanaka and Timothy Baldwin. 2003. Noun-noun compound machine translation a feasibility study on shallow processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 17–24, Sapporo, Japan.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 948–957, Stroudsburg, PA, USA.

David J. Weir, Julie Weeds, Jeremy Reffin, and Thomas Kober. 2016. Aligning packed dependency trees: A theory of composition for distributional semantics. *Computational Linguistics*, 42(4):727–761.