Weighted compositional vectors for translating collocations using monolingual corpora.*

Marcos Garcia[1,2][0000−0002−6557−0210]
Marcos García-Salido[1][0000−0002−5667−2119]
Margarita Alonso-Ramos[1,2][0000−0002−1353−9270]

[1] Universidade da Coruña, Grupo LyS, Departamento de Letras
Campus da Zapateira, 15071, Coruña, Galicia, Spain
[2] Universidade da Coruña, CITIC, Campus de Elviña, 15701, Coruña, Galicia, Spain

NOTICE: this is the final peer-reviewed manuscript that was accepted for publication in *EUROPHRAS 2019: Computational and Corpus-Based Phraseology*. Changes resulting from the publishing process, such as editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version will be published in *Computational and Corpus-Based Phraseology. EUROPHRAS 2019*, LNCS.
DOI: 10.1007/978-3-030-30135-4_9.

Abstract. This paper presents a method to automatically identify bilingual equivalents of collocations using only monolingual corpora in two languages. The method takes advantage of cross-lingual distributional semantics models mapped into a shared vector space, and of compositional methods to find appropriate translations of non-congruent collocations (e.g., *pay attention–prestar atenção* in English–Portuguese). This strategy is evaluated in the translation of English–Portuguese and English–Spanish collocations belonging to two syntactic patterns: *adjective-noun* and *verb-object*, and compared to other methods proposed in the literature. The results of the experiments performed show that the compositional approach, based on a weighted additive model, behaves better than the other strategies that have been evaluated, and that both the asymmetry and the compositional properties of collocations are captured by the combined vector representations. This paper also contributes with two freely available gold-standard data sets which are useful to evaluate the performance of automatic extraction of multilingual equivalents of collocations.

Keywords: multilingual collocations · distributional semantics · compositional semantics.

* This research was supported by a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation), by Ministerio de Economía, Industria y Competitividad (project with reference FFI2016-78299-P), and by the Galician Government (Xunta de Galicia grant ED431B-2017/01). Marcos Garcia has been funded by a Juan de la Cierva-incorporación grant (IJCI-2016-29598), and Marcos García-Salido by a post-doctoral grant from Xunta de Galicia (ED481D 2017/009).

1 Introduction

Bilingual models of distributional semantics have been used to automatically find word translations from both parallel and comparable corpora between several languages [8,10,32]. Besides, and also using parallel and comparable corpora, several attempts have been made to identify translations of different types of multiword expressions (MWEs), using external resources such as bilingual dictionaries or cross-lingual models of distributional semantics [2,11,15,34,38]. Most of these approaches, however, often consider the semantic load of each MWE component as similar, so that they obtain candidates in the target languages as word-to-word translations, which are then filtered and ranked using different metrics. For instance, the decomposition of the French MWE *examen clinique* may first generate different English candidate translations (*clinic inspection*, *clinical test*, *clinical review*, etc.), among which the most probable translations are then selected (e.g., *clinical examination*) [30].

Even if these strategies are well suited for some types of collocations (e.g., those in which the translation of each component is predictable), they fail to capture other cases where the meaning of one of their components is not a direct translation in the target language, that is, it is not a congruent equivalent. In this respect, we follow [31] and consider congruent equivalents those where a word-to-word translation generates sound combinations in both languages (e.g., English–Portuguese *formulate [a] hypothesis–formular [uma] hipótese*), and non-congruent those cases where this definition does not apply, such as the English–French translation *take [a] walk–faire une promenade* (where *faire* 'to do' is not the literal translation of *to take*). This is the case of some collocations, which are asymmetric combinations of two lexical units, where one of them (the *base*) is freely selected by the speaker due to its meaning, while the other (the *collocate*, whose selection is restricted by the base) conveys a particular meaning in this specific combination [24,25]. In the previous examples, *hypothesis/hipótese* and *walk/promenade* are the bases, while *to formulate/formular* and *to take/faire* are the collocates. From this point of view, the bases are often defined as *autosemantic* and the collocates as *synsemantic* [19].

For non-congruent collocations, finding the appropriate collocates in the target language is a challenging task both using bilingual dictionaries or distributional semantics models, since the source and the target words may not be semantically related except in that particular combinations.

In the field of distributional compositional semantics, several studies have proposed different methods to obtain vector representations of large expressions such as phrases and sentences [3,28,29,33]. These approaches vary from vector multiplication and addition to the creation of contextualized representations for each component word, later combined in compositional vectors which produce better representations of words in a particular context [42].

With that in mind, this paper explores the compositionality of collocations in distributional semantics by using compositional weighted vectors to translate this type of MWEs in English, Portuguese, and Spanish. For a given collocation in the source language, we obtain a compositional vector which allows us to

identify candidate collocates which are semantically similar to the whole combination, but not necessarily related to the source collocate. Thus, the proposed strategy is able to find both non-congruent and congruent candidate collocates, used to create equivalent collocations in the target language. Aimed at having a better view of the compositional properties of these expressions, we evaluate the impact of different weights to construct the compositional vector, by giving more relevance to the base than to the collocate, and vice versa.

To evaluate our approach, we manually created test sets in English–Portuguese and English–Spanish of two types of collocations (based on syntactic dependencies): *adj-noun* (e.g., *fresh water*), and *verb-obj* (e.g., *pay attention*). For each relation and language we selected 50 input collocations, obtaining a final resource of 200 inputs with 365 possible translations. The results of several tests, using only cross-lingual word embeddings trained on monolingual corpora, improve previous translation methods, and confirm both the compositional character of collocations as well as their internal asymmetry, whereby collocates convey a particular meaning in each specific combination.

The rest of this paper is organized as follows. Section 2 presents some related work about the different tasks carried out in our study. Then, in Section 3 we introduce the proposed method to translate collocations using monolingual corpora. The experiments to evaluate our approach are addressed in Section 4, while Section 5 concludes this work.

2 Related work

This paper deals with different tasks to perform unsupervised translation of collocations. Thus, this section briefly presents representative studies about bilingual collocation extraction, distributional compositional semantics as well as cross-lingual vector spaces.

The first approaches to identify bilingual equivalents of multiword expressions took advantage of parallel corpora and statistical measures to find translations of noun phrases or collocations [18,21,36]. Other strategies using bilingual corpora make use of syntactic information to identify predefined MWEs patterns [35,43]. More recently, [34] presented a system to extract collocation equivalents from parallel or comparable corpora using bilingual dictionaries and WordNet, while [39] uses a non-compositional approach by training cross-lingual distributional models in which collocations are treated as single units.

Grefenstette [17] describes a method which performs a word-to-word translation of MWEs using bilingual dictionaries, as well as ranking approach which relies on web frequency. Other works use comparable as well as unrelated corpora to extract MWEs equivalents between two languages [2,7,38]. They first identify the MWEs in a source language using pattern-based approaches and statistical metrics. Then, they generate candidate translations by applying a word-to-word approach, or extracting MWEs in the target data which are ranked using bilingual dictionaries or cross-lingual distributional models [15,30]. Cross-lingual models were also exploited in recent studies to find bilingual equivalents both

using single-word vectors [41] and contextualized representations [12,13]. In general, most of these studies select candidate MWEs in the target languages using bilingual dictionaries or vectors representing single words, which make difficult to correctly identify non-congruent translations.

Regarding compositionality in distributional semantics, classical methods combine the vectors of syntactically related words using algebraic operations such as addition or multiplication [28,29]. Other approaches obtain the compositional representation of a given expression using a functional approach, where the semantic arguments are represented by vectors while functional words (including verbs) are functions operating on those vectors [4,5].

Finally, our work also takes advantage of cross-lingual distributional semantics models which map in the same vector space representations of different languages. Apart from bilingual models trained in parallel corpora [40], both count-based techniques [9,32] and recent neural network algorithms [1,22,27] obtain high quality bilingual models using comparable and unrelated corpora.

The method proposed in this paper follows the phraseological perspective about the internal properties of collocations [24], and takes advantage of weighted additive models [29] to find candidate collocates in the target language for a given source combination. This strategy allows us to evaluate non-congruent combinations where the components of each MWE are not word-to-word translations nor distributionally similar to the source ones. It is worth noting that this paper deals only with collocation-to-collocation equivalents, so other translations (e.g., *obtain [a] doctorate–doutorar-se*, in English–Portuguese) are not addressed.

3 Compositional distributional semantics for collocations

In this section we first discuss some particular issues that should be taken into account when dealing with multilingual collocations equivalents, and then we present the method proposed to tackle those questions.

3.1 Multilingual collocation equivalents

As several studies have shown, the word-to-word approach for selecting candidate collocations (and other MWEs) in a target language obtains good results in different languages and morphosyntactic patterns [15,30,38]. One of the reasons of this success is that a considerable part of compositional MWEs are to some extent congruent in several language pairs (e.g., *deep analysis–análise profunda*, in English–Portuguese, or *jefe de policía–police chief* in Spanish–English), so bilingual dictionaries or distributional models, in combination with statistical filters are suitable to avoid many incorrect translations which do not occur in the target language. As an example, the English verb *to fill* is usually translated as *encher* (in Portuguese) when collocates with physical objects (*encher o copo*, 'fill the glass'), but as *preencher* in figurative contexts (*preencher um relatório*, 'fill a report'). In Spanish, these translations are similar (*llenar* and *rellenar*), but the latter is also used as equivalent to *refill* (*rellenar un vaso*, 'refill the

glass'). These examples show the importance of considering the conventionality of these combinations even in those cases in which the candidate translations are relatively easy to find using word-to-word approaches.

A different case is the above mentioned non-congruent equivalents, in which the candidate collocates are not found using standard techniques. In this respect, the heads of light verb constructions (LVCs) such as in *take [a] picture*, or in *have breakfast* are often not semantically similar (when isolated) to their translations in the equivalent collocations in other languages (e.g., *tirar [uma] fotografia*, 'take a picture' in Portuguese, or *tomar [el] desayuno* 'have breakfast' in Spanish). Nevertheless, as these collocates are frequent support verbs in various languages, filtering strategies may help to find appropriate translations. However, other less frequent cases such as *pay attention* (where *to pay* is translated as *prestar*—literally 'to lend'—in both Portuguese and Spanish) are more difficult to identify, since the support verb has in most contexts a full meaning which is different than the one in that particular combination.

Similarly, many *adj-noun* collocation equivalents are non-congruent in several language pairs: for example, in *fresh water*, the adjective is usually translated as *doce* (Portuguese) or *dulce* (Spanish) (literally 'sweet' in both languages). In this respect, bilingual dictionaries (except those with collocational information) do not include *doce* or *dulce* as translations of *fresh*. Also, in standard cross-lingual distributional models the similarity between these words is close to 0.5, so it would be hard to select *água doce* or *agua dulce* as candidate collocations using word-to-word approaches.

These examples allow us to have a better view about some relevant properties concerning multilingual collocation equivalents, and also to note the importance of the idiosynchratic character of these expressions. Moreover, it is worth remembering that, although collocations are compositional, collocates often have particular meanings, in contrast to the bases (except, for obvious reasons, in polysemous words).

3.2 Weighted compositional vectors

To obtain a compositional vector of a given collocation, we follow [29] by using the *weighted additive models*. However, as we are focusing on the semantic properties of a base–collocate pair instead of the syntactic roles of a head–dependent relation,[3] we define a compositional vector v as the addition of the base and collocate weighted vectors b and c:

$$v = \alpha b + \beta c \tag{1}$$

where $\beta = 1 - \alpha$ (so that $\alpha + \beta = 1$). Thus, variations in α and β make the compositional vector asymmetric, allowing to better represent the semantic load of both the collocate and the base. Therefore, if we set $\alpha = 0.2$ and $\beta = 0.8$, the collocate will contribute more to the compositional vector, and vice versa.

[3] In this regard, note that in an *adj-noun* collocation, the syntactic head is occupied by the base, while in *verb-obj* collocations the syntactic head is the collocate.

Table 1: Collocate candidates (top) and extracted collocations (bottom) in Portuguese from the English input *pay attention*. Differences between using the vector of the *pay* and the weighted additive vector of *pay+attention* (0.5/0.5) to select collocate candidates in Portuguese. *Filter* means filtering out collocation candidates with frequency lower than 10 in the reference corpus. Floating-point numbers (in brackets) are the cosine distance between the source and target vectors, while ordinal numbers are the rank position in each list of candidates.

Input vector	**Collocate candidates**
pay	pagar (0.9), cobrar (0.7), custear (0.7), desembolsar (0.7), quantiar (0.7) [...], **prestar (0.5) ($37^{\mathbf{th}}$)**
pay+attention	pagar (0.7), cobrar (0.6), investir (0.6) [...], **prestar (0.6) ($7^{\mathbf{th}}$)**

Input vector	**Extracted collocations**
pay (no filter)	pagar atenção (0.9), cobrar atenção (0.8), pagar repercussão (0.8), custear atenção (0.8) [...]
pay (filter)	$\emptyset$
pay+attention (no filter)	pagar atenção (0.9), pagar elogio (0.8) pagar repercussão (0.8), cobrar atenção (0.8) [...] **prestar atenção (0.7)**
pay+attention (filter)	**prestar atenção (0.7)**

Taking the above into account, we define the method to identify bilingual collocations as follows: we use cross-lingual word embeddings to search for both base and collocate candidates in the target languages. For the bases, we simply use the source vector to retrieve the n most similar words (e.g., between 3 and 5) with the same PoS-tag. To search for candidate collocates we use the weighted compositional vectors described in Equation 1, which allow us to find collocate candidates (e.g., 10, also with the same PoS-tag) whose meaning is closer to the one of the whole combination than to the collocate alone. Then, both base and collocate candidates are combined to generate collocations in the target language, with the same pattern as the source one. For each of these collocation candidates, we also obtain the weighted average vector in order to compute its translation confidence (in terms of cosine similarity) with respect to the input collocation. This confidence value is used to rank the target collocations with regard to the source one. Optionally, we use a corpus-based filter method to remove unconventional combinations in the target language.

Table 1 shows an example of the proposed approach to generate the Portuguese translation of the English collocation *pay attention* (using the models and data described in Section 4). First, top rows contain the collocate candidates selected using the vector of the collocate *pay* as well as the weighted additional vector of *pay+attention* (α and β were set to 0.5 in this example). On one hand, it can be seen that most similar verbs to *pay* in Portuguese have economical or financial meanings, and that *prestar* appears as the candidate 37^{th}. On the other hand, the compositional vector reduces the load of its "economical mean-

ing" (e.g., *pagar* 'to pay' drops from 0.9 to 0.7 of cosine similarity) and promotes *prestar* to a higher position (7^{th} position).

Second, bottom rows display the selected collocation candidates from each set of collocates (generated using the top 10 and the top 5 candidates for the collocate and the base, respectively). The combinations created with the vector of *pay* do not include the correct translation *prestar atenção*, and none of the candidates passed the frequency filter ($f > 10$), so this method did not extract a suitable translation of *pay attention*. The compositional approach, however, generated the translation *prestar atenção* as one of the collocation candidates. This combination was the only one which passed the frequency filter and therefore was correctly selected as the Portuguese equivalent of *pay attention*.[4]

4 Experiments

In this section we carry out a set of experiments to evaluate the presented compositional approach using different weights for both the base and the collocate. This weighted model is compared to the following methods, all of them using the same cross-lingual models to select the translation candidates:[5] (i) baseline (*bas*), which creates a collocation selecting the most similar equivalents of both the base and the collocate in the target language (it is therefore a similar approach to those presented in [15,30]); (ii) addition (*add*), which generates at most 100 collocation candidates from the top 10 bases and collocates in the target language, and rank them by cosine similarity of the source and target compositional vectors $v = b + c$; (iii) multiplication (*mult*), the same as *add* but obtaining the collocation vectors by multiplication instead of addition ($v = b \cdot c$). Furthermore, we also evaluated a *non-compositional* strategy (*ncp*, explained below) which learns single vector representations for a given MWE. However, this method should not be directly compared to the other approaches, since the MWEs were added to the model from the gold-standard data sets (and not automatically selected). The frequency filter (applied to every model except for the baseline and non-compositional ones), selects only those collocations with a frequency $f > 10$ (using lemmas in the same dependency relation) in the 250M corpora used to train the distributional models (see below).

4.1 Test data

To evaluate the different methods we created two new gold-standard data sets as follows: First, we randomly extracted 100 English collocations (50 *adj-noun* and

[4] It is worth noting that the candidate *cobrar atenção*, which exists in Brazilian Portuguese with a slightly different meaning, could be selected if we had used more resources from this variety and from other typologies. In our data, mostly composed of corpora from Portugal (besides the Wikipedia, with mixed varieties), this combination has a frequency of only 1.

[5] Both the gold-standard data as well as the output of each system can be downloaded at `https://github.com/marcospln/bilingual_collocations`.

50 *verb-obj*) from an annotated corpus [16]. Then, we manually translated these 100 examples into the most natural collocations in Portuguese and Spanish, obtaining a total of 365 equivalents, 227 in Portuguese and 138 in Spanish (as said, this paper is focused on collocation-to-collocation equivalents).[6] The Portuguese data set has more translations mainly because we have added spelling variants with regard to the Orthographic Agreement of Portuguese Language[7] and to the Brazilian orthography (e.g., *actual/atual* or *crónico/crônico*).

4.2 Distributional models

To create the monolingual distributional models we compiled three corpora (for English, Portuguese, and Spanish) with 250 million tokens and similar properties: each of them is composed by 200 million tokens from Wikipedia, 20 millions from Europarl [20], 20 millions from news and web pages, and 10 million tokens from OpenSubtitles [23]. The sentences of each corpus were randomly selected, so the degree of comparability is low. Except for Wikipedia (which often contains mixed varieties), the European varieties (from UK, Portugal, and Spain) were preferred.

The texts were processed with LinguaKit [14] to convert them into *lemma_PoS-tag* corpora. They were also parsed with UDPipe [37] to obtain dependency triples used for the frequency filter. Then, the *lemma_PoS-tag* corpora were used to learn monolingual models using *word2vec* [26] (with the skip-gram algorithm, 300 dimensions, a window of 5 tokens, and a frequency threshold of 5). Finally, the monolingual models were mapped into a shared vector space with the semi-supervised mode of *vecmap* [1]: we used 100 numbers (0 to 99) and 300 randomly selected words (100 adjectives, 100 nouns, and 100 verbs) automatically translated and then reviewed by the authors.

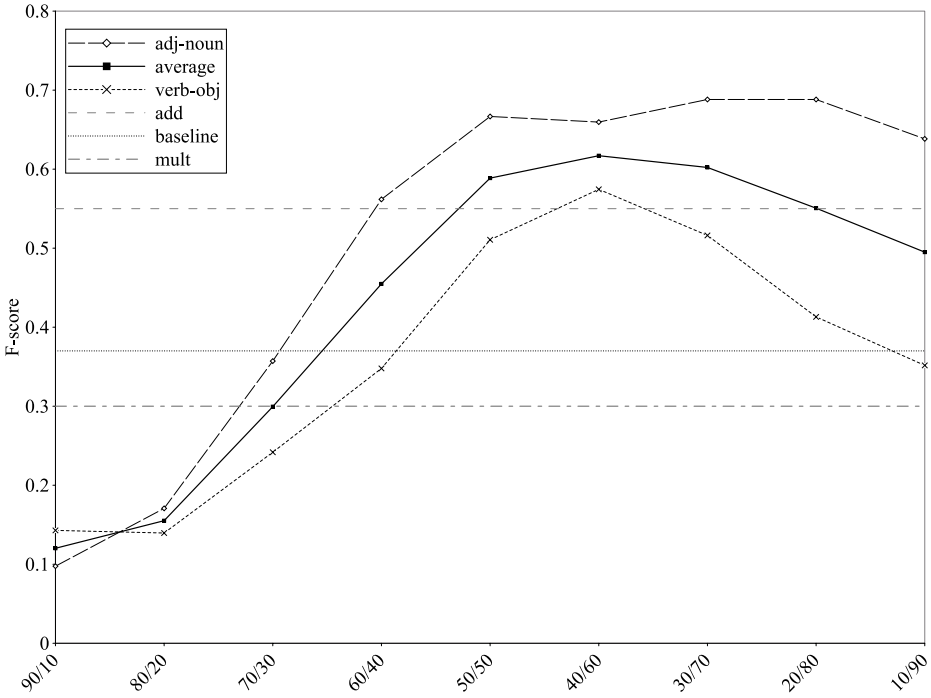
For the *ncp* method we used the same lemmatized corpora (without PoS-tags) representing the gold (source and target) collocations as single tokens (e.g., *fresh_water*), and trained *word2vec* models with them. For non-adjacent collocations —linked by a dependency relation— we replaced the first element by the whole combination and removed the intermediate lemmas (e.g., "to take a brief walk" → *take_walk*). To map the models we used the same bilingual pair words as well as 10 additional bilingual collocation equivalents (e.g., *take_walk–dar_passeio* in English–Portuguese), different from those of the gold-standard.

4.3 Results

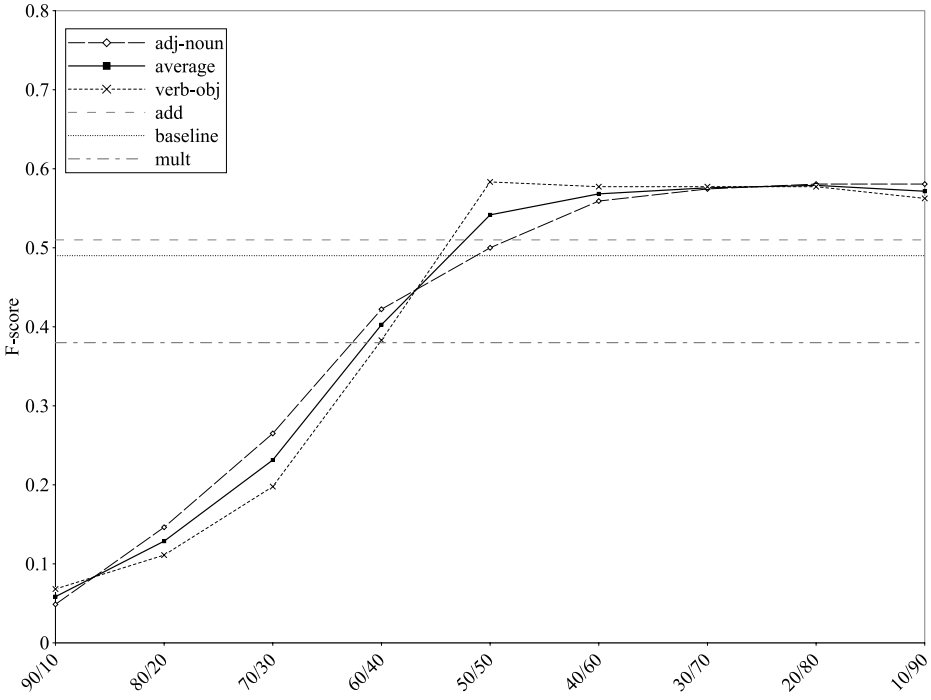
For each run we computed precision, recall, and f-score values for the top-1 and top-3 candidates. For the proposed compositional approach we evaluated 9 different weight ratios of the bases (α) and collocates (β): from $\alpha = 0.9, \beta = 0.1$ (90), to $\alpha = 0.1, \beta = 0.0$ (10).

[6] This gold-standard includes an average of 3.65 translations for each collocation, but, as discussed in Section 4.4, there may be more suitable equivalents for some of them.

[7] `https://en.wikipedia.org/wiki/Portuguese_Language_Orthographic_Agreement_of_1990`



(a) English–Portuguese



(b) English–Spanish

Fig. 1: F-score (top-3 results) *versus* base/collocate weight of the compositional vectors (from 90/10 to 10/90) for each collocation pattern in English–Portuguese (top) and English–Spanish (bottom) translations.

Table 2: Average precision, recall, and f-score using the top-1 translations (top), and the top-3 (bottom) in English-Portuguese and English-Spanish. Each column of *weighted vectors* shows the results of the proposed strategy using different base/collocate ratios (from 90/10 to 10/90). *mult* and *add* methods use multiplication and addition to obtain compositional vectors, while the baseline (*bas*) uses selects the most similar base and collocate from the source collocation. *ncp* (non compositional) learns a single vector for each source and target collocation.

		Weighted vectors									*mult*	*add*	*bas*	*ncp*
		90	80	70	60	50	40	30	20	10				
top-1	Precision	0.05	0.10	0.22	0.40	0.49	0.49	**0.50**	0.47	0.36	0.28	0.41	0.43	*0.95*
	Recall	0.03	0.07	0.16	0.33	0.42	**0.44**	**0.44**	0.42	0.32	0.28	0.41	0.43	*0.46*
	F-score	0.04	0.08	0.19	0.36	0.45	**0.47**	**0.47**	0.44	0.34	0.28	0.41	0.43	*0.61*
top-3	Precision	0.11	0.17	0.32	0.48	0.61	**0.63**	**0.63**	0.61	0.57	0.34	0.53	0.43	*0.98*
	Recall	0.08	0.12	0.23	0.39	0.53	**0.56**	**0.56**	0.53	0.50	0.34	0.53	0.43	*0.50*
	F-score	0.09	0.14	0.27	0.43	0.57	**0.59**	**0.59**	0.57	0.53	0.34	0.53	0.43	*0.65*

Table 2 displays the top-1 and top-3 results of each weighted model together with the other evaluated systems (average values of English–Portuguese and English–Spanish translations). In general, the weighted method achieved the best results both in top-1 and in top-3 translations (except for the non-compositional model, which obtained high precision values between 0.95 and 0.98).

About the base–collocate weights, the results show that to select collocate candidates for a given collocation the base should not contribute more than 30% or 40% to the compositional vector, being about 60% or 70% obtained by the collocate. With higher weights of the base, the performance is dramatically lower, while increasing the contribution of the collocate involves smaller drops.

The baseline achieved better results than *mult* (top-1 and top-3), and also than the *add* models (in the top-1 evaluation). Between these two compositional approaches, the addition method was better than *mult* in every scenario (in contrast to the results of other studies—not dealing specifically with collocations—such as [29]).

With regard to the different collocation patterns, Figures 1a and 1b plot the f-score values (top-3) of the translation of *verb-obj* and *adj-noun* collocations in English–Portuguese and English–Spanish, respectively. In the first case, *adj-noun* collocations were better translated than *verb-obj*, while in English–Spanish the results of both types were very similar. Also, in English–Portuguese, the baseline obtained noticeably worse results than both the weighted and the additive methods. If we look at the data, some of these differences in the English–Portuguese pair have arisen due to the incorrect translation of the support verb *to have* (e.g., *have [a] drink*) as *possuir* 'to own', thus generating incorrect combinations of LVCs using an incorrect collocate. These *verb-obj* results also involve a drop in the English–Portuguese values with collocate ratios of 70% and higher, which did not occur in the other language pair.

Finally, it is important to mention that, even if we should not compare the non-compositional approach to the compositional ones, the high precision achieved by this method suggests that this is a promising research line to translate both non-compositional and compositional (specially non-congruent) MWEs.[8] To make a fair comparison, we should have identified the source collocations using automatic methods, which will extract thousands of candidate collocations (many of them incorrect), thus introducing a lot of noise in the final models.

4.4 Error analysis

In order to have a better view of the performance of our method, we have carried out a brief error analysis of the translations of the weighted models with the best average results (top-3 outputs with a 30%/70% base/collocate ratio). They had a total of 65 errors with respect to the gold-standard data, which were classified into the following 4 groups (see Table 3 for the percentage of each type):

1. Distributional model: the most frequent error type was caused by the distributional method, which represents with very similar vectors word pairs such as antonyms. For example, *put spell* was translated to Spanish as *deshacer [el] hechizo* 'break [the] spell'.
2. Compositional method: the weighted compositional method produced ≈22% of the errors. Some of them appeared due to a non-optimal weight ratio, such as the English–Portuguese *make [a] step→fazer [um] passo* instead the correct *dar [um] passo*, which was selected by the models with a higher weight of the base ($b > 0.3$). Some other mistranslations were produced because the weighted vectors (with any base/collocate ratio) could not find a proper collocate candidate among the top 10 selected.
3. Filter: few errors (4) were due to the frequency filter method, which removed several good translations with low frequency in our corpora. Thus, the English collocation *excellent mother* was translated to Spanish as *buena madre* instead of *madre excelente* because the correct equivalent only appears once in the corpus. Similarly, *dark sorcerer* was translated to Portuguese as *feiticeiro escuro* since *feiticeiro negro* also has a frequency of 1 in the corpus.

* Disagreement between the gold-standard: the second most frequent source of errors is actually formed by suitable translations which differ from those of the gold-standard, so that they are not really mistranslations. For instance, *meet [the] requirement* was correctly translated as *reunir [el] requisito* in Spanish (instead of the gold *cumplir [el] requisito*), and *loud noise* produced the proper equivalent in Portuguese *barulho ensurdecedor* (but not *ruído alto* or *barulho alto*, the two translations present in the gold-standard).

On one hand, these results indicate that out of 65 errors, 24 were actually good translations which differ from those in the gold-standard. This fact also

[8] Note that the high precision of the non-compositional method is an evidence of the good performance of the distributional approach and of the cross-lingual mapping, but it does not necessarily imply that collocations are non-compositional.

Table 3: Percentage of each error type of the weighted additive model with a 30/70 base/collocate ratio. *Disagreement* are not actual errors, but correct translations not found in the gold-standard. Average results in both language pairs and collocation patterns.

	Distrib. model	**Composit. method**	**Filter**	***Disagreement***
Percentage	36.9	21.5	6.2	*35.4*

helps us to understand why the results of our evaluations are lower than others in the literature, which were obtained by manually classifying the output of the systems instead of using gold-standard data sets. In this regard, the average precision results of the analyzed model would approximately increase from 0.63 to 0.75. On the other hand, the analysis sheds some light on some disadvantages of the proposed approach. Thus, the issues caused by the distributional method (which are frequent in these types of strategies) may be reduced by using contextualized models. Besides, the compositional method could be improved by dynamically adapting the weights of the additive models using unsupervised approaches which predict the compositionality of a given combination [6]. Also, it could be interesting to use different weight ratios for selecting collocate candidates and for computing the similarity between two combinations. With respect to the non-compositional model, it seems a good alternative for some combinations which are not well captured by the weighted additive models. Finally, the errors produced by the filter method were due to the relatively small size of the corpora. Further experiments are necessary to verify whether using large resources actually improves the accuracy of the filter.

5 Conclusions and further work

This paper has presented a compositional distributional semantics strategy to find bilingual equivalents of collocations using only monolingual corpora. The method consists of learning monolingual word embeddings and map them into a bilingual space using an (almost) unsupervised approach. To find the collocation equivalents in the target language, we first obtain a weighted compositional vector of the source collocation, which allows us to select only those collocates which are similar to the whole combination. Both to weigh the components of each collocation and to filter out unconventional combinations, we follow a phraseological approach which states that collocations are syntactically related and lexically restricted combinations of two lexical units.

Different experiments using two collocation patterns in English–Portuguese and English–Spanish have shown that a ratio of about 35%/65% (base/collocate) in the compositional vector achieves the best results in most scenarios. Besides, the evaluations have shown that the proposed method works better not only for non-congruent combinations, but also for other collocations which are also well covered by word-to-word approaches. These results confirm both the compositionality and the asymmetry of collocations, where the meaning of the col-

locate depends on the particular combination in which it occurs. Apart from that, we have also implemented a non-compositional strategy which obtained high precision values, being a promising method to translate compositional and non-compositional MWEs such as idioms.

In further work we plan to combine the weighted strategy with contextualized models of distributional semantics, such as the use of selectional preferences to represent the lexical units [13]. We believe that these approaches obtain more accurate distributional representations of words, so this may be an interesting line for further research, together with the referred non-compositional method. Besides, it would be interesting to separately evaluate the translation of congruent and non-congruent collocations (and eventually of non-compositional expressions) in order to identify the best method for each type, or whether simple approaches could deal with different sorts of MWEs.

Finally, it is worth mentioning that this study also contributes with two new freely available data sets for evaluating the translation of *adj-noun* and *verb-obj* collocations in English–Portuguese and English–Spanish.

References

1. Artetxe, M., Labaka, G., Agirre, E.: A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 789–798 (2018)
2. Baldwin, T., Tanaka, T.: Translation by Machine of Complex Nominals: Getting it Right. In: Second ACL Workshop on Multiword Expressions: Integrating Processing. pp. 24–31. ACL, Barcelona, Spain (2004)
3. Baroni, M.: Composition in distributional semantics. Language and Linguistics Compass **7**(10), 511–522 (2013)
4. Baroni, M., Zamparelli, R.: Frege in space: A program of compositional distributional semantics. LiLT (Linguistic Issues in Language Technology) **9** (2014)
5. Coecke, B., Sadrzadeh, M., Clark, S.: Mathematical foundations for a compositional distributional model of meaning. Linguistic Analysis (36), 345–384 (2010)
6. Cordeiro, S., Villavicencio, A., Idiart, M., Ramisch, C.: Unsupervised compositionality prediction of nominal compounds. American Journal of Computational Linguistics **45**(1), 1–57 (2019)
7. Delpech, E., Daille, B., Morin, E., Lemaire, C.: Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. In: Proceedings of COLING 2012. pp. 745–762 (2012)
8. Fung, P.: A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In: Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas. Machine Translation and the Information Soup. pp. 1–17. AMTA 1998, Springer (1998)
9. Fung, P., McKeown, K.: Finding Terminology Translations from Non-parallel Corpora. In: Fifth Workshop on Very Large Corpora. pp. 192–202. ACL (1997)
10. Fung, P., Yee, L.Y.: An IR approach for translating new words from nonparallel, comparable texts. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 414–420. ACL (1998)

11. Gamallo, P.: Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. Language Resources and Evaluation **51**(3), 727–743 (2017)
12. Gamallo, P.: The role of syntactic dependencies in compositional distributional semantics. Corpus Linguistics and Linguistic Theory **13**(2), 261–289 (2017)
13. Gamallo, P., Garcia, M.: Unsupervised Compositional Translation of Multiword Expressions. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). ACL, Florence (2019)
14. Gamallo, P., Garcia, M., Pineiro, C., Martinez-Castaño, R., Pichel, J.C.: LinguaKit: a Big Data-based multilingual tool for linguistic analysis and information extraction. In: Proceedings of the Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS). pp. 239–244. IEEE (2018)
15. Garcia, M., García-Salido, M., Alonso-Ramos, M.: Using bilingual word-embeddings for multilingual collocation extraction. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 21–30. ACL (2017)
16. Garcia, M., García-Salido, M., Sotelo, S., Mosqueira, E., Alonso-Ramos, M.: *Pay attention when you pay the bills.* A multilingual corpus with dependency-based and semantic annotation of collocations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). ACL, Florence (2019)
17. Grefenstette, G.: The world wide web as a resource for example-based machine translation tasks. In: Proceedings of the ASLIB Conference on Translating and the Computer. vol. 21 (1999)
18. Haruno, M., Ikehara, S., Yamazaki, T.: Learning bilingual collocations by word-level sorting. In: Proceedings of the 16th Conference on Computational Linguistics. COLING 1996, vol. 1, pp. 525–530 (1996)
19. Hausmann, F.J.: Le dictionnaire de collocations. In: Hausmann, F.J., Reichmann, O., Wiegand, H., Zgusta, L. (eds.) Wörterbücher: ein internationales Handbuch zur Lexikographie. Dictionaries. Dictionnaires, pp. 1010–1019. Mouton De Gruyter, Berlin (1989)
20. Koehn, P.: Europarl: a parallel corpus for Statistical Machine Translation. In: Proceedings of the 10th Machine Translation Summit. pp. 79–86. Phuket (2005)
21. Kupiec, J.: An algorithm for finding noun phrase correspondences in bilingual corpora. In: Proceedings of the 31st Annual Meeting on Association for Computational Linguistics (ACL 1993). pp. 17–22. ACL (1993)
22. Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only. In: Proceedings of the Sixth International Conference on Learning Representations (ICLR 2018) (2018)
23. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). pp. 923–929. European Language Resources Association (ELRA) (2016)
24. Mel'čuk, I.: Collocations and lexical functions. In: Cowie, A.P. (ed.) Phraseology. Theory, analysis and applications, pp. 23–53. Clarendon Press, Oxford (1998)
25. Mel'čuk, I.: Phraseology in the language, in the dictionary, and in the computer. Yearbook of Phraseology **3**(1), 31–56 (2012)
26. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013 (2013), arXiv preprint arXiv:1301.3781

27. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. CoRR **abs/1309.4168** (2013)
28. Mitchell, J., Lapata, M.: Vector-based Models of Semantic Composition. In: Proceedings of ACL-08: HLT. pp. 236–244. ACL (2008)
29. Mitchell, J., Lapata, M.: Composition in distributional models of semantics. Cognitive science **34**(8), 1388–1429 (2010)
30. Morin, E., Daille, B.: Revising the compositional method for terminology acquisition from comparable corpora. In: Proceedings of COLING 2012. pp. 1797–1810. Mumbai, India (2012)
31. Nesselhauf, N.: The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. Applied linguistics **24**(2), 223–242 (2003)
32. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999). pp. 519–526. ACL (1999)
33. Reddy, S., Klapaftis, I., McCarthy, D., Manandhar, S.: Dynamic and static prototype vectors for semantic composition. In: Proceedings of 5th International Joint Conference on Natural Language Processing. pp. 705–713. Asian Federation of Natural Language Processing (2011)
34. Rivera, O., Mitkov, R., Pastor, G.: A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In: Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology. pp. 18–25 (2013)
35. Seretan, V., Wehrli, E.: Collocation translation based on sentence alignment and parsing. In: Actes de la 14e conference sur le traitement automatique des langues naturelles. pp. 401–410. TALN 2007, IRIT Press (2007)
36. Smadja, F.: How to compile a bilingual collocational lexicon automatically. In: Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques. pp. 57–63. AAAI Press (1992)
37. Straka, M., Straková, J.: Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 88–99. ACL (Aug 2017)
38. Tanaka, T., Baldwin, T.: 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. pp. 17–24. ACL (2003)
39. Taslimipoor, S., Mitkov, R., Corpas Pastor, G., Fazly, A.: Bilingual contexts from comparable corpora to mine for translations of collocations. In: Computational Linguistics and Intelligent Text Processing. pp. 115–126. Springer, Cham (2018)
40. Upadhyay, S., Faruqui, M., Dyer, C., Roth, D.: Cross-lingual models of word embeddings: An empirical comparison. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1661–1670 (2016)
41. Vargas, N., Ramisch, C., Caseli, H.: Discovering light verb constructions and their translations from parallel corpora without word alignment. In: Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017). pp. 91–96. ACL (2017)
42. Weir, D., Weeds, J., Reffin, J., Kober, T.: Aligning Packed Dependency Trees: A Theory of Composition for Distributional Semantics. American Journal of Computational Linguistics **42**(4), 727–761 (2016)
43. Wu, C.C., Chang, J.S.: Bilingual collocation extraction based on syntactic and statistical analyses. In: Proceedings of the 15th Conference on Computational Linguistics and Speech Processing. pp. 1–20. Association for Computational Linguistics and Chinese Language Processing (2003)