24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

The Impact of Linguistic Knowledge in Different Strategies to Learn Cross-Lingual Distributional Models

Pablo Gamallo[1]

Abstract. In recent years, with the emergence of neural networks and word embeddings, there has been a growing interest in working on cross-lingual distributional models learned from monolingual corpora to induce bilingual lexicons However, interest in these models existed prior to the emergence of deep learning. In this article, we will study the differences between the recent strategies, which are based on the alignment of models, as opposed to the old methods focused on the use of bilingual anchors aligning the text itself. We will also analyze the impact of including different levels of linguistic knowledge (e.g. lemmatization, PoS tagging, syntactic dependencies) in the process of building cross-lingual models for English and Spanish. Our experiments show that syntactic information benefits traditional models based on text alignment but harms mapped cross-lingual embeddings.

1 Introduction

Cross-lingual distributional models learned from monolingual corpora (i.e., without using parallel texts) have been subject of study for more than 20 years, as the first works date from the end of the last century [11, 12, 36]. In order to learn cross-lingual distributional models from monolingual corpora, it is required some kind of alignment or mapping, either of the corpus or of the models themselves. Until the appearance of word embeddings and deep learning techniques [28], the alignment of monolingual corpora has been carried out by placing bilingual anchorage marks within the text before learning the models. Since the work by Mikolov et al. [27] in 2013, the most popular approach is to build the distributional models (word embeddings) separately from the monolingual corpora and then align these models. The alignment or mapping consists of learning a linear projection from one distributional space to the other by using bilingual anchors. Anchors are seed bilingual word pairs or other linguistic marks shared by the two languages that may appear in the same position in the text and, consequently, are surrounded by similar context words. Anchors can be provided directly by external resources as bilingual dictionaries (supervised approach) [18, 2] or can be automatically learned (unsupervised) from monolingual corpora [4]. Between both supervised and unsupervised methods, there are semi-supervised strategies, using as starting point a small set of seed words [3], or distant supervised methods relying on parallel texts to automatically learn bilingual anchors [14].

Although the first studies on cross-lingual models have made experiments with different linguistic features (e.g., syntactic information) [23], the most recent have not yet worked with linguistically analyzed text as they are based on simply models of word tokens or sub-words (common substrings). One of the main objectives of our article is to learn cross-lingual models with more abstract linguistic representations than just tokenization, such as lemmatization, PoS tagging, and dependency syntactic analysis. We will try to find out to what linguistic level it is convenient to work so as to build efficient cross-lingual word models.

The other objective of the article is to compare the performance of the two main strategies to learn cross-lingual distributional models. That is, we will compare the method that make the alignment on monolingual texts, what we call *text-aligned* approach, with the more recent strategy based on aligning the models themselves, which we call *model-aligned*. As our experiments will show, the model-aligned strategy outperforms the text-aligned one to induce bilingual lexicons. In addition, the experiments also show that syntactic information benefits text-aligned methods but harms model-aligned embeddings.

The two main contributions of our work are the following. First, to the best of our knowledge, this is the first study explicitly comparing traditional text-aligned strategies with more recent model-aligned methods. Second, as the latter usually only works with tokens, our study is innovative in adding various levels of linguistic knowledge to enrich cross-lingual embeddings.

The proposed work could be framed within the Hybrid Intelligence approach, which consists of inserting human knowledge into machine/deep learning architectures to try to overcome the limits of existing Artificial Intelligence systems [8]. The use of structured linguistic knowledge may help make the language models more transparent, easy to interpret by humans, and more efficient for specific purposes.

The remaining of the article is organized as follows. In the next section (2), we will introduce related work on both text and model-aligned strategies. Then, in Section 3, we describe the specific methods used in the experiments and how they are configured with several levels of linguistic features. In Section 4, the experiments are described and the results are discussed. Finally, some conclusions and future work are addressed in Section 5.

2 Text and Model-Aligned Strategies

2.1 Text-Aligned

The text-aligned strategy to learn cross-lingual models was mainly designed to extract candidate translations from monolingual corpora [11, 12, 36, 7, 14]. The starting point is a list of bilingual word pairs (called *anchors*) provided by external resources, e.g., existing bilingual dictionaries or probabilistic lexicons learned from parallel corpora. The basic idea is simple: given the word x_i and its translation z_i, the set of bilingual anchors, $\{(x_i, z_i) \in DIC\} : 1 \leq i \leq n$,

[1] CiTIUS, University of Santiago de Compostela, email: pablo.gamallo@usc.es

are the contexts used to define the n dimensions of the bilingual vector space. So, a word vector $\vec{z}_k$ of the target language is similar (and then a candidate translation) to $\vec{x}_j$ in the source language if the two vectors tend to share the same bilingual anchors (x_i, z_i) within the corpus. The strategy consists of learning just one bilingual vector space where the dimensions representing word contexts are the bilingual pairs of DIC. These bilingual pairs are the anchors used to align the monolingual corpora. Let us take an example, if $(president, presidente)$ is an English-Spanish bilingual anchor and it frequently co-occurs with both the English word *Republic* and the Spanish one *república*, it would be one of the most relevant dimensions allowing to approach both words in the vector space.

Concerning the type of context used to build the distributional models, there are some text-aligned works that use dependency-based contexts [39, 14, 20, 22, 23] and compare its efficiency with those standard approaches using just tokens, which are known as bag-of-words techniques. In the studies cited above, syntax-based methods outperform bag-of-words techniques.

In most text-aligned approaches, the vector space tends to be transparent, interpretable and small, as the vector size coincides with the size of the external resource, namely the number of bilingual pairs used as contextual anchors.

However, recent work [25] have used a sort of implicit text alignment of two monolingual corpora to infer a dense vector space shared by the two languages. The method, known as *unsupervised joint training*, is very simple. If two languages are related and share a good fraction of tokens, a single set of cross-lingual word embeddings can be learned from the result of just concatenating the two monolingual corpora. To maximize the token sharing across two languages, Lample et al. [25] makes use of of *byte pair encoding* tokenization, which is a sort of subword tokenization. In [40], the authors propose a framework that combines both strategies: text and model alignment. More precisely, they use unsupervised joint training (text alignment) as initialization and linear mapping on the vector space (model alignment) as refinement.

Unsupervised joint training is behind the construction of multilingual contextualized word embeddings [10], which are in fact a single contextualized representation pre-trained on the concatenation of monolingual corpora in several languages. These multilingual representations are fine-tuned for specific tasks (e.g., parsing, named entity recognition, text classification, machine translation,...) on small amounts of supervised data from one particular language. Then, the task is evaluated in a different language as the multilingual model generalizes information across languages [35].

In our work, we will not use multilingual models based on joint training from corpus concatenation as generalization across languages only works with tokens or subwords. This methodology splits words into the most common sub-words across all languages so as to maximize the shared vocabulary between languages. However, our aim is to introduce more abstract linguistic levels with more complex linguistic units (e.g., syntactic dependencies) which goes in the opposite direction to subword tokenization.

2.2 Model-Aligned

The first model-aligned strategy to learn cross-lingual word embeddings was proposed by Mikolov et al. [27]. They focus on learning a translation mapping between two word embeddings learned from two monolingual corpora. For this purpose, they use an external dictionary DIC of n word pairs such that $\{(x_i, z_i) \in DIC\} : 1 \leq i \leq n$. The pairs in DIC are the bilingual anchors used to align the monolingual models, which is done by learning a linear mapping W between the source and the target vector space. The mapping W is the weight that best approximates the vectors of the word translations in DIC. To approximate the translations, the strategy is based on searching for the minimum value of W such that $\vec{z}_i = \vec{x}_i W$, by iterating on the n dictionary pairs:

$$\arg\min_{W} \sum_{i}^{n} ||W\vec{x}_i - \vec{z}_i|| \quad (1)$$

The algorithm to make the search is based on stochastic gradient descent. Even though this simple linear mapping works very well to find word translations, [41, 2] showed that the results may be improved by adding an orthogonality constraint on W and length-normalized embeddings. In addition, instead of using gradient descent, [2] propose an efficient analytic algorithm to compute the minimum value of W.

As opposed to the supervised method, based on large dictionaries (over 5000 pairs), semi-supervised strategies start with a small set of seed word pairs. [3] used an iterative self-learning method to bootstrap an acceptable mapping function by making use of small seed bilingual anchors (e.g., 25 bilingual word pairs). Other (almost) unsupervised approaches are based on words shared by the two languages (e.g., proper names) and cognates [38], or even just shared numerals as in [3]. However, in order to take advantage of these natural anchors, the languages to be used must share the same spelling and numeral system, which is not always the case between many pairs of languages all over the world.

To avoid the need of external bilingual resources or alphabetic dependencies and thereby making the strategy totally unsupervised, it is possible to design a technique to create bilingual anchors from the similarity matrix of words in the two languages. [4] induce the initial set of bilingual anchors by considering the similarity distributions of the most frequent words in each language. More precisely, it was observed that equivalent translations (for instance word *two* in English and *due* in Italian) have a more similar distribution than words without any semantic relation (e.g. *two* and *cane* - meaning dog). This observation is used to build an initial set of bilingual anchors that is later reinforced with a robust self-learning method by iteratively improving the mapping. In [24], an adversarial training procedure is performed to learn the linear mapping in an unsupervised way.

3 The Use of Linguistic Knowledge in Cross-Lingual Strategies

As it was said in the Introduction, the objective of our work is, on the one hand, to compare the performance of text and model-aligned strategies and, on the other, to verify if the use of different levels of linguistic knowledge improves or impoverishes the basic vector space composed by just word tokens. For this purpose, two state-of-the-art systems has been configured to work with tokens, lemmas, PoS tags, and syntactic dependencies. In this section, we first present the main characteristics of the two systems and then introduce the different linguistic properties that we will be integrated in the systems.

3.1 Text-Aligned *versus* Model-Aligned

The state-of-the-art technique we have chosen to align texts and build bilingual model spaces was designed and implemented by the author of the article, and defined in [16]. One particular configuration of

this technique turned out to the best system using only monolingual corpora at the SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity [6].

The starting point of the system is a seed list of word pairs that are used as bilingual anchors to align the texts and build the cross-lingual vector space. The seed list is provided by an external bilingual dictionary. The bilingual anchors are the word contexts in the vector space, which is a transparent count-based model with explicit and sparse dimensions. To reduce sparseness, we apply a technique to filter out contexts by relevance [17]. The filtering technique consists in selecting, for each word, the R (relevant) contexts with the highest lexical association (e.g. log-likelihood) scores. The top R contexts are considered to be the most *relevant* and informative for each word. R is a global, arbitrarily defined constant whose usual values range from 10 to 1000 [5, 32]. The data structure we have chosen to store the explicit matrix is a hash table with a key-value representation. Keys are structured as a two-dimensional array containing only those row-column pairs with non-zero values (relevant contexts).

This standard text-aligned strategy can be considered supervised as it relies on an external bilingual dictionary. It is also a count-based method, as vectors are learned by counting the co-occurrence of all word-context pairs. A particular implementation of this strategy aimed at building cross-lingual dependency-based models is freely available.[2]

Concerning the model-aligned strategy, we have chosen VecMap to align monolingual models into a shared vector space.[3] This tools yields state-of-the-art results using different methods:

Supervised: A technique relying on an external bilingual dictionary [2].

Semi-supervised: It just requires a small seed list (e.g., 25) of bilingual pairs [3].

Identical: It takes advantage of words shared by the languages in question, for instance: numbers and named entities.

Unsupervised: It does not require an external bilingual dictionary as it is automatically built [4].

Figure 1 shows how the modules and resources are organized in the two strategies. On the one hand, the text-aligned method does not require monolingual models as the alignment is made on the text corpora. Then, a common vector space represented by a single cross-lingual model is built from the aligned corpora. On the other hand, the model-aligned strategy directly aligns the two monolingual models so as to put them in the same cross-lingual vector space. In both cases, the alignment is made by means of a bilingual dictionary providing bilingual anchors.

3.2 Linguistic Knowledge

Although there have been many works on the text-aligned approach that used distributional models with all kinds of linguistic information, current model-aligned embeddings usually contain only token-level information.

Table 1 shows the linguistic analysis of a sentence. The linguistic knowledge provided by the analysis includes tokenization (first column), lemmatization (second column), morphological analysis (third column), PoS tagging (third and forth columns), and dependency analysis (fifth and sixth columns). On the basis of that, we make use of four levels of linguistic knowledge:

[2] https://github.com/gamallo/CrossLingual
[3] https://github.com/artetxem/vecmap

token	lemma	PoS+morph	PoS	head	relation
Unions	union	NNS	N	4	nsubj
across	across	IN	IN	3	case
America	america	NNP	N	1	nmod
prepared	prepare	VBD	V	0	root
a	a	DT	DT	7	det
general	general	JJ	JJ	7	amod
strike	strike	NN	N	4	dobj

Table 1. Linguistic representation on the sentence "Unions across America prepared a general strike". Concerning the PoS tags, NNS means plural common noun, IN means preposition, NNP means proper noun, VBD means verb in past tense, JJ means adjective, and NN means common noun. On the other hand, nsubj, nmod, amod, and dobj are Universal Dependency labels meaning nominal subject, noun modifier, adjective modifier, and direct object, respectively.

Tokens : Only the first column of the analysis is selected and then the token text is normalized to lower case.

Lemmas : Only the second column of the analysis is selected as such (without further normalization).

Lemma-PoS : The second and forth columns (lemmas and simplified PoS tags) are selected to create $(lemma, PoS)$ pairs, e.g., $(union, N)$ or $(prepare, V)$.

Dependencies : The second, fifth and sixth columns are selected so as to assign syntactic contexts to lemmas.

To extract contexts from syntactic dependencies, we use the co-compositional methodology defined in [17]. Contexts can be derived from the dependency relations the lemmas participate in (e.g. nominal subject, direct object, nominal modifier, etc). For a target lemma l related to a set of dependents $d_1, ..., d_k$ and to a head h (since each lemma is only dependent of only one head), we extract the contexts:

$$(d_1, \downarrow rel_1)..., (d_k, \downarrow rel_k), (h, \uparrow rel_h)$$

where $\downarrow rel$ is a type of dependency relation containing a specific dependent lemma, and $\uparrow rel$ stands for the inverse relation: a dependency containing a specific head. For instance, taking into account the analysis in Table 1 $(union, \downarrow subject)$ is a lexico-syntactic context of the verb *prepare*, while $(prepare, \uparrow subject)$ is a lexico-syntactic context of *union*.

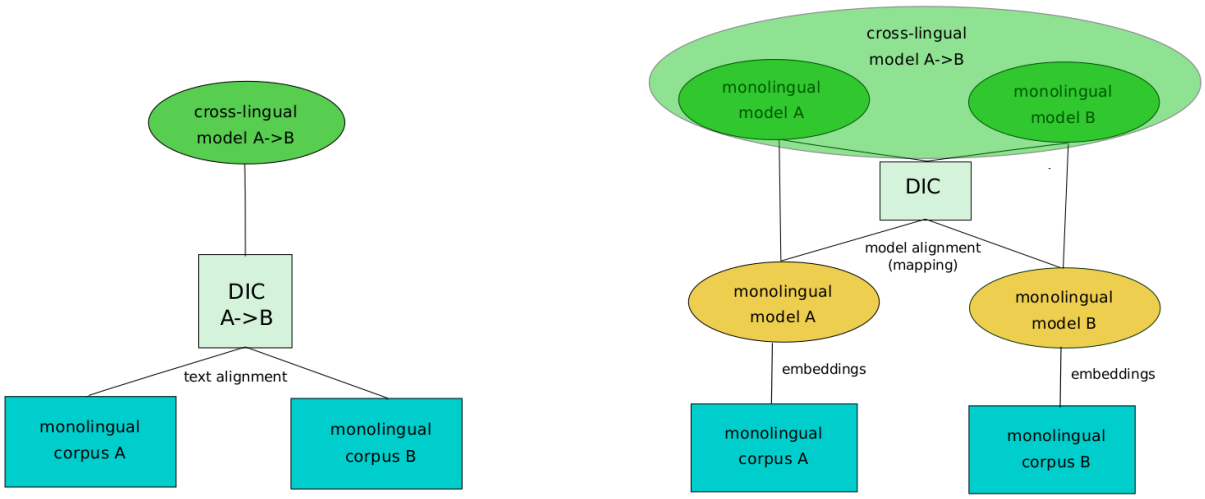
We must emphasize that the linguistic knowledge we use is incremental: syntactic analysis is based on PoS tagging, which in turn contains lemmatization, which takes tokenization and a lexicon as input.

4 Experiments

After describing the different types of linguistic information and the characteristics of the two systems (and corresponding strategies), we present below the experiments carried out to compare their performance.

4.1 Resources and Text Corpora

The experiments were focused on English $\rightarrow$ Spanish translations. All monolingual models were generated from both the English and Spanish Wikipedia (January 2018 dump), containing 2.1B and 518M word tokens, respectively. To process the corpora and create the four levels of linguistic knowledge, we used the lemmatizer and PoS tagger of LinguaKit [13] and the syntactic analyzer DepPattern, a rule-based and multilingual dependency parser [19] also taking part of



(a) Text-aligned architecture

(b) Model-aligned architecture

Figure 1. Architecture of two strategies (text-aligned and model-aligned) to build cross-lingual models from two monolingual corpora.

LinguaKit.[4] Both tools were mostly developed by the author of the article.

To generate the cross-lingual models in a supervised way, we made use of 5,000 bilingual $(lemma, PoS)$ pairs randomly extracted from the English-Spanish dictionary integrated in Apertium, an open source machine translation system, where the entries are lemmas with their corresponding PoS tags.[5] From this initial list, we also created another one with 4,934 bilingual pairs of lemmas only (once the PoS tags are removed). We need to work with lemmas because they are the only word forms that are shared by the four linguistic levels we are evaluating. All lemmas are included in the set of tokens as well as in the set of all $(lemma, PoS)$ pairs. Besides, dependencies were defined between lemmas. By contrast, Not all tokens are included in the set of all lemmas. For this reason, it is not possible for us to use the usual benchmarks with token-based dictionaries, which are common in the evaluations of cross-lingual word embeddings.

To test the models, a different list of 1,500 bilingual $(lemma, PoS)$ pairs was selected from the same dictionary. On the basis of it, another list containing 1,474 pairs with just lemmas was also created. The train and test bilingual datasets are freely available.[6]

4.2 Configuration

To build the distributional models, English lexical units (tokens and lemmas) appearing less than 200 times were filtered out after we have analyzed the text. As the Spanish Wikipedia is four times smaller, we removed those units with frequency less than 100. In the case of dependency-based models, lexico-syntactic contexts with frequency less than 50 were removed.

Regarding the cross-lingual models built using the text-aligned strategy, we selected the 300 most relevant contexts for each word. Note that in this strategy, there is no need to build independent monolingual models. Instead, cross-lingual models are elaborated directly from texts aligned with bilingual anchors, being these anchors the contexts that form the dimensions of the cross-lingual vector space. The anchors correspond to the 5,000 bilingual pairs of the train dictionary. According to the type of linguistic information targeted (i.e., tokens, lemmas, $(lemma, PoS)$ pairs, or dependencies), four different cross-lingual models were created. For those models containing tokens, lemmas, or $(lemma, PoS)$ pairs, we used a window of 5 units. In the case of syntactic-based models, we use the lexico-syntactic contexts (i.e., lemma in a syntactic position) defined in Section 3.

Concerning the model-aligned strategy, the first step is to build independent monolingual embeddings. The monolingual embeddings of tokens, lemmas and $(lemma, PoS)$ pairs were created with Word2Vec, configured with CBOW algorithm, window of 5 tokens, negative-sampling parameter (how many negative contexts to sample for every correct one) of 15, and 300 dimensions [29]. For the embeddings of syntactic dependencies, we used Word2Vecf [26], a tool that allows to generate embeddings from pairs of lemmas and lexico-syntactic contexts. The algorithm behind Word2Vecf is a generalization of the Skip-Gram model that moves from linear bag-of-words to arbitrary word contexts, as in the case of lexico-syntactic contexts. The tool has been configured with negative-sampling parameter of 15 and 300 dimensions. In the second step, the monolingual embeddings were aligned using different learning methods of VecMap: *supervised*, *semi-supervised*, *identical*, and *unsupervised* learning. Supervised and semi-supervised methods were set up to train and map the models with the list of 5,000 train translations. By contrast, the identical and unsupervised methods do not require the training dic-

[4] https://github.com/citiususc/Linguakit
https://github.com/citiususc/Deppattern
[5] https://sourceforge.net/projects/apertium/
[6] https://github.com/gamallo/CrossLingual/tree/master/dicos

tionary. The former is based on mapping the monolingual models by using lemmas shared in the two languages, while the later automatically builds translation equivalents with the most frequent words in each language.

4.3 Results

We used the translation task and a test dictionary with 1,500 bilingual pairs to evaluate the quality of all the distributional models generated with the two strategies and different linguistic knowledge. The translation task consists of learning a cross-lingual model from a train bilingual dictionary and two monolingual corpora, and measures its accuracy on predicting the translation of new words in a test dictionary. In the case of unsupervised (and identical) methods, the training dictionary is not required.

Table 2 shows the results obtained with four cross-lingual models (one per linguistic level) learned by using the text-aligned strategy. The coverage is the percentage of words actually evaluated with regard to the total number of pairs in the test dictionary. The last three columns show the precision values P@k, where k is the number of similar words returned for each test word. More precisely, P@1, P@5 and P@10 correspond to the number of correct translations among the top 1, top 5 and top 10 of similar words returned. The final scores show that the cross-lingual model based on syntactic dependencies clearly outperforms the other models.

We have also been carrying out experiments using multilingual models based on joint training from corpus concatenation. In particular, we concatenated the English and Spanish wikipedias to learn multilingual word embeddings (with Word2Vec) in the same vector space. This strategy merely achieved 10.01 P@1. We also induced bilingual lexicons with the pre-trained multilingual BERT model [10] by encoding tokens instead of sentences. As expected, we did not get any positive results (P@1 = 0.00). The cross-lingual ability of Transformers such as BERT, based on sentence embeddings, is not suitable for translating words out of context, but to allow for zero-shot cross-lingual model transfer in specific NLP tasks [35].

Table 3 shows the results obtained with four cross-lingual models learned using the model-aligned strategy. The syntactic strategy builds the monolingual embeddings with Word2Vecf by taking as input the same syntactic dependencies as those used in the text-aligned experiments. In this table, the columns show, for each linguistic configuration, the coverage and the P@1 precision with four different methods: supervised (sup), semi-supervised (semi), identical (id), and unsupervised (uns). The best results were obtained with $(lemma, PoS)$ pairs. To simplify the display of the results, the values of P@5 and P@10 are missing. The best method, $(lemma, PoS)$, reaches 77.43 P@5 and 79.84 P@10.

Methods	coverage	P@1	P@5	P@10
tokens	90.81	23.73	36.13	41.30
lemmas	90.90	27.23	44.67	50.45
(lemma,PoS)	90.90	27.85	46.01	52.31
syntax	94.65	**62.33**	75.91	78.39

Table 2. Results for the English-Spanish translation task obtained with the text-aligned strategy across four levels of linguistic knowledge. The scores represent word translation accuracy evaluated on the top-1, top-5, and top-10 words as ranked by the method.

Methods	cov	sup(P@1)	semi(P@1)	id(P@1)	uns(P@1)
tokens	92.91	66.88	63.91	63.82	64.18
lemmas	92.91	70.74	72.08	72.17	72.08
(lem,PoS)	91.34	72.24	**72.90**	72.49	72.65
syntax	95.00	48.90	46.71	47.32	40.12

Table 3. Results for the English-Spanish translation task obtained with the model-aligned strategy across four levels of linguistic knowledge, and making use of four learning methods: supervised (sup), semi-supervised (semi), identical (id), and unsupervised (uns) learning.

4.4 Discussion

From the results obtained, we observe the following patterns and trends:

- Most configurations of the model-aligned strategy improves the results of the best configuration of the text-aligned strategy.
- In the text-aligned strategy, the more structured the linguistic knowledge, the better the results. In the model-based strategy, on the other hand, the same happens until the syntactic level is reached, since precision is strongly reduced at this level.
- In the model-based strategy, there are no significant differences between supervised and unsupervised methods, including intermediate ones.
- Syntactic information seems to be extremely useful in the text-aligned strategy, but, on the contrary, it considerably damages the model-aligned strategy.

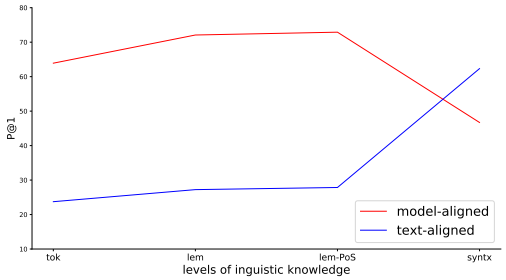


Figure 2. Comparing the best text-aligned configuration (syntax) with the best model-aligned one $(lemma, PoS)$ across the four linguistic levels.

Figure 2 allows us to observe how the incremental integration of linguistic knowledge improves the text-aligned strategy to reach the peak with the use of syntactic dependencies. Concerning the model-aligned method, it improves smoothly until it reaches the $(lemma, PoS)$ level, but drops drastically with the syntactic dependencies.

The fact that text-aligned models work better with syntactic dependencies than with lower linguistic structure seems to be in accordance with previous studies using count-based vector spaces [37, 31, 34, 15]. All of them state that syntax-based methods outperform bag-of-words techniques with tokens or lemmas, in particular when the objective is to compute semantic similarity between functional equivalent words so as to detect co-hyponym/hypernym word relations (i.e. near synonymy). This objective is actually very similar to that of finding word translations in a bilingual vector space.

The poor results obtained with syntactic information in the model-aligned strategy have no obvious explanation. It could be because the use of Word2Vecf produces low quality monolingual models. However, when we compare the dependency-based model with that built with lemmas in other tasks, there are no significant differences. For instance, in the task consisting of computing the correlation between the WordSim353 dataset [1] and the word similarity extracted from the models, we obtain very close Spearman correlations: 0.732 (dependency-based model) *vs* 0.738 (lemma-based model). With other experiments, such as TOEFL synonym test questions, we also find no significant differences between the two models. This means that the quality of the dependency-based models generated with Word2Vecf seems to be satisfactory, so the problem might be at the alignment/mapping stage.

We also made a more qualitative analysis of the results of the two best configurations for each strategy: dependency-based for the text-aligned strategy and $(lemma, PoS)$ for the model-aligned. This analysis allows us to observe the following:

- The model-aligned method tends to find correct translations with less frequent words. Taking into account the ranked list (from 1 to N, where N is the vocabulary size) of all English words, the average ranking of the words for which a correct translation was found is 6,224. However, in the case of the text-aligned strategy, this average ranking is much lower: 3,777.
- 14.95% of the correct translations detected with the best text-aligned strategy are not among the correct translations of the best model-aligned method. It means that there is an important number of translations with frequent words that are detected by the text-aligned method but not by the model-aligned method. A sample of these translations is depicted in Table 4. There is, therefore, considerable room for improvement if some new model-based strategy is found that takes advantage of syntactic information.

English	Spanish
patent	patentar
penalty	pena
perpetual	perpetuo
place	colocar
plane	plano
practice	practicar
press	pulsar
prison	prisión
prominent	prominente
promise	promesa
promising	prometedor
promotion	ascenso
provide	proporcionar
provoke	provocar
psychiatric	psiquiátrico
pump	bombear
push	pulsar

Table 4. Sample of correct translations found by the text-aligned method which are not among the correct translations of the best model-aligned strategy.

Since the text-aligned strategy is based on transparent vector spaces, it allows us to observe and analyze problems and errors more easily. Besides, transparency also allows us to inquire into why a particular case was correctly translated. For instance, in the case of the translation pair *psychiatric* $\rightarrow$ *psiquiátrico*, which was correctly translated by the text-aligned method (but not by the model-aligned), some of the 70 shared lexico-syntactic contexts are depicted in Table 5. All contexts are typical of adjectives, specifically, they all are nouns modified by the adjective ($amod$ dependency). This means that, in English, the system only pays attention to the nouns that appear to the right of the adjective, while in Spanish, the attention is focused on the postponed noun. Syntax helps pay attention on relevant contextual elements, which should have a positive effect on detecting distributional similarity. If this fails, it is probably due to errors propagated from previous linguistic analysis (lemmatization, PoS tagging, and syntactic parsing), or because direct dependencies are not able to cover indirect relationships between words that are semantically related. For instance, given the sentence *research on psychiatric issues*, there is no direct syntactic dependency detecting the semantic relation between *psychiatric* and *research*. By contrast, the approaches relying on window-based contexts (including Skip-Gram and CBOW algorithms) account for these indirect relations without problem even if they also insert noise in the models. What is needed to detect semantic relationships in an intelligent way is to combine dependencies and compositionality in an iterative way. The most recent neural models, specifically contextualized embeddings [9], simulate this compositional strategy, but we still do not know if it is based on how humans understand language, or just on the brute force of distributed iterations.

psychiatric	psiquiátrico
$(jail, \uparrow amod)$	$(prisión, \uparrow amod)$
$(internment, \uparrow amod)$	$(internamiento, \uparrow amod)$
$(institution, \uparrow amod)$	$(institución, \uparrow amod)$
$(institute, \uparrow amod)$	$(instituto, \uparrow amod)$
$(inquiry, \uparrow amod)$	$(consulta, \uparrow amod)$
$(hospital, \uparrow amod)$	$(hospital, \uparrow amod)$
$(hospitalization, \uparrow amod)$	$(hospitalización, \uparrow amod)$
$(help, \uparrow amod)$	$(ayuda, \uparrow amod)$
$(file, \uparrow amod)$	$(expediente, \uparrow amod)$
$(feature, \uparrow amod)$	$(característica, \uparrow amod)$
$(examination, \uparrow amod)$	$(examen, \uparrow amod)$
$(doctor, \uparrow amod)$	$(médico, \uparrow amod)$
$(disorder, \uparrow amod)$	$(desorden, \uparrow amod)$
$(care, \uparrow amod)$	$(cuidado, \uparrow amod)$
$(assistance, \uparrow amod)$	$(asistencia, \uparrow amod)$

Table 5. Sample of bilingual lexico-syntactic contexts shared by the pair *psychiatric* $\rightarrow$ *psiquiátrico* in the text-aligned method.

5 Conclusions

In this work, we have configured two very different methods with different types of linguistic knowledge so as to construct cross-lingual models from English and Spanish monolingual corpora. We have found that, in general, the gradual increase in linguistic information helps to improve models, except in one very specific case: the use of syntactic dependencies in recent model-aligned methods. In addition, we have also found that aligning monolingual embeddings is a more effective technique than aligning monolingual texts using bilingual anchors.

Model-aligned methods can work well because it is assumed that embedding spaces in different languages have a similar structure, i.e. they are approximately isometric. Otherwise, it would not be possible to find a linear map from one space to another [30]. However, according to a recent study [33], the isometric assumption weakens,

and then the quality of the mapping decreases, as the languages are distant. On the basis of this observation, new lines of research are opened which will need to be explored. For instance, it will probably be necessary to adapt the alignment strategy according to the structural distance that separates the languages in question. And so, it will also be necessary to apply a certain level of linguistic knowledge to the method in function of that language distance. The more distance there is, the more abstract will be the linguistic structure we need. As the concept of language distance becomes essential, it will be necessary to compute the linguistic distance between any two languages [21]. In order to be able to deal with these new issues, models and strategies will have to be fine-tuned and applied to different pairs of languages without neglecting linguistic knowledge.

ACKNOWLEDGEMENTS

This work has received financial support from DOMINO project (PGC2018-102041-B-I00 , MCIU/AEI/FEDER, UE), eRisk project (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditations ED431G/08 2016-2019 and ED431B 2017/39) and the European Regional Development Fund (ERDF).

REFERENCES

[1] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa, 'A study on similarity and relatedness using distributional and wordnet-based approaches', in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pp. 19–27, Stroudsburg, PA, USA, (2009). Association for Computational Linguistics.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, '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*, pp. 2289–2294, (2016).

[3] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, 'Learning bilingual word embeddings with (almost) no bilingual data', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 451–462, (2017).

[4] Mikel Artetxe, Gorka Labaka, and Eneko Agirre, '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)*, (2018).

[5] Chris Biemann and Martin Riedl, 'Text: Now in 2d! a framework for lexical expansion with contextual similarity', *Journal of Language Modelling*, **1**(1), 55–95, (2013).

[6] José Camacho-Collados, Mohammad Pilehvar, Nigel Collier, and Roberto Navigli, 'Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity', in *Proceedings of SemEval*, Vancouver, Canada, (2017).

[7] Y-C. Chiao and P. Zweigenbaum, 'Looking for candidate translational equivalents in specialized, comparable corpora', in *19th COLING'02*, (2002).

[8] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister, 'Hybrid intelligence', *Business & Information Systems Engineering*, **61**, 637–643, (2019).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: pre-training of deep bidirectional transformers for language understanding', *CoRR*, **abs/1810.04805**, (2018).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, (June 2019). Association for Computational Linguistics.

[11] Pascale Fung and Kathleen McKeown, 'Finding terminology translation from non-parallel corpora', in *5th Annual Workshop on Very Large Corpora*, pp. 192–202, Hong Kong, (1997).

[12] Pascale Fung and Lo Yuen Yee, 'An IR Approach for Translating New Words from Nonparallel, Comparable Texts', in *Coling'98*, pp. 414–420, Montreal, Canada, (1998).

[13] P. Gamallo, M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel, '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)*, pp. 239–244, (2018).

[14] Pablo Gamallo, 'Learning Bilingual Lexicons from Comparable English and Spanish Corpora', in *Machine Translation SUMMIT XI*, Copenhagen, Denmark, (2007).

[15] Pablo Gamallo, 'Comparing different properties involved in word similarity extraction', in *14th Portuguese Conference on Artificial Intelligence (EPIA'09), LNCS, Vol. 5816*, pp. 634–645, Aveiro, Portugal, (2009). Springer-Verlag.

[16] Pablo Gamallo, 'Citius at semeval-2017 task 2: Cross-lingual similarity from comparable corpora and dependency-based contexts', in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 226–229. Association for Computational Linguistics, (2017).

[17] Pablo Gamallo, 'Comparing explicit and predictive distributional semantic models endowed with syntactic contexts', *Language Resources and Evaluation*, **51**(3), 727–743, (2017).

[18] Pablo Gamallo, 'Strategies for building high quality bilingual lexicons from comparable corpora', in *Parallel corpora for contrastive and translation studies: New resources and applications. Studies in Corpus Linguistics*, eds., Irene Doval and María Teresa Sánchez-Nieto, 251–266, John Benjamins, (2019).

[19] Pablo Gamallo and Marcos Garcia, 'Dependency parsing with finite state transducers and compression rules', *Information Processing & Management*, **54**(6), 1244–1261, (2018).

[20] Pablo Gamallo and José Ramom Pichel, 'Learning Spanish-Galician Translation Equivalents Using a Comparable Corpus and a Bilingual Dictionary', in *International Conference on Intelligent Text Processing and Computational Linguistics*, volume 4919 of *Lecture Notes in Computer Science*, 423–433, Springer, (2008).

[21] Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria, 'From Language Identification to Language Distance', *Physica A*, **484**, 162–172, (2017).

[22] Nikesh Garera, Chris Callison-Burch, and David Yarowsky, 'Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences', in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 129–137, Boulder, Colorado, (June 2009). Association for Computational Linguistics.

[23] Amir Hazem and Emmanuel Morin, 'Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models', *Lecture Notes in Computer Science*, **8404**, 310–323, (2014).

[24] Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou, 'Word translation without parallel data', in *International Conference on Learning Representations*, (2018).

[25] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc Aurelio Ranzato. Phrase-Based & neural unsupervised machine translation, April 2018.

[26] Omer Levy and Yoav Goldberg, 'Dependency-based word embeddings', in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA*, pp. 302–308, (2014).

[27] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever, 'Exploiting similarities among languages for machine translation', *CoRR*, **abs/1309.4168**, (2013).

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean, 'Distributed representations of words and phrases and their compositionality', in *Advances in Neural Information Processing Systems*, pp. 3111–3119, (2013).

[29] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig, '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*, pp. 746–751, Atlanta, Georgia, (2013).

[30] Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre, 'Analyzing the limitations of cross-lingual word embedding mappings', in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4990–4995, (2019).

[31] Sebastian Padó and Mirella Lapata, 'Dependency-Based Construction of Semantic Space Models', *Computational Linguistics*, **33**(2), 161–199, (2007).

[32] Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch, 'Nothing like good old frequency: Studying context filters for distributional thesauri', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 419–424, (2014).

[33] Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig, 'Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 184–193, Florence, Italy, (July 2019). Association for Computational Linguistics.

[34] Yves Peirsman, Kris Heylen, and Dirk Speelman, 'Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts', in *CoSMO Workshop*, pp. 9–16, Roskilde, Denmark, (2007).

[35] Telmo Pires, Eva Schlinger, and Dan Garrette, 'How multilingual is multilingual BERT?', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, (July 2019). Association for Computational Linguistics.

[36] Reinhard Rapp, 'Automatic Identification of Word Translations from Unrelated English and German Corpora', in *ACL'99*, pp. 519–526, (1999).

[37] Violeta Seretan and Eric Wehrli, 'Accurate Collocation Extraction Using a Multilingual Parser', in *21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 953–960, (2006).

[38] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla, 'Offline bilingual word vectors, orthogonal transformations and the inverted softmax', in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, (2017).

[39] T. Tanala, 'Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora', in *19th COLING'02*, pp. 981–987, (2002).

[40] Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework, 2019.

[41] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin, 'Normalized word embedding and orthogonal transform for bilingual word translation', in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1006–1011, Denver, Colorado, (May–June 2015). Association for Computational Linguistics.