Improving Cross-Lingual Word Embeddings by Meeting in the Middle

Yerai Doval
Grupo COLE, Escola Superior
de Enxeñaría Informática
Universidade de Vigo, Spain
`yerai.doval@uvigo.es`

**Jose Camacho-Collados, Luis Espinosa-Anke
and Steven Schockaert**
School of Computer Science and Informatics
Cardiff University, UK
`camachocolladosj@cardiff.ac.uk`
`espinosa-ankel@cardiff.ac.uk`
`schockaerts1@cardiff.ac.uk`

Abstract

Cross-lingual word embeddings are becoming increasingly important in multilingual NLP. Recently, it has been shown that these embeddings can be effectively learned by aligning two disjoint monolingual vector spaces through linear transformations, using no more than a small bilingual dictionary as supervision. In this work, we propose to apply an additional transformation after the initial alignment step, which moves cross-lingual synonyms towards a middle point between them. By applying this transformation our aim is to obtain a better cross-lingual integration of the vector spaces. In addition, and perhaps surprisingly, the monolingual spaces also improve by this transformation. This is in contrast to the original alignment, which is typically learned such that the structure of the monolingual spaces is preserved. Our experiments confirm that the resulting cross-lingual embeddings outperform state-of-the-art models in both monolingual and cross-lingual evaluation tasks.

1 Introduction

Word embeddings are one of the most widely used resources in NLP, as they have proven to be of enormous importance for modeling linguistic phenomena in both supervised and unsupervised settings. In particular, the representation of words in cross-lingual vector spaces (henceforth, *cross-lingual word embeddings*) is quickly gaining in popularity. One of the main reasons is that they play a crucial role in transferring knowledge from one language to another, specifically in downstream tasks such as information retrieval (Vulić and Moens, 2015b), entity linking (Tsai and Roth, 2016) and text classification (Mogadala and Rettinger, 2016), while at the same time providing improvements in multilingual NLP problems such as machine translation (Zou et al., 2013).

There exist different approaches for obtaining these cross-lingual embeddings. One of the most successful methodological directions, which constitutes the main focus of this paper, attempts to learn bilingual embeddings via a two-step process: first, word embeddings are trained on monolingual corpora and then the resulting monolingual spaces are aligned by taking advantage of bilingual dictionaries (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Xing et al., 2015).

These *alignments* are generally modeled as linear transformations, which are constrained such that the structure of the initial monolingual spaces is left unchanged. This can be achieved by imposing an orthogonality constraint on the linear transformation (Xing et al., 2015; Artetxe et al., 2016). Our hypothesis in this paper is that such approaches can be further improved, as they rely on the assumption that the internal structure of the two monolingual spaces is identical. In reality, however, this structure is influenced by language-specific phenomena, e.g., the fact that Spanish distinguishes between masculine and feminine nouns (Davis, 2015) as well as the specific biases of the different corpora from which the monolingual spaces were learned. Because of this, monolingual embedding spaces are not isomorphic (Søgaard et al., 2018; Kementchedjhieva et al., 2018). On the other hand, simply dropping the orthogonality constraints leads to overfitting, and is thus not effective in practice.

The solution we propose is to start with existing state-of-the-art alignment models (Artetxe et al., 2017; Conneau et al., 2018), and to apply a further transformation to the resulting initial alignment. For each word w with translation w', this additional transformation aims to map the vector representations of both w and w' onto their average, thereby creating a cross-lingual vector space which intuitively corresponds to the average of the

two aligned monolingual vector spaces. Similar to the initial alignment, this mapping is learned from a small bilingual lexicon.

Our experimental results show that the proposed additional transformation does not only benefit cross-lingual evaluation tasks, but, perhaps surprisingly, also monolingual ones. In particular, we perform an extensive set of experiments on standard benchmarks for bilingual dictionary induction and monolingual and cross-lingual word similarity, as well as on an extrinsic task: cross-lingual hypernym discovery.

Code and pre-trained embeddings to reproduce our experiments and to apply our model to any given cross-lingual embeddings are available at `https://github.com/yeraidm/meemi`.

2 Related Work

Bilingual word embeddings have been extensively studied in the literature in recent years. Their nature varies with respect to the supervision signals used for training (Upadhyay et al., 2016; Ruder et al., 2018). Some common signals to learn bilingual embeddings come from parallel (Hermann and Blunsom, 2014; Luong et al., 2015; Levy et al., 2017) or comparable corpora (Vulić and Moens, 2015a; Søgaard et al., 2015; Vulić and Moens, 2016), or lexical resources such as WordNet, ConceptNet or BabelNet (Speer et al., 2017; Mrksic et al., 2017; Goikoetxea et al., 2018). However, these sources of supervision may be scarce, limited to certain domains or may not be directly available for certain language pairs.

Another branch of research exploits pre-trained monolingual embeddings with weak signals such as bilingual lexicons for learning bilingual embeddings (Mikolov et al., 2013b; Faruqui and Dyer, 2014; Ammar et al., 2016; Artetxe et al., 2016). Mikolov et al. (2013b) was one of the first attempts into this line of research, applying a linear transformation in order to map the embeddings from one monolingual space into another. They also noted that more sophisticated approaches, such as using multilayer perceptrons, do not improve with respect to their linear counterparts. Xing et al. (2015) built upon this work by normalizing word embeddings during training and adding an orthogonality constraint. In a complementary direction, Faruqui and Dyer (2014) put forward a technique based on canonical correlation analysis to obtain linear mappings for both monolin-

gual embedding spaces into a new shared space. Artetxe et al. (2016) proposed a similar linear mapping to Mikolov et al. (2013b), generalizing it and providing theoretical justifications which also served to reinterpret the methods of Faruqui and Dyer (2014) and Xing et al. (2015). Smith et al. (2017) further showed how orthogonality was required to improve the consistency of bilingual mappings, making them more robust to noise. Finally, a more complete generalization providing further insights on the linear transformations used in all these models can be found in Artetxe et al. (2018a).

These approaches generally require large bilingual lexicons to effectively learn multilingual embeddings (Artetxe et al., 2017). Recently, however, alternatives which only need very small dictionaries, or even none at all, have been proposed to learn high-quality embeddings via linear mappings (Artetxe et al., 2017; Conneau et al., 2018). More details on the specifics of these two approaches can be found in Section 3.1. These models have in turn paved the way for the development of machine translation systems which do not require any parallel corpora (Artetxe et al., 2018b; Lample et al., 2018). Moreover, the fact that such approaches only need monolingual embeddings, instead of parallel or comparable corpora, makes them easily adaptable to different domains (e.g., social media or web corpora).

In this paper we build upon these state-of-the-art approaches by applying an additional transformation, which aims to map each word and its translation onto the average of their vector representations. This strategy bears some resemblance with the idea of learning meta-embeddings (Yin and Schütze, 2016). Meta-embeddings are vector space representations which aggregate several pre-trained word embeddings from a given language (e.g., trained using different corpora and/or different word embedding models). Empirically it was found that such meta-embeddings can often outperform the individual word embeddings from which they were obtained. In particular, it was recently argued that word vector averaging can be a highly effective approach for learning such meta-embeddings (Coates and Bollegala, 2018). The main difference between such approaches and our work is that because we rely on a small dictionary, we cannot simply average word vectors, since for most words we do not know the corresponding

translation. Instead, we train a regression model to predict this average word vector from the vector representation of the given word only, i.e., without using the vector representation of its translation.

3 Methodology

Our approach for improving cross-lingual embeddings consists of three main steps, where the first two steps are the same as in existing methods. In particular, given two monolingual corpora, a word vector space is first learned independently for each language. This can be achieved with common word embedding models, e.g., Word2vec (Mikolov et al., 2013a), GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017). Second, a linear alignment strategy is used to map the monolingual embeddings to a common bilingual vector space (Section 3.1). Third, a final transformation is applied on the aligned embeddings so the word vectors from both languages are refined and further integrated with each other (Section 3.2). This third step is the main contribution of our paper.

3.1 Aligning monolingual spaces

Once the monolingual word embeddings have been obtained, a linear transformation is applied in order to integrate them into the same vector space. This linear transformation is generally carried out using a supervision signal, typically in the form of a bilingual dictionary. In the following we explain two state-of-the-art models performing this linear transformation.

VecMap (Artetxe et al., 2017). VecMap uses an orthogonal transformation over normalized word embeddings. An iterative two-step procedure is also implemented in order to avoid the need of starting with a large seed dictionary (e.g., in the original paper it was tested with a very small bilingual dictionary of just 25 pairs). In this procedure, first, the linear mapping is estimated using a small bilingual dictionary, and then, this dictionary is augmented by applying the learned transformation to new words from the source language. Lastly, the process is repeated until some convergence criterion is met.

MUSE (Conneau et al., 2018). In this case, the transformation matrix is learned through an iter-

ative Procrustes alignment (Schönemann, 1966).[1] The anchor points needed for this alignment can be obtained either through a supplied bilingual dictionary or through an unsupervised model. This unsupervised model is trained using adversarial learning to obtain an initial alignment of the two monolingual spaces, which is then refined by the Procrustes alignment using the most frequent words as anchor points. A new distance metric for the embedding space, referred to as cross-domain similarity local scaling, is also introduced. This metric, which takes into account the nearest neighbors of both source and target words, was shown to better handle high-density regions of the space, thus alleviating the hubness problem of word embedding models (Radovanović et al., 2010; Dinu et al., 2015), which arises when a few points (known as hubs) become the nearest neighbors of many other points in the embedding space.

3.2 Meeting in the middle

After the initial alignment of the monolingual word embeddings, our proposed method leverages an additional linear model to refine the resulting bilingual word embeddings. This is because the methods presented in the previous section apply constraints to ensure that the structure of the monolingual embeddings is largely preserved. As already mentioned in the introduction, conceptually this may not be optimal, as embeddings for different languages and trained from different corpora can be expected to be structured somewhat differently. Empirically, as we will see in the evaluation, after applying methods such as VecMap and MUSE there still tend to be significant gaps between the vector representations of words and their translations. Our method directly attempts to reduce these gaps by moving each word vector towards the middle point between its current representation and the representation of its translation. In this way, by bringing the two monolingual fragments of the space closer to each other, we can expect to see an improved performance on cross-lingual evaluation tasks such as bilingual dictionary induction. Importantly, the internal structure of the two monolingual fragments themselves is also affected by this step. By aver-

---

[1]Very recently, Kementchedjhieva et al. (2018) showed that projecting both monolingual embedding spaces onto a third space (instead of directly onto each other) using a generalized Procrustes analysis facilitates the learning of alignments.

aging between the representations obtained from different languages, we hypothesize that the impact of language-specific phenomena and corpus specific biases will be reduced, thereby ending up with more "neutral" monolingual embeddings.

In the following, we detail our methodological approach. First, we leverage the same bilingual dictionary that was used to obtain the initial alignment (Section 3.1). Specifically, let $D = \{(w, w')\}$ be the given bilingual dictionary, where $w \in V$ and $w' \in V'$, with V and V' representing the vocabulary of the first and second language, respectively. For pairs $(w, w') \in D$, we can simply compute the corresponding average vector $\vec{\mu}_{w,w'} = \frac{\vec{v}_w + \vec{v}_{w'}}{2}$. Then, using the pairs in D as training data, we learn a linear mapping X such that $X\vec{v}_w \approx \vec{\mu}_{w,w'}$ for all $(w, w') \in D$. This mapping X can then be used to predict the averages for words outside the given dictionary. To find the mapping X, we solve the following least squares linear regression problem:

$$E = \sum_{(w,w') \in D} \| X\vec{w} - \vec{\mu}_{w,w'} \|^2 \qquad (1)$$

Similarly, for the other language, we separately learn a mapping X' such that $X'\vec{v}_{w'} \approx \vec{\mu}_{w,w'}$.

It is worth pointing out that we experimented with several variants of this linear regression formulation. For example, we also tried using a multilayer perceptron to learn non-linear mappings, and we experimented with several regularization terms to penalize mappings that deviate too much from the identity mapping. None of these variants, however, were found to improve on the much simpler formulation in (1), which can be solved exactly and efficiently. Furthermore, one may wonder whether the initial alignment is actually needed, since e.g., Coates and Bollegala (2018) obtained high-quality meta-embeddings without such an alignment set. However, when applying our approach directly to the initial monolingual non-aligned embedding spaces, we obtained results which were competitive but slightly below the two considered alignment strategies.

4 Evaluation

We test our bilingual embedding refinement approach on both intrinsic and extrinsic tasks. In Section 4.1 we describe the common training setup for all experiments and language pairs. The languages we considered are English, Spanish, Italian, German and Finnish. Throughout all the experiments we use publicly available resources in order to make comparisons and reproducibility of our experiments easier.

4.1 Cross-lingual embeddings training

Corpora. In our experiments we make use of web-extracted corpora. For English we use the 3B-word UMBC WebBase Corpus (Han et al., 2013), while we chose the Spanish Billion Words Corpus (Cardellino, 2016) for Spanish. For Italian and German, we use the itWaC and sdeWaC corpora from the WaCky project (Baroni et al., 2009), containing 2 and 0.8 billion words, respectively.[2] Lastly, for Finnish, we use the Common Crawl monolingual corpus from the Machine Translation of News Shared Task 2016[3], composed of 2.8B words. All corpora are tokenized and lowercased.

Monolingual embeddings. The monolingual word embeddings are trained with the Skipgram model from FastText (Bojanowski et al., 2017) on the corpora described above. The dimensionality of the vectors was set to 300, with the default FastText hyperparameters.

Bilingual dictionaries. We use the bilingual dictionaries packaged together by Artetxe et al. (2017), each one conformed by 5000 word translations. They are used both for the initial bilingual mappings and then again for our linear transformation.

Initial mapping. Following previous works, for the purpose of obtaining the initial alignment, English is considered as source language and the remaining languages are used as target. We make use of the open-source implementations of **VecMap**[4] (Artetxe et al., 2017) and **MUSE**[5] (Conneau et al., 2018), which constitute strong baselines for our experiments (cf. Section 3.1). Both of them were used with the recommended parameters and in their supervised setting, using the aforementioned bilingual dictionaries.

Meeting in the Middle. Then, once the initial cross-lingual embeddings are trained, and as explained in Section 3.2, we obtain our linear transformation by using the exact solution to the least

---

[2]UMBC, Spanish Billion-Words and ItWaC are the official corpora of the hypernym discovery SemEval task (Section 4.2.3) for English, Spanish and Italian, respectively.

[3]http://www.statmt.org/wmt16/translation-task.html

[4]github.com/artetxem/vecmap

[5]github.com/facebookresearch/MUSE

Model	EN-ES			EN-IT			EN-DE			EN-FI		
	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$
VecMap	36.0	59.8	65.6	35.5	57.2	63.9	31.7	54.2	60.2	17.2	36.4	43.7
VecMap$_\mu$	**37.8**	**61.5**	**67.1**	**36.3**	**59.2**	**66.3**	**33.5**	**57.3**	**61.7**	**18.5**	**40.9**	**48.3**
MUSE	37.1	59.0	65.2	36.3	57.3	62.9	32.5	53.7	59.0	18.2	35.2	42.4
MUSE$_\mu$	**38.3**	**62.3**	**67.2**	**37.0**	**59.0**	**65.7**	**33.7**	**57.0**	**62.2**	**19.4**	**41.1**	**49.0**

Table 1: Bilingual dictionary induction results. Precision at k ($P@K$) performance for Spanish (ES), Italian (IT), German (DE) and Finnish (FI), using English (EN) as source language.

squares linear regression problem. To this end, we use the same bilingual dictionaries as in the previous step. Henceforth, we will refer to our transformed models as **VecMap**$_\mu$ and **MUSE**$_\mu$, depending on the initial mapping.

4.2 Experiments

We test our cross-lingual word embeddings in two intrinsic tasks, i.e., bilingual dictionary induction (Section 4.2.1) and word similarity (Section 4.2.2), and an extrinsic task, i.e., cross-lingual hypernym discovery (Section 4.2.3).

4.2.1 Bilingual dictionary induction

The dictionary induction task consists in automatically generating a bilingual dictionary from a source to a target language, using as input a list of words in the source language.

Experimental setting For this task, and following previous works, we use the English-Italian test set released by Dinu et al. (2015) and those released by Artetxe et al. (2017) for the remaining language pairs. These test sets have no overlap with respect to the training and development sets, and contain around 1900 entries each. Given an input word from the source language, word translations are retrieved through a nearest-neighbor search of words in the target language, using cosine distance. Note that this gives us a ranked list of candidates for each word from the source language. Accordingly, the performance of the embeddings is evaluated with the precision at k ($P@k$) metric, which evaluates for what percentage of test pairs, the correct answer is among the k highest ranked candidates.

Results As can be seen in Table 1, our refinement method consistently improves over the baselines (i.e., VecMap and MUSE) on all language pairs and metrics. The higher scores indicate that the two monolingual embedding spaces become more tightly integrated because of our additional

transformation. It is worth highlighting here the case of English-Finnish, where the gains obtained in $P@5$ and $P@10$ are considerable. This might indicate that our approach is especially useful for morphologically richer languages such as Finnish, where the limitations of the previous bilingual mappings are most apparent.

Analysis When analyzing the source of errors in $P@1$, we came to similar conclusions as Artetxe et al. (2017).[6] Several source words are translated to words that are closely related to the one in the gold reference in the target language; e.g., for the English word *essentially* we obtain *básicamente* (*basically*) instead of *fundamentalmente* (*fundamentally*) in Spanish, both of them closely related, or the closest neighbor for *dirt* being *mugre* (*dirt*) instead of *suciedad* (*dirt*), which in fact was among the five closest neighbors. We can also find multiple examples of the higher performance of our models compared to the baselines. For instance, in the English-Spanish cross-lingual models, after the initial alignment, we can find that *seconds* has *minutos* (*minutes*) as nearest neighbour, but after applying our additional transformation, *seconds* becomes closest to *segundos* (*seconds*). Similarly, *paint* initially has *tintado* (*tinted*) as the closest Spanish word, and then *pintura* (*paint*).

4.2.2 Word similarity

We perform experiments on both monolingual and cross-lingual word similarity. In monolingual similarity, models are tested in their ability to determine the similarity between two words in the same language, whereas in cross-lingual similarity the words belong to different languages. While in the monolingual setting the main objective is to test the quality of the monolingual subsets of the bilin-

---

[6]The results on this task are lower than those reported in Artetxe et al. (2017). This is due to the different corpora and embedding algorithms used to train the monolingual embeddings. In particular, they use corpora including Wikipedia, which is comparable across languages.

Model	English								Spanish				Italian				German					
	SemEval		WordSim		SimLex		RG-65		SemEval		RG-65		SemEval		WordSim		SemEval		WordSim		RG-65	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
VecMap	74.1	73.9	67.9	67.0	42.0	40.7	77.8	**77.5**	70.0	71.4	86.6	88.0	67.2	**69.0**	64.0	66.9	70.1	70.1	72.7	72.2	80.2	79.7
VecMap$_\mu$	75.0	74.8	70.5	70.1	43.8	41.8	78.0	76.6	71.5	72.1	87.6	89.4	68.4	68.9	65.3	67.3	70.9	70.7	72.7	72.4	81.0	81.3
MUSE	74.2	74.2	68.3	67.6	42.6	41.5	78.6	**78.4**	70.5	71.9	86.6	88.3	67.4	**69.2**	64.1	66.9	69.8	69.8	72.5	72.5	80.3	**80.1**
MUSE$_\mu$	75.0	74.8	70.8	70.4	44.2	42.4	78.3	77.5	71.8	72.3	87.7	89.3	68.6	69.1	65.3	67.2	70.4	70.2	72.2	72.1	80.3	80.0

Table 2: Monolingual word similarity results. Pearson (r) and Spearman (ρ) correlation.

gual vector space, the cross-lingual setting constitutes a straightforward benchmark to test the quality of bilingual embeddings.

Experimental setting For monolingual word similarity we use the English SimLex-999 (Hill et al., 2015), and the language-specific versions of SemEval-17[7] (Camacho-Collados et al., 2017), WordSim-353[8] (Finkelstein et al., 2002), and RG-65 (Rubenstein and Goodenough, 1965). The corresponding cross-lingual datasets from SemEval-18, WordSim-353 and RG-65 were considered for the cross-lingual word similarity evaluation[9]. Cosine similarity is again used as comparison measure.

Results Tables 2 and 3 show the monolingual[10] and cross-lingual word similarity results[11], respectively. For both the monolingual and cross-lingual settings, we can notice that our models generally outperform the corresponding baselines. Moreover, in cases where no improvement is obtained, the differences tend to be minimal, with the exception of RG-65, but this is a very small test set for which larger variations can thus be expected. In contrast, there are a few cases where substantial gains were obtained by using our model. This is most notable for English WordSim and SimLex in the monolingual setting.
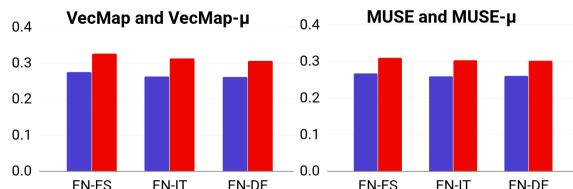

Figure 1: Comparative average similarity between VecMap and MUSE (blue) and our proposed model (red) on the SemEval cross-lingual similarity datasets.

Analysis In order to further understand the movements of the space with respect to the original VecMap and MUSE spaces, Figure 1 displays the average similarity values on the SemEval cross-lingual datasets (the largest among all benchmarks) of each model. As expected, the figure clearly shows how our model consistently brings the words from both languages closer on all language pairs. Furthermore, this movement is performed smoothly across all pairs, i.e., our model does not make large changes to specific words but rather small changes overall. This can be verified by inspecting the standard deviation of the difference in similarity after applying our transformation. These standard deviation scores range from 0.031 (English-Spanish for VecMap) to 0.039 (English-Italian for MUSE), which are relatively small given that the cosine similarity scale ranges from -1 to 1.

As a complement of this analysis we show some qualitative results which give us further insights on the transformations of the vector space after our average approximation. In particular, we analyze the reasons behind the higher quality displayed by our bilingual embeddings in monolingual settings. While VecMap and MUSE do not transform the initial monolingual spaces, our model transforms both spaces simultaneously. In this analysis we focus on the source language of our experiments (i.e., English). We found interesting patterns which are learned by our model and

---

[7]The original datasets of SemEval-17 contained also multiwords, but for consistency we use the version containing single words only.

[8]WordSim datasets consist of the similarity re-scoring for several languages of Leviant and Reichart (2015), downloaded from http://leviants.com/ira.leviant/MultilingualVSMdata.html

[9]The WordSim-353 and RG-65 cross-lingual datasets (Camacho-Collados et al., 2015) were downloaded at http://lcl.uniroma1.it/similarity-datasets/

[10]The English results correspond to the averaged performance of the English fragments of English-Spanish, English-Italian and English-German cross-lingual embeddings.

[11]The results of the original VecMap in cross-lingual similarity are comparable or better to those reported in Artetxe et al. (2017) on the three datasets used in their evaluation.

Model	English-Spanish				English-Italian				English-German					
	SemEval		RG-65		SemEval		WordSim		SemEval		WordSim		RG-65	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
VecMap	71.7	**71.6**	82.1	82.4	69.6	69.6	60.2	**63.1**	71.6	71.3	64.1	**65.9**	78.1	78.8
VecMap$_\mu$	71.7	71.3	82.1	**82.8**	**70.2**	**69.9**	**61.3**	63.0	**72.0**	**71.5**	**64.2**	65.4	**78.6**	**79.7**
MUSE	72.0	**72.0**	81.9	82.3	69.4	69.4	59.9	62.7	70.4	70.1	63.5	65.1	78.4	79.5
MUSE$_\mu$	**72.2**	71.8	**82.3**	**82.5**	**70.5**	**70.1**	**61.2**	62.7	**71.9**	71.4	**64.1**	65.3	**78.8**	**80.5**

Table 3: Cross-lingual word similarity results. Pearson (r) and Spearman (ρ) correlation.

help understand these monolingual gains. For example, a recurring pattern is that words in English which are translated to the same word, or to semantically close words, in the target language end up closer together after our transformation. For example, in the case of English-Spanish the following pairs were among the pairs whose similarity increased the most by applying our transformation: *cellphone-telephone*, *movie-film*, *book-manuscript* or *rhythm-cadence*, which are either translated to the same word in Spanish (i.e., *teléfono* and *película* in the first two cases) or are already very close in the Spanish space. More generally, we found that word pairs which move together the most tend to be semantically very similar and belong to the same domain, e.g., *car-bicycle*, *opera-cinema*, or *snow-ice*.

4.2.3 Cross-lingual hypernym discovery

Modeling hypernymy is a crucial task in NLP, with direct applications in diverse areas such as semantic search (Hoffart et al., 2014; Roller and Erk, 2016), question answering (Prager et al., 2008; Yahya et al., 2013) or textual entailment (Geffet and Dagan, 2005). Hypernyms, in addition, are the backbone of lexical ontologies (Yu et al., 2015), which are in turn useful for organizing, navigating and retrieving online content (Bordea et al., 2016). Thus, we propose to evaluate the contribution of cross-lingual embeddings towards the task of hypernym discovery, i.e., given an input word (e.g., *cat*), retrieve or discover its most likely (set of) valid hypernyms (e.g., *animal*, *mammal*, *feline*, and so on). Intuitively, by leveraging a bilingual vector space condensing the semantics of two languages, one of them being English, the need for large amounts of training data in the target language may be reduced.

Experimental setting We follow Espinosa-Anke et al. (2016) and learn a (cross-lingual) linear transformation matrix between the hyponym and hypernym spaces, which is afterwards used to predict the most likely (set of) hypernyms, given an unseen hyponym. Training and evaluation data come from the SemEval 2018 Shared Task on Hypernym Discovery (Camacho-Collados et al., 2018). Note that current state-of-the-art systems aimed at modeling hypernymy (Shwartz et al., 2016; Bernier-Colborne and Barriere, 2018) combine large amounts of annotated data along with language-specific rules and cue phrases such as Hearst Patterns (Hearst, 1992), both of which are generally scarcely (if at all) available for languages other than English. Therefore, we report experiments with training data only from English (11,779 hyponym-hypernym pairs), and "enriched" models informed with relatively few training pairs (500, 1k and 2k) from the target languages. Evaluation is conducted with the same metrics as in the original SemEval task, i.e., Mean Reciprocal Rank (MRR), Mean Average Precision (MAP) and Precision at 5 (P@5). These measures explain a model's behavior from complementary prisms, namely how often at least one valid hypernym was highly ranked (MRR), and in cases where there is more than one correct hypernym, to what extent they were all correctly retrieved (MAP and P@5). Finally, as in the previous experiments, we report comparative results between our proposed models and the two competing baselines (VecMap and MUSE). As an additional informative baseline, we include the highest scoring unsupervised system at the SemEval task for both Spanish and Italian (BestUns), which is based on the distributional models described in Shwartz et al. (2017).

Results The results listed in Table 4 indicate several trends.[12] First and foremost, in terms of

---

[12] Note that this task is harder than hypernymy detection (Upadhyay et al., 2018). Hypernymy detection is framed as a binary classification task, while in hypernym discovery hypernyms have to be retrieved from the whole vocabulary.

Train data	Model	Spanish			Italian		
		MAP	MRR	P@5	MAP	MRR	P@5
-	**BestUns**	2.4	5.5	2.5	3.9	8.7	3.9
EN	**VecMap**	**6.4**	**16.5**	**6.0**	4.5	10.6	4.3
	VecMap$_\mu$	6.1	15.4	5.7	**5.6**	**13.3**	**5.4**
	MUSE	5.9	14.1	5.5	4.9	11.1	4.7
	MUSE$_\mu$	**6.2**	**14.8**	**5.8**	**5.1**	**11.7**	**4.9**
EN + 500	**VecMap**	**7.3**	**18.2**	**7.0**	6.1	14.0	5.8
	VecMap$_\mu$	7.0	17.6	6.6	**6.8**	**16.2**	**6.4**
	MUSE	6.4	15.9	6.1	5.3	12.0	5.0
	MUSE$_\mu$	**6.9**	**16.9**	**6.6**	**6.0**	**13.4**	**5.7**
EN + 1k	**VecMap**	**7.9**	19.2	**7.6**	7.0	16.4	6.6
	VecMap$_\mu$	7.8	19.2	7.4	**7.5**	**18.1**	**7.0**
	MUSE	7.2	17.3	6.9	6.2	13.8	5.8
	MUSE$_\mu$	**7.8**	**18.8**	**7.5**	**6.5**	**14.2**	**6.3**
EN + 2k	**VecMap**	8.0	19.1	7.7	8.2	19.1	7.5
	VecMap$_\mu$	**8.2**	**19.9**	**7.9**	**8.7**	**20.7**	**8.1**
	MUSE	7.2	17.2	6.8	7.2	15.8	7.0
	MUSE$_\mu$	**8.3**	**19.5**	**8.0**	**7.6**	**17.0**	**7.2**

Table 4: Results on the hypernym discovery task.

model-wise comparisons, we observe that our proposed alterations of both `VecMap` and `MUSE` improve their quality in a consistent manner, across most metrics and data configurations. In Italian our proposed model shows an improvement across all configurations. However, in Spanish `VecMap` emerges as a highly competitive baseline, with our model only showing an improved performance when training data in this language abounds (in this specific case there is an increase from 17.2 to 19.5 points in the MRR metric). This suggests that the fact that the monolingual spaces are closer in our model is clearly beneficial when hybrid training data is given as input, opening up avenues for future work on weakly-supervised learning. Concerning the other baseline, `MUSE`, the contribution of our proposed model is consistent for both languages, again becoming more apparent in the Italian split and in a fully cross-lingual setting, where the improvement in MRR is almost 3 points (from 10.6 to 13.3). Finally, it is noteworthy that even in the setting where no training data from the target language is leveraged, all the systems based on cross-lingual embeddings outperform the best unsupervised baseline, which is a very encouraging result with regards to solving tasks for languages on which training data is not easily accessible or not directly available.

Analysis A manual exploration of the results obtained in cross-lingual hypernym discovery reveals a systematic pattern when comparing, for example, `VecMap` and our model. It was shown in Table 4 that the performance of our model gradually increased alongside the size of the training data in the target language until surpassing `VecMap` in the most informed configuration (i.e., EN+2k). Specifically, our model seems to show a higher presence of generic words in the output hypernyms, which may be explained by these being closer in the space. In fact, out of 1000 candidate hyponyms, our model correctly finds *person* 143 times, as compared to the 111 of `VecMap`, and this systematically occurs with generic types such as *citizen* or *transport*. Let us mention, however, that the considered baselines perform remarkably well in some cases. For example, the English-only `VecMap` configuration (EN), unlike ours, correctly discovered the following hypernyms for *Francesc Macià* (a Spanish politician and soldier): *politician*, *ruler*, *leader* and *person*. These were missing from the prediction of our model in all configurations until the most informed one (EN+2k).

5 Conclusions and Future Work

We have shown how to refine bilingual word embeddings by applying a simple transformation which moves cross-lingual synonyms closer towards their average representation. Before applying this strategy, we start by aligning the monolingual embeddings of the two languages of interest. For this initial alignment, we have considered two state-of-the-art methods from the literature, namely `VecMap` (Artetxe et al., 2017) and `MUSE` (Conneau et al., 2018), which also served as our baselines. Our approach is motivated by the fact that these alignment methods do not change the structure of the individual monolingual spaces. However, the internal structure of embeddings is, at least to some extent, language-specific, and is moreover affected by biases of the corpus from which they are trained, meaning that after the initial alignment significant gaps remain between the representations of cross-lingual synonyms. We tested our approach on a wide array of datasets from different tasks (i.e., bilingual dictionary induction, word similarity and cross-lingual hypernym discovery) with state-of-the-art results.

This paper opens up several promising avenues for future work. First, even though both languages are currently being treated symmetrically, the initial monolingual embedding of one of the

languages may be more reliable than that of the other. In such cases, it may be of interest to replace the vectors $\vec{\mu}_{w,w'}$ by a weighted average of the monolingual word vectors. Second, while we have only considered bilingual scenarios in this paper, our approach can naturally be applied to scenarios involving more languages. In this case, we would first choose a single target language, and obtain alignments between all the other languages and this target language. To apply our model, we can then simply learn mappings to predict averaged word vectors across all languages. Finally, it would also be interesting to use the obtained embeddings in downstream applications such as language identification or cross-lingual sentiment analysis, and extend our analysis to other languages, with a particular focus on morphologically-rich languages (after seeing our success with Finnish), for which the bilingual induction task has proved more challenging for standard cross-lingual embedding models (Søgaard et al., 2018).

Acknowledgments

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

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

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

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Gabriel Bernier-Colborne and Caroline Barriere. 2018. Crim at semeval-2018 task 9: A hybrid approach to hypernym discovery. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1081–1091.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*, New Orleans, LA, United States.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A Framework for the Construction of Monolingual and Cross-lingual Word Similarity Datasets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing – Short Papers*, pages 1–7, Beijing, China.

Cristian Cardellino. 2016. Spanish Billion Words Corpus and Embeddings. *http://crscardellino.me/SBWCE/*.

Joshua Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embedding–computing meta-embeddings by averaging source word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 194–198.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of ICLR*.

Ernest Davis. 2015. Does the world look different in different languages? *Artif. Intell.*, 229:202–209.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR, Workshop track*.

Luis Espinosa-Anke, Jose Camacho-Collados, Claudio Delli Bovi, and Horacio Saggion. 2016. Supervised distributional hypernym discovery via domain adaptation. In *Proceedings of EMNLP*, pages 424–435.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Lev Finkelstein, Gabrilovich Evgeniy, Matias Yossi, Rivlin Ehud, Solan Zach, Wolfman Gadi, and Ruppin Eytan. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Maayan Geffet and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 107–114. Association for Computational Linguistics.

Josu Goikoetxea, Aitor Soroa, and Eneko Agirre. 2018. Bilingual embeddings with random walks over multilingual wordnets. *Knowledge-Based Systems*.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages 44–52.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING 1992*, pages 539–545.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 58–68.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Johannes Hoffart, Dragan Milchevski, and Gerhard Weikum. 2014. Stics: searching with strings, things, and cats. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1247–1248. ACM.

Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing procrustes analysis for better bilingual dictionary induction. In *Proceedings of the Conference on Computational Natural Language Learning*.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations*.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *arXiv preprint arXiv:1508.00106*.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 765–774.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Aditya Mogadala and Achim Rettinger. 2016. Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 692–702.

Nikola Mrksic, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gai, Anna Korhonen, and Steve Young. 2017. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. *TACL*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

John Prager, Jennifer Chu-Carroll, Eric W Brown, and Krzysztof Czuba. 2008. Question answering by predictive annotation. In *Advances in Open Domain Question Answering*, pages 307–347. Springer.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Stephen Roller and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. In *Proceedings of EMNLP*, pages 2163–2172, Austin, Texas.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2018. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research.*

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of ACL*, Berlin, Germany.

Vered Shwartz, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of EACL*, Valencia, Spain. Association for Computational Linguistics.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of ACL*, pages 1713–1722.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. *arXiv preprint arXiv:1805.03620.*

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. 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)*, volume 1, pages 1661–1670.

Shyam Upadhyay, Yogarshi Vyas, Marine Carpuat, and Dan Roth. 2018. Robust cross-lingual hypernymy detection using dependency context. In *Proceedings of NAACL*.

Ivan Vulić and Marie-Francine Moens. 2015a. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 719–725.

Ivan Vulić and Marie-Francine Moens. 2015b. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372. ACM.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research*, 55:953–994.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. 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*, pages 1006–1011.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. 2013. Robust question answering over the web of linked data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1107–1116. ACM.

Wenpeng Yin and Hinrich Schütze. 2016. Learning word meta-embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. 2015. Learning term embeddings for hypernymy identification. In *Proceedings of IJCAI*, pages 1390–1397.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of EMNLP*, pages 1393–1398, Seattle, USA.