Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4013–4023
Marseille, 11–16 May 2020

On the Robustness of Unsupervised and Semi-supervised Cross-lingual Word Embedding Learning

Yerai Doval*,[1] Jose Camacho-Collados*,[2] Luis Espinosa-Anke,[2] Steven Schockaert[2]
[1]Grupo COLE, Escola Superior de Enxeñaría Informática, Universidade de Vigo, Spain
yerai.doval@uvigo.es
[2]School of Computer Science and Informatics, Cardiff University, UK
{camachocolladosj, espinosa-ankel, schockaerts1}@cardiff.ac.uk

Abstract

Cross-lingual word embeddings are vector representations of words in different languages where words with similar meaning are represented by similar vectors, regardless of the language. Recent developments which construct these embeddings by aligning monolingual spaces have shown that accurate alignments can be obtained with little or no supervision, which usually comes in the form of bilingual dictionaries. However, the focus has been on a particular controlled scenario for evaluation, and there is no strong evidence on how current state-of-the-art systems would fare with noisy text or for language pairs with major linguistic differences. In this paper we present an extensive evaluation over multiple cross-lingual embedding models, analyzing their strengths and limitations with respect to different variables such as target language, training corpora and amount of supervision. Our conclusions put in doubt the view that high-quality cross-lingual embeddings can always be learned without much supervision.

Keywords: Multilinguality, Evaluation Methodologies, Semantics, Semi-supervised, weakly-supervised and unsupervised learning

1. Introduction

The standard approach for training word embeddings is to rely on monolingual corpora, which means in particular that a separate embedding model is learned for each language. There is an increasing interest, however, in learning cross-lingual word embeddings, where words from different languages are mapped onto a single space. Such representations are attractive, for instance, for dealing with the multilingual nature of text on the Web, but also as a vehicle for transferring knowledge (e.g., labelled training data) from resource-rich languages such as English to other languages (Ruder et al., 2018).

Initially, the main obstacle to learning such cross-lingual embeddings was the need for large multilingual parallel corpora (Klementiev et al., 2012; Chandar et al., 2014; Luong et al., 2015). This limitation, however, was alleviated by the development of methods that only need comparable data (e.g., Wikipedia corpora in different languages) as the main source of supervision (Vulić and Moens, 2015). In a complementary direction, it has recently been shown that high-quality cross-lingual embeddings can be obtained by aligning two independently learned monolingual embedding spaces. This strategy is appealing, as it means that one only needs access to monolingual corpora and a bilingual dictionary as supervision signal. Surprisingly, perhaps, it turns out that dictionaries with less than 100 word pairs are sufficient to obtain good alignments (Artetxe et al., 2017). In fact, recent works have shown that cross-lingual embeddings can even be learned without any user-provided dictionary (Conneau et al., 2018a; Artetxe et al., 2018b; Xu et al., 2018).

Despite the promising results reported in the literature, it remains unclear under which conditions the aforementioned methods succeed. For example, Artetxe et al. (2017) and Conneau et al. (2018a) achieved promising results in the word translation task (i.e., bilingual lexicon induction), but their experiments relied on the availability of high-quality monolingual source corpora, namely Wikipedia, which is also the case in a more recent analysis on cross-lingual embeddings performance (Glavas et al., 2019). In fact, there exists a significant number of settings which have been largely ignored, and which might challenge models that excel in idealized environments. For instance, Ahmad et al. (2019) found that for dissimilar languages with different word orderings than English, cross-lingual transfer is still challenging. Similarly, it remains unclear how well existing methods would perform on language pairs with significant differences in morphology (e.g., English-Finnish, the latter being an agglutinative language) or with different alphabets (e.g., English-Farsi or English-Russian). Moreover, settings with different kinds of corpora (e.g. noisy user-generated) have not been fully explored. This means, among others, that it is not clear how current cross-lingual embedding models would behave for transferring knowledge in environments such as social media centred tasks, given that such tasks usually benefit from embeddings that have been trained on social media corpora (Tang et al., 2014; Godin et al., 2015; Yang et al., 2018).

In this work, we broaden the empirical evaluation of state-of-the-art techniques for learning cross-lingual embeddings, by using several types of training corpora, various amounts of supervision, languages from different families and different alignment strategies in three different tasks. The results obtained cast some doubt on the view that high-quality cross-lingual embeddings can always be learned without much supervision.

Authors marked with an asterisk (*) contributed equally.

2. Related Work

Cross-lingual embeddings have become increasingly popular in the past few years (Smith et al., 2017; Artetxe et al., 2017; Artetxe et al., 2018a; Conneau et al., 2018a). Recent efforts have focused on reducing the need for large amounts of resources (e.g., parallel corpora), which could be difficult to obtain for most languages and language pairs. However, the evaluation of these approaches has tended to be somewhat limited, often using only one type of training corpus, including only similar languages, and considering only one evaluation task. The most similar work to ours is that of Søgaard et al. (2018), which included an in-depth analysis of two of the factors that we also considered, namely language family and corpus type, but they only considered a single model, i.e., `MUSE` (Conneau et al., 2018a). Moreover, they studied each factor in isolation. In our case the analysis is also extended to more languages (covering up to 5 language pairs), systems (two unsupervised, two supervised, and a postprocessing technique), evaluation tasks (cross-lingual word similarity), and the impact of external bilingual dictionaries.

Another similar contribution is the analysis by Vulić and Korhonen (2016), where the impact of bilingual dictionaries on cross-lingual alignments was examined. However, they only considered closely-related languages using the same alphabet and one type of corpus (i.e., Wikipedia). Also, given the publication date, this analysis does not account for the important developments in cross-lingual embeddings from recent years, such as the methods we cover in this paper. Other empirical comparisons focused mostly on the need for different degrees of supervision, such as (Upadhyay et al., 2016), which has been extended in a more recent survey by Ruder et al. (2018).

In this paper, we complement those studies by analyzing and discussing empirical findings of the most recent state-of-the-art unsupervised and semi-supervised methods in a broader experimental setting, more in line with the recent concurrent analysis of Glavas et al. (2019). The main differences between this empirical evaluation and the contributions of our work lie in the scope of the survey, since: (1) they only consider Wikipedia data for training; (2) they do not consider postprocessing techniques such as `Meemi` (Doval et al., 2018), which we found to improve the performance of cross-lingual models, especially in the case of distant languages and non-comparable corpora; (3) in our analysis we also consider additional settings with scarce training data such as small seed dictionaries and automatically-constructed dictionaries; and (4) we include a more exhaustive intrinsic evaluation (including cross-lingual semantic similarity).

3. Learning Cross-lingual Word Embeddings

The focus of our evaluation is on methods that start off with monolingual embedding models and then integrate these in a shared cross-lingual space.

Hence, 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., 2013), 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). In some cases, a third transformation is applied to already aligned embeddings so the word vectors from both languages are refined and further re-positioned (Section 3.2). Regardless of the overall methodology, however, these linear transformations are all learned based on a bilingual dictionary. This dictionary may be manually curated or, in some cases, automatically generated as part of the alignment process.

3.1. Alignment methods

In this paper we analyze two well-known orthogonal models for aligning monolingual embedding models: the corresponding version of `VecMap` and `MUSE`, plus an unsupervised non-orthogonal variant of the former. Basically, both methods use a linear transformation learned through an iterative procedure in which a seed bilingual dictionary is iteratively refined. They can be used with an empty initial seed dictionary, in which case the alignment process is fully unsupervised.

`VecMap` (Artetxe et al., 2017) uses an orthogonal transformation over normalized word embeddings. Its semi-supervised two-step procedure is specifically designed to avoid the need for a large seed dictionary. For instance, in the original paper, a seed dictionary with 25 word pairs was used. This seed dictionary is then augmented by applying the learned transformation to new words from the source language. The process is repeated until some convergence criterion is met. The unsupervised variant (Artetxe et al., 2018b) obtains the initial seed dictionary automatically by exploiting the similarity distribution of words, and then applies the same method followed by a refinement step that reweights the embeddings based on the cross-correlation of their components, which makes it the only non-orthogonal method tested in this work. `MUSE` (Conneau et al., 2018a) obtains its transformation matrix in a similar way. In this case, the seed dictionary is used as-is (supervised setting) or obtained in a fully automatically way through an adversarial learning method (unsupervised setting).

3.2. Limitations and postprocessing

By restricting transformations to orthogonal linear mappings, `VecMap` and `MUSE` rely on the assumption that the monolingual embeddings spaces are approximately isomorphic (Barone, 2016). However, it has been argued that this assumption is overly restrictive, as the isomorphism assumption is not always satisfied (Søgaard et al., 2018; Kementchedjhieva et al., 2018). For this reason, it has been proposed to go beyond orthogonal transformations by modifying the internal structure of the monolingual spaces, either by giving more weight to highly correlated embedding components, as is the case for the unsupervised variant of `VecMap` in this work (Artetxe et al., 2018a), or by complementing the orthogonal transformation with other forms of post-processing. As an example of this latter strategy, Doval et al. (2018) fine-tune the initial alignment by learning an unconstrained linear transformation which aims to map each word vector onto the average of that vector and the corresponding word vector from the other language.

4. Variables

Our main aim is to explore how the choice of corpora (Section 4.1), supervision signals (Section 4.2) and languages (Section 4.3) impacts the performance of cross-lingual word embedding models. In Section 4.4 we also list some other variables which were not directly studied in this paper.

4.1. Monolingual corpora

It is reasonable to assume that accurate word-level alignments will be easier to obtain from corpora from similar domains with similar vocabularies and register. Wikipedia has been the mainstream monolingual source in cross-lingual word embedding training so far (Artetxe et al., 2017; Conneau et al., 2018a). It provides a particularly reliable bilingual signal because of the highly comparative nature of Wikipedia corpora from different languages. As we will see, this makes finding high-quality alignments considerably easier.

In our analysis we use three different types of corpora: Wikipedia[1] (as a prototypical example of comparable monolingual corpora), Web corpora from different sources[2] (as a prototypical example of non-comparable but generally high-quality corpora) and social media[3] (as a prototypical example of noisy text). Statistics of these corpora are provided in Table 1.[4]

4.2. Bilingual supervision

Early approaches for learning bilingual embeddings relied on large parallel corpora (Klementiev et al., 2012; Luong et al., 2015), which limited their applicability. More recent approaches instead rely on (often small) bilingual dictionaries as the only source of bilingual supervision. In fact, some methods remove the need for a user-supplied bilingual dictionary altogether (Conneau et al., 2018a; Artetxe et al., 2018b; Hoshen and Wolf, 2018; Xu et al., 2018), relying instead on synthetic dictionaries that are obtained fully automatically. In our experiments we consider a wide range of signals, including no supervision as well as automatically generated dictionaries of identical words. In the latter case, we rely on the assumption that words that occur in both of the monolingual corpora tend to have the same meaning. While this may seem naive, this strategy has been reported in the literature to perform well in practice (Smith et al., 2017; Søgaard et al., 2018).

Domain	Corpus	Language	Size	Words
Wikipedia	Wiki_{en}	English	1.7B	12.0M
	Wiki_{es}	Spanish	407M	3.4M
	Wiki_{it}	Italian	338M	3.3M
	Wiki_{de}	German	605M	7.4M
	Wiki_{fi}	Finnish	68M	2.8M
	Wiki_{ru}	Russian	313M	5.4M
	Wiki_{fa}	Farsi	48M	1.0M
Web corpora	UMBC	English	3.5B	8.1M
	1-billion	Spanish	1.9B	5.5M
	itWaC	Italian	1.3B	4.2M
	sdeWaC	German	438M	1.5M
	Comm-crawl	Finnish	2.8B	1.8M
	Comm-crawl	Russian	1.1B	18.8M
	Hamshahri	Farsi	167M	0.8M
Social media	Twitter_{en}	English	294M	5.5M
	Twitter_{es}	Spanish	144M	3.3M
	Twitter_{it}	Italian	63M	1.6M
	Twitter_{de}	German	114M	2.3M
	Twitter_{fi}	Finnish	29M	1.7M
	Twitter_{fa}	Farsi	90M	1.0M

Table 1: Statistics of the corpora used to train monolingual word embeddings: size (measured in total number of tokens) and words (number of unique tokens).

4.3. Languages

In most previous work, the evaluation of cross-lingual embeddings has been limited to a small set of closely-related languages. For instance, Smith et al. (2017) evaluated their model on the English-Italian pair only, while the evaluation of Artetxe et al. (2017) was performed on three languages, all of which share the same alphabet. Moreover, as the considered language pairs vary from one study to another, the relative performance of different methods for particular types of languages remains unclear. More recently, however, Søgaard et al. (2018) have extended the usual evaluation framework by covering additional Eastern European languages. We similarly expand the range of languages by considering: Spanish (ES), Italian (IT), German (DE), Finnish (FI), Farsi (FA) and Russian (RU). In all cases we use English (EN) as source language. This set of languages represents not only the usual family of Indo-European languages (all of them except Finnish), but also agglutinative languages (German, Farsi and Finnish, the latter being non-Indo-European), as well as languages with different alphabets (Farsi and Russian).

4.4. Other variables

It is worth mentioning that there are several other external factors that may affect the quality of cross-lingual embeddings, beyond the ones considered in this study. For instance, in our experiments we use FastText (Bojanowski et al., 2017), since morphological information might be useful for agglutinative languages as noted by its authors, with default values and dimensionality[5], but the impact of other

[1]All Wikipedia text dumps were downloaded from the Polyglot project (Al-Rfou et al., 2013): `https://sites.google.com/site/rmyeid/projects/polyglot`

[2]The sources of the web-corpora are: UMBC (Han et al., 2013), 1-billion (Cardellino, 2016), itWaC and sdeWaC (Baroni et al., 2009), Hamshahri (AleAhmad et al., 2009), and Common Crawl downloaded from `http://www.statmt.org/wmt16/translation-task.html`.

[3]Social media corpora are based on Twitter, at different dates between 2015 and 2018 (Camacho-Collados et al., 2020). Monolingual embeddings were downloaded at `https://github.com/pedrada88/crossembeddings-twitter`

[4]Due to some restrictions, we were not able to compile a reliable Twitter corpus for Russian.

[5]300 dimensions in the case of Wikipedia and web corpora, and 100 dimensions in the smaller social media corpora.

`Wikipedia`

Sup.	Model	Spanish			Italian			German			Finnish			Farsi			Russian			Avg
		$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@5$
`Unsup`	**VecMap**	**39.6**	66.1	72.3	42.7	65.7	71.6	**28.6**	**48.3**	**54.8**	19.6	40.4	48.3	20.5	37.0	42.8	19.5	45.3	54.5	50.5
	MUSE	39.3	64.7	71.3	41.6	63.2	69.9	28.3	46.5	53.3	0.0	0.0	0.0	0.0	0.0	0.0	14.9	36.0	46.5	35.1
`Ident`	**VecMap**	**39.5**	66.0	72.4	42.7	65.8	71.7	**28.6**	**48.3**	54.7	21.6	43.7	51.6	23.4	40.3	46.1	19.6	46.0	55.1	51.7
	MUSE	35.9	60.6	67.3	37.8	60.4	68.5	24.8	41.9	49.5	13.4	25.5	32.0	6.7	16.6	21.3	7.8	19.9	26.1	37.5
`8K`	**VecMap**	**39.6**	66.2	72.3	**42.6**	65.9	71.8	**28.6**	**48.3**	**54.8**	22.4	44.5	52.5	22.8	39.7	46.2	20.0	46.3	55.6	51.8
	MUSE	39.1	65.4	72.3	41.1	63.3	70.1	27.6	45.9	53.2	19.5	40.4	49.5	19.7	35.4	42	21.3	43.7	52.9	49.0
	Meemi$_{\mathrm{VM}}$	39.3	**67.4**	**73.7**	41.6	66.5	72.5	28	47.8	**54.8**	**23.8**	**48.7**	**57.0**	**23.4**	**41.7**	**47.7**	23.0	49.3	58.3	**53.4**
	Meemi$_{\mathrm{MS}}$	39.3	**67.4**	**73.7**	41.3	**66.8**	**72.8**	27.1	46.3	53.9	21.7	45.0	53.6	20.7	38.6	45.1	**24.4**	**50.3**	**59.3**	52.4

`Web corpora`

Sup.	Model	Spanish			Italian			German			Finnish			Farsi			Russian			Avg
		$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@5$
`Unsup`	**VecMap**	**34.8**	60.6	67.0	31.4	53.7	60.7	23.2	42.7	50.2	0.0	0.0	0.0	19.7	34.6	40.4	13.8	30.9	38.6	37.1
	MUSE	31.4	51.2	57.7	31.4	51.2	57.7	20.8	38.7	46.6	17.7	35.7	42.8	18.1	32.8	37.8	0.0	0.0	0.0	34.9
`Ident`	**VecMap**	34.7	60.4	67.0	31.4	54.0	60.7	23.1	42.9	50.5	18.6	41.6	49.3	20.0	35.3	40.3	14.1	31.2	38.8	44.2
	MUSE	26.1	46.7	53.8	24.7	45.1	52.4	17.4	32.8	40.5	12.6	26.0	33.8	3.0	8.3	5.8	0.1	0.2	0.2	26.5
`8K`	**VecMap**	34.6	60.6	66.9	31.9	54.2	60.4	23.1	42.7	50.5	18.9	40.9	48.8	19.6	35.8	41.4	14.6	31.7	39.6	44.3
	MUSE	32.5	58.2	65.9	32.5	56.0	63.2	22.4	40.9	48.9	20.0	40.1	48.3	17.4	31.6	37.6	15.5	35.6	44.1	43.7
	Meemi$_{\mathrm{VM}}$	34.5	**61.6**	67.9	33.6	58.3	**65.6**	**23.7**	**45.4**	**53.2**	22.3	**46.7**	**55.0**	**21.7**	**39.0**	**43.8**	18.2	40.0	47.5	**48.9**
	Meemi$_{\mathrm{MS}}$	33.9	60.7	**68.4**	**33.8**	**58.4**	**65.6**	**23.7**	45.3	52.3	**23.0**	46.1	54.0	19.3	36.0	41.7	**18.7**	**40.5**	**49.7**	47.8

`Social media`

Sup.	Model	Spanish			Italian			German			Finnish			Farsi			Avg
		$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@5$
`Unsup`	**VecMap**	8.1	16.4	20.4	8.8	17.0	22.3	0.1	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0	6.8
	MUSE	0.0	0.0	0.0	7.3	14.5	18.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	2.9
`Ident`	**VecMap**	8.1	16.4	20.4	8.8	17.0	22.3	0.1	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0	8.2
	MUSE	0.0	0.0	0.0	7.3	14.5	18.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	3.7
`8K`	**VecMap**	8.7	16.6	21.6	8.9	17.3	22.4	3.2	6.8	9.5	0.2	0.8	1.2	0.4	1.6	2.0	8.6
	MUSE	8.1	17.6	22.7	8	16.4	21.1	2.2	6.0	8.4	0.6	2.2	3.2	1.2	4.5	6.3	9.3
	Meemi$_{\mathrm{VM}}$	**9.8**	**21.3**	**26.9**	**10.6**	**20.0**	**25.6**	**3.7**	**9.6**	**13.2**	1.3	3.6	5.5	**1.8**	5.1	7.0	**12.1**
	Meemi$_{\mathrm{MS}}$	9.5	20.5	26.3	9.5	19.1	24.5	3.0	7.6	11.1	**1.5**	**4.3**	**6.4**	1.6	**5.3**	**8.1**	11.4

Table 2: Bilingual dictionary induction results using English as source language. Performance measured by $P@k$. Overall average $P@5$ is shown in the last column.

word embedding models such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) could also be analyzed, in the line of Søgaard et al. (2018). Likewise, all cross-lingual models and post-processing technique we evaluate are used *as is*, with their default configurations.

5. Evaluation

We use two standard tasks for evaluating cross-lingual word embeddings: bilingual dictionary induction (Section 5.1) and cross-lingual word similarity (Section 5.2). In addition, we also consider a downstream application: cross-lingual natural language inference (Section 5.3).

The systems we compare are two well-known cross-lingual embedding methods which can be used in unsupervised and semi-supervised settings, namely the orthogonal version of `VecMap`[6] (Artetxe et al., 2018b) and `MUSE`[7] (Conneau et al., 2018a). As seed dictionaries we consider three samples of varying sizes, considering 8K, 1K and 100 word pairs, to test the robustness of the models regarding the amount of supervision available.[8] For the sake of clarity, in this section we only present results for the largest dictionary (i.e., with 8K word pairs). The results for all the other dictionary sizes are included in the appendix (these results are also considered in the analysis in Section 6). Additionally, we also leverage synthetic dictionaries, consisting of identical words that are found in the corpora for both languages. Lastly, using those same bilingual dictionaries, we apply the postprocessing proposed in Doval et al. (2018)[9] to refine the cross-lingual embeddings obtained by `VecMap` and `MUSE`. We will refer to these postprocessed vectors as `Meemi`$_{\mathrm{VM}}$ and `Meemi`$_{\mathrm{MS}}$, respectively.

5.1. Bilingual dictionary induction

This task consists in automatically obtaining the word translations in a target language for words in a source language. To obtain the translation candidates, we use the standard cosine distance measure, selecting the nearest neighbors from the target language to the source word in the cross-lingual embedding space. The performance is measured with precision at k ($P@k$), that is, the proportion of test instances where the correct translation candidate for a given source word was among the k highest ranked candidates. Table 2 summarizes the results obtained by all comparison systems on the test dictionaries published by Conneau et al. (2018a). Note that the test dictionaries do not overlap with the dictionaries used for training.

5.2. Cross-lingual semantic word similarity

Given a pair of words from two different languages, the task of cross-lingual semantic word similarity consists in measuring to what extent both words are semantically similar. For the evaluation we make use of the cross-lingual word similarity datasets of the SemEval 2017 task (Camacho-Collados et al., 2017). In this dataset each word from one language is paired with another word from the other language. This evaluation task has been found to correlate

[6] `https://github.com/artetxem/vecmap`

[7] `https://github.com/facebookresearch/MUSE`

[8] These dictionaries were obtained by splitting the training dictionaries provided by Conneau et al. (2018a)

[9] `https://github.com/yeraidm/meemi`

Wikipedia						
Sup.	**Model**	**EN-ES**	**EN-IT**	**EN-DE**	**EN-FA**	**Avg**
Unsup	**VecMap**	72.1	70.6	69.3	61.3	68.3
	MUSE	72.6	71.2	68.9	6.5	54.8
Ident	**VecMap**	71.8	70.6	69.3	61.9	68.4
	MUSE	71.9	70.5	68.4	51.3	65.5
8K	**VecMap**	71.8	70.6	69.3	61.7	68.4
	MUSE	72.6	70.9	68.9	58.7	67.8
	Meemi$_{VM}$	71.9	70.9	**70.3**	**63.4**	69.1
	Meemi$_{MS}$	**72.9**	**71.9**	70.1	62.0	**69.2**
Web corpora						
Sup.	**Model**	**EN-ES**	**EN-IT**	**EN-DE**	**EN-FA**	**Avg**
Unsup	**VecMap**	70.5	68.8	70.4	33.4	60.8
	MUSE	71.6	69.4	70.0	23.8	58.7
Ident	**VecMap**	70.6	68.8	70.4	33.0	60.7
	MUSE	70.1	67.5	69.7	14.5	55.5
8K	**VecMap**	70.6	68.8	70.4	**33.5**	60.8
	MUSE	71.9	70.4	70.2	23.9	59.1
	Meemi$_{VM}$	70.9	70.0	71.8	**39.0**	**62.9**
	Meemi$_{MS}$	**72.3**	**71.1**	**72.1**	33.0	62.1
Social media						
Sup.	**Model**	**EN-ES**	**EN-IT**	**EN-DE**	**EN-FA**	**Avg**
Unsup	**VecMap**	46.9	51.5	31.2	2.4	33.0
	MUSE	10.9	49.7	13.0	4.7	19.6
Ident	**VecMap**	47.1	51.9	50.3	26.5	44.0
	MUSE	47.7	49.8	46.8	32.4	44.2
8K	**VecMap**	47.4	51.8	49.5	30.3	44.8
	MUSE	47.6	49.3	48.6	42.2	46.9
	Meemi$_{VM}$	50.1	**53.6**	**53.8**	43.1	50.2
	Meemi$_{MS}$	**50.4**	52.5	52.0	**46.6**	**50.4**

Table 3: Spearman correlation performance of various cross-lingual word embedding models in the cross-lingual word similarity task.

better with downstream performance than other intrinsic benchmarks (Bakarov et al., 2018). The results are reported in terms of the Pearson and Spearman correlation with respect to human similarity judgments. The cross-lingual word similarity results for all the systems are displayed in Table 3. The languages available for this dataset are English, Spanish, Italian, German and Farsi, hence Finnish and Russian were not evaluated in this task.

5.3. Cross-lingual natural language inference

The task of natural language inference (NLI) consists in detecting entailment, contradiction and neutral relations between pairs of sentences. We test a zero-shot cross-lingual transfer setting where a system is trained with English corpora and is then evaluated on a different language. It is important to highlight that in this evaluation our main aim is to compare the quality of the cross-lingual word embeddings, and not to develop a state-of-the-art NLI system. Therefore, since this is a downstream task evaluated at the sentence level (and not at the word level as in dictionary induction and semantic word similarity), we develop a simple bag-of-words approach where a sentence embedding is obtained by word vector averaging. We then train a linear classifier[10] to obtain the predicted label for each pair of sentences: entailment, contradiction or neutral. We use the full MultiNLI (Williams et al., 2018) English corpus for training and the Spanish, German and Russian test sets from XNLI (Conneau et al., 2018b) for testing. Accuracy results are shown in Table 4.[11]

Wikipedia					
Sup.	**Model**	**EN-ES**	**EN-DE**	**EN-RU**	**Avg**
Unsup	**VecMap**	**49.6**	46.3	**34.1**	43.3
	MUSE	48.4	47.4	33.3	43.0
Ident	**VecMap**	43.0	42.9	33.2	39.7
	MUSE	39.5	35.8	33.3	36.2
8K	**VecMap**	49.2	46.7	33.4	43.1
	MUSE	47.7	47.1	33.1	42.6
	Meemi$_{VM}$	49.5	**47.6**	33.8	**43.6**
	Meemi$_{MS}$	44.2	46.7	33.3	41.4
Web corpora					
Sup.	**Model**	**EN-ES**	**EN-DE**	**EN-RU**	**Avg**
Unsup	**VecMap**	**48.5**	47.9	33.4	43.3
	MUSE	47.7	47.1	33.6	42.8
Ident	**VecMap**	45.5	44.4	33.4	41.1
	MUSE	35.2	36.6	33.3	35.0
8K	**VecMap**	48.4	47.5	33.2	43.0
	MUSE	47.3	**48.6**	33.1	43.0
	Meemi$_{VM}$	47.8	**48.6**	**33.8**	**43.4**
	Meemi$_{MS}$	47.3	48.2	33.2	42.9

Table 4: Accuracy in the cross-lingual natural language inference task (XNLI) using different cross-lingual word embedding models.

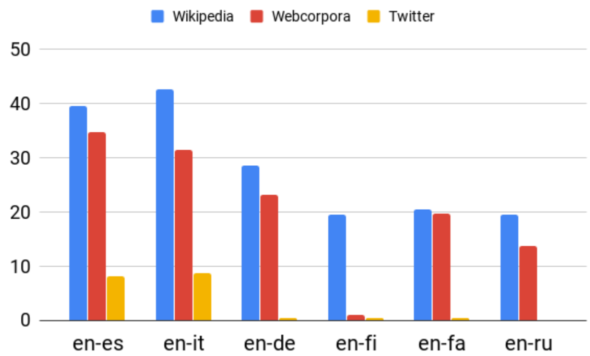

Figure 1: P@1 performance of the unsupervised version of `VecMap` on dictionary induction across corpus types and language pairs.

6. Analysis

Supervision signals. Unsurprisingly, the best alignments of monolingual spaces tend to be obtained with the largest bilingual dictionaries. The unsupervised variants of `VecMap` (see Figure 1) and `MUSE` attain competitive performance in most cases, especially for comparable corpora where alignments are easier to obtain. However, they struggle in the case of noisy social media corpora and unrelated languages (e.g. both `VecMap` and `MUSE` obtain inferior results, close to 0, on both Finnish and Farsi), which challenges conclusions from previous work (Conneau et al., 2018a; Artetxe et al., 2018b). Overall, the results obtained when using social media are clearly inferior, suggesting that

[10]The codebase for these experiments is that of SentEval (Conneau and Kiela, 2018)

[11]For this task we focused on the better performing embeddings learned from Wikipedia and web corpora.

there is still room for improvement when it comes to dealing with noisy corpora, regardless of the supervision.

VecMap vs. MUSE. One of the main differences between these two models relates to their robustness. The results of VecMap are largely stable across the different types of the supervision. In fact, the best performance for Spanish and Russian on the XNLI task is even obtained in its unsupervised mode. In contrast, MUSE does not perform well with small dictionaries. Figure 2 illustrates this trend. In addition, MUSE also suffers from some stability issues, as it does not always converge to the optimal solution, which confirms findings from previous work (Artetxe et al., 2018b; Søgaard et al., 2018; Hartmann et al., 2018).[12] In terms of overall results, when given a sufficiently large dictionary training data, the performance of both methods is comparable, which is perhaps unsurprising as they both rely on the solution of the orthogonal Procrustes problem to learn the transformation between the monolingual spaces.

Impact of corpora. As can be observed throughout all the experiments, the more comparable and less noisy the monolingual data is, the better the bilingual alignments. For instance, VecMap goes from an average of 31.2% in P@1 on Wikipedia down to 4.3% on social media, considering all language pairs. In word similarity, we observe an analogous performance drop, from 68.4% to 44.8% in Spearman correlation. Additionally, in Figure 1 we can observe the negative influence of noisy corpora and distant languages on the performance of the unsupervised version of VecMap on dictionary induction. In terms of error analysis, unsurprisingly we find that the low performance of the models trained on Twitter data is largely due to the noise and the informal nature of the conversation topics. For instance, for the word *discover*, instead of *descubren* (one of the correct Spanish translations obtained by the models trained on Wikipedia), the translation given by VecMap corresponds to a misspelling of the correct translation: *descubr*. As another example, *timeline* is not translated to *cronología* in Spanish, but to *instas*, which refers to the social network Instagram. This is clearly due to the specific use of the word *timeline* on Twitter.

Distant languages. As expected, the more different the languages are, the harder it is to obtain a reliable alignment of the monolingual spaces. This is particularly noticeable in the case of Farsi, Russian and Finnish (and German to a lesser extent). For instance, in the bilingual dictionary task, while most models are over 30.0% in P@1 (excluding social media text which causes performance drops in all languages), for Finnish, Farsi and Russian the results are below 20% in most cases. A similar tendency can be observed for Farsi on the word similarity task, where the differences are even more pronounced. In addition to its idiosyncrasies (Farsi is considered agglutinative and has a noun compounding formation similar to German), the fact that it uses a different alphabet may explain this large performance gap, noting that FastText takes subword units into account. Finally, while the poor performance could be partially explained by the small size of the monolingual training corpora for some languages, it is interesting to see notable performance differences in cases where a distant language has similar or even greater amounts of training data available; e.g., Italian and Russian on Wikipedia, or Italian, Finnish or Russian on web data.

Distant supervision. As far as the synthetic dictionary of identical words is concerned, MUSE seems to have more difficulties coping with its noisy nature than VecMap (obtaining an average of 16.9% versus 23.6% in P@1 overall in dictionary induction in the Wikipedia and web corpora domains). In fact, using MUSE in its unsupervised setting or with a small dictionary generally provides better results. However, on social media using the dictionary of identical words appears to help MUSE considerably in the word similarity task compared to the unsupervised setting, going from 19.6% to 44.2% on average in Spearman correlation overall. This can be attributed to the multilinguality of social media data, where phenomena like code-switching often occur. On the other hand, the consistency of the VecMap semi-supervised algorithm is highlighted again, as using the identical dictionary in this case yields similar results to using external bilingual dictionaries.

Postprocessing. As explained in Section 3.2, for our analysis we experimented with Meemi (Doval et al., 2018), a recent postprocessing technique which can be applied to any cross-lingual embedding space. There are two main conclusions regarding this technique. First, a clean and relatively big bilingual dictionary is needed in order to get improvements over the base methods VecMap and MUSE (for instance, $+1.2\%$ P@1 and $+3.1\%$ Spearman correlation scores on social media on average, using the 8K dictionary), with the performance otherwise ending up significantly lower. In general, the best overall results are achieved when using this postprocessing technique in combination with the largest dictionary (i.e., 8K pairs). Table 5 shows the performance gains or drops by using Meemi in the cross-lingual word similarity task, clearly showing the need for a reasonably large dictionary. This performance variability depending on the size of the dictionary was not addressed in the original paper. Second, Meemi appears to be particularly useful when the monolingual corpora are not comparable, as shown by the larger improvements attained on web-based data.

Evaluation tasks. The performance variability in bilingual dictionary induction, cross-lingual word similarity and cross-lingual inference seems to be very similar across the board, with the main difference being the lower variability in results in the cross-lingual NLI task (which can be expected given that it is a downstream task where additional factors are also involved). The factors with the greatest impact on performance, namely monolingual corpora and language pairs, are clearly reflected in both cases, with analogous drops when going from training on Wikipedia to social media, and also when testing on Finnish, Farsi or Russian. To test our intuition, we computed Pearson correlation values from all overlapping results between task pairs. In this case, similarity and dictionary induction attain the highest correlation ($r = 0.78$), with cross-lingual NLI and dictionary induction also attaining a high correlation

[12]This feature was not explicitly tested in this work, as in our experiments models were run until convergence.

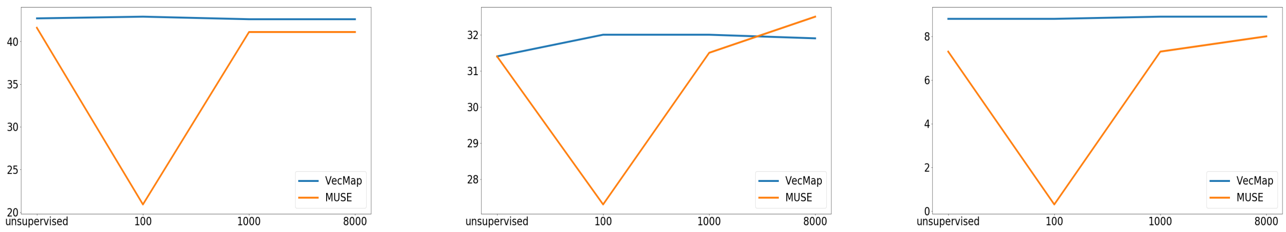

Figure 2: Comparison between the dictionary induction performance ($P@1$) of `VecMap` (blue) and `MUSE` (red) in English-Italian on Wikipedia (left), web corpora (middle) and social media text (right). The figure clearly shows how `VecMap` produces similar results irrespective of the seed supervision, while the results of `MUSE` fluctuate depending on the size of the seed dictionary (with its unsupervised variant being better than using a small dictionary).

		100	**1K**	**8K**
VecMap	**EN-ES**	-70.4	-2.0	**+0.3**
	EN-IT	-68.8	-1.0	**+1.2**
	EN-DE	-70.5	-0.3	**+1.4**
	EN-FA	-30.5	-32.8	**+5.5**
MUSE	**EN-ES**	-71.5	-1.8	**+1.7**
	EN-IT	-65.1	-0.8	**+0.7**
	EN-DE	-68.8	-0.1	**+1.9**
	EN-FA	-7.4	-22.3	**+9.1**

Table 5: Absolute improvement (in percentage points) by applying the postprocessing (`Meemi`) over the two base models VecMap and MUSE on the cross-lingual word similarity task using web corpora.

score ($r = 0.73$). The lowest correlation score corresponds to cross-lingual similarity and NLI, with a lower figure of $r = 0.28$. Despite being positive, this relatively low correlation may suggest that dictionary induction would be a better proxy to test cross-lingual embedding performance in downstream tasks. We should note, however, that these correlation figures are only indicative and particular to the methods tested in our analysis and therefore should not be taken as the global correlation between tasks.

7. Conclusions

We have presented an extensive evaluation of state-of-the-art cross-lingual embedding models in a wide variety of experimental settings. The variables explored in this paper were: the choice of training corpus, the type of supervision signal (including different types of bilingual dictionaries), and the language pairs considered. Likewise, the evaluation procedure included two standard benchmarks for cross-lingual embedding evaluation, namely bilingual dictionary induction and cross-lingual word similarity, as well as cross-lingual natural language inference as an extrinsic task. The set of languages considered included not only related languages such as English, Spanish, Italian and German, but also languages from different families such as Finnish, Farsi and Russian.

Our analysis highlights a particularly marked variability in the performance of the considered methods concerning (1) the monolingual training corpora used (e.g., between comparable corpora such as Wikipedia and non-comparable or noisy user-generated corpora) and (2) language pairs (distant language pairs still constitute a major challenge). We may also conclude that bilingual supervision signals constitute a key component for most models in non-ideal settings (i.e., non-comparable corpora or distant languages). In general, our analysis and the results show that supervised cross-lingual word embedding learning is more robust than purely unsupervised cross-lingual learning, challenging claims from previous works on this regard (Conneau et al., 2018a; Artetxe et al., 2018b; Chen and Cardie, 2018; Hoshen and Wolf, 2018; Xu et al., 2018) and in line with a concurring analysis showing a similar trend (Vulić et al., 2019).

As future work, it would be interesting to analyze multilingual embeddings that involve more than two languages, along the lines of recent multilingual approaches (Chen and Cardie, 2018; Heyman et al., 2019; Doval et al., 2019).

8. Acknowledgments

Steven Schockaert is supported by ERC Starting Grant 637277. Yerai Doval has been supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the ANSWER-ASAP project (TIN2017-85160-C2-2-R); by the Spanish State Secretariat for Research, Development and Innovation (which belongs to MINECO) and the European Social Fund (ESF) through a FPI fellowship (BES-2015-073768) associated to TELEPARES project (FFI2014-51978-C2-1-R); and by the Xunta de Galicia through TELGALICIA research network (ED431D 2017/12).

Appendix: Additional Experimental Results

This appendix contains additional experimental results not included in the main body of the paper. In particular, it contains supplementary results for the dictionary induction (Table 6) and cross-lingual word similarity (Table 7) tasks, using all sources of supervision: no supervision, dictionary of identical words, and dictionaries containing 100, 1K and 8K translation pairs. The methods included in these tables are explained in the paper (Section 3).

`Wikipedia`

Model	Supervision	English-Spanish			English-Italian			English-German			English-Finnish			English-Farsi		
		$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$
`VecMap`	**8K**	**39.6**	66.2	72.3	42.6	65.9	71.8	28.6	**48.3**	**54.8**	22.4	44.5	52.5	22.8	39.7	46.2
	1K	**39.6**	66.2	72.3	42.6	65.7	71.6	**28.7**	**48.3**	54.7	22.2	43.9	51.7	23.2	40.2	46.1
	100	**39.6**	66.2	72.4	**42.9**	65.7	71.6	28.6	**48.3**	**54.8**	21.6	43.4	51.7	22.7	**40.6**	**46.4**
	Identical	39.5	66.0	72.4	42.7	65.8	71.7	28.6	**48.3**	54.7	21.6	43.7	51.6	**23.4**	40.3	46.1
	Unsupervised	**39.6**	66.1	72.3	42.7	65.7	71.6	28.6	**48.3**	**54.8**	19.6	40.4	48.3	20.5	37.0	42.8
`MUSE`	**8K**	39.1	65.4	72.3	41.1	63.3	70.1	27.6	45.9	53.2	19.5	40.4	49.5	19.7	35.4	42
	1K	39.2	65.4	72.1	41.1	63.3	70.1	27.6	46.0	53.1	18.1	36.8	44.9	19.8	35.3	41.5
	100	24.8	47.5	54.6	20.9	39.2	48.1	0.8	3.4	5.2	0.3	1.3	2.2	6.2	16.1	22.8
	Identical	35.9	60.6	67.3	37.8	60.4	68.5	24.8	41.9	49.5	13.4	25.5	32	6.7	16.6	21.3
	Unsupervised	39.3	64.7	71.3	41.6	63.2	69.9	28.3	46.5	53.3	0.0	0.0	0.0	0.0	0.0	0.0
Meemi$_{\text{VM}}$	**8K**	39.3	**67.4**	**73.7**	41.6	66.5	72.5	28	47.8	**54.8**	**23.8**	**48.7**	**57.0**	0.0	0.0	0.0
	1K	35.5	63.7	69.4	38.6	64.0	70.1	23.1	42.5	49.9	17.8	40.1	48.6	0.0	0.0	0.0
	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Identical	38.7	63.7	70.1	40.6	64.1	70.2	27.5	46.6	53.1	19.3	37.7	45.5	7.1	14.3	17.9
Meemi$_{\text{MS}}$	**8K**	39.3	**67.4**	**73.7**	41.3	**66.8**	**72.8**	27.1	46.3	53.9	21.7	45.0	53.6	0.0	0.0	0.0
	1K	35.4	63.1	69.3	38.2	63.6	70.2	22.4	40.4	47.9	14.7	33.6	41.8	0.0	0.0	0.0
	100	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Identical	35.4	58.9	65.4	37.0	59.0	65.9	24.0	40.0	47.0	13.0	25.5	32.0	2.5	6.2	8.3

`Web corpora`

Model	Supervision	English-Spanish			English-Italian			English-German			English-Finnish			English-Farsi		
		$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$
`VecMap`	**8K**	34.6	60.6	66.9	31.9	54.2	60.4	23.1	42.7	50.5	18.9	40.9	48.8	19.6	**35.8**	**41.4**
	1K	34.6	60.5	67.0	32.0	54.0	60.5	23.1	42.7	50.3	19.4	42.0	49.4	19.5	35.6	41.2
	100	**38.5**	61.2	67.5	32.0	54.2	60.5	23.0	43.0	50.2	19.3	41.6	49.6	19.7	35.5	41.3
	Identical	34.7	60.4	67.0	31.4	54.0	60.7	23.1	42.9	50.5	18.6	41.6	49.3	**20.0**	35.3	40.3
	Unsupervised	34.8	60.6	67.0	31.4	53.7	60.7	23.2	42.7	50.2	0.0	0.0	0.0	19.7	34.6	40.4
`MUSE`	**8K**	32.5	58.2	65.9	32.5	56.0	63.2	22.4	40.9	48.9	20.0	40.1	48.3	17.4	31.6	37.6
	1K	32.9	56.8	64.2	31.5	52.7	60.6	22.1	41.0	48.2	18.4	39.1	47.7	16.6	31.1	36.4
	100	32.3	56.1	63.9	27.3	48.0	55.3	17.8	35.0	41.6	2.7	7.9	11.1	0.0	0.5	0.7
	Identical	26.1	46.7	53.8	24.7	45.1	52.4	17.4	32.8	40.5	12.6	26.0	33.8	3.0	8.3	5.8
	Unsupervised	31.4	51.2	57.7	31.4	51.2	57.7	20.8	38.7	46.6	0.0	0.0	0.0	18.1	32.8	37.8
Meemi$_{\text{VM}}$	**8K**	34.5	**61.6**	67.9	33.6	58.3	**65.6**	**23.7**	**45.4**	**53.2**	22.3	**46.7**	**55.0**	0.0	0.0	0.0
	1K	30.2	55.0	62.7	30.7	54.0	61.1	19.4	38.9	45.9	18.2	39.9	48.1	0.0	0.0	0.0
	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Identical	34.1	58.3	64.8	31.6	54.6	62.5	22.5	42.0	49.0	21.1	43.2	51.3	11.2	23.9	28.6
Meemi$_{\text{MS}}$	**8K**	33.9	60.7	**68.4**	**33.8**	**58.4**	**65.6**	**23.7**	45.3	52.3	**23.0**	46.1	54.0	0.0	0.0	0.0
	1K	29.1	54.6	62.3	29.9	52.7	60.3	18.3	37.0	44.1	17.2	37.4	45.6	0.0	0.0	0.0
	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Identical	24.3	43.0	49.8	21.8	41.1	48.9	16.4	30.7	37.3	13.1	26.1	33.9	2.0	4.2	5.8

`Social media`

Model	Supervision	English-Spanish			English-Italian			English-German			English-Finnish			English-Farsi		
		$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$	$P@1$	$P@5$	$P@10$
`VecMap`	**8K**	8.7	16.6	21.6	8.9	17.3	22.4	3.2	6.8	9.5	0.2	0.8	1.2	0.4	1.6	2.0
	1K	8.3	17.0	21.3	8.9	17.5	22.0	2.9	6.5	9.3	0.0	0.4	1.1	0.3	1.0	1.4
	100	7.9	15.9	20.2	8.8	17.6	22.3	2.8	6.0	8.6	0.0	0.0	0.1	0.1	0.3	0.4
	Identical	8.5	16.9	21.6	9.1	16.8	21.8	2.6	6.7	9.6	0.0	0.0	0.0	0.2	0.5	1.1
	Unsupervised	8.1	16.4	20.4	8.8	17.0	22.3	0.1	0.4	0.5	0.0	0.0	0.0	0.0	0.0	0.0
`MUSE`	**8K**	8.1	17.6	22.7	8.0	16.4	21.1	2.2	6.0	8.4	0.6	2.2	3.2	**1.2**	**4.5**	**6.3**
	1K	7.2	15.9	20.5	7.3	14.6	18.4	0.9	3.0	4.5	0.6	1.5	2.1	0.9	2.1	3.4
	100	0.4	1.1	1.9	0.3	1.1	1.8	0.1	0.3	0.6	0.0	0.2	0.4	0.1	0.3	0.4
	Identical	2.5	5.2	7.1	3.9	10.1	13.7	1.1	2.6	3.7	0.1	0.1	0.2	0.1	0.3	0.8
	Unsupervised	0.0	0.0	0.0	7.3	14.5	18.3	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1
Meemi$_{\text{VM}}$	**8K**	**9.8**	**21.3**	**26.9**	**10.6**	**20.0**	**25.6**	**3.7**	**9.6**	**13.2**	1.3	3.6	5.5	0.0	0.1	0.1
	1K	8.3	17.7	22.6	8.6	18.2	23.6	3.0	7.5	10.6	0.5	2.4	3.7	0.0	0.0	0
	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Identical	3.8	9.1	11.8	6.6	14.2	18.2	2.0	4.1	5.9	0.0	0.1	0.2	0.1	0.3	0.5
Meemi$_{\text{MS}}$	**8K**	9.5	20.5	26.3	9.5	19.1	24.5	3.0	7.6	11.1	**1.5**	**4.3**	**6.4**	0.0	0.1	0.2
	1K	7.6	16.9	22.3	7.8	15.9	21	1.7	4.1	6.2	0.8	2.3	3.7	0	0.0	0.0
	100	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Identical	2.9	5.3	6.7	3.5	9.9	13.2	1.2	2.8	3.6	0.2	0.3	0.3	0.0	0.1	0.2

Table 6: Bilingual dictionary induction results in the test sets of Conneau et al. (2018a).

Wikipedia

Model	Dictionary	English-Spanish		English-Italian		English-German		English-Farsi	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
VecMap	8K	72.1	71.8	71.2	70.6	70.0	69.3	63.7	61.7
	1K	72.1	71.8	71.2	70.6	70.0	69.3	63.9	61.9
	100	72.1	71.8	71.2	70.6	70.0	69.3	63.9	62.0
	Identical	72.1	71.8	71.2	70.6	70.0	69.3	63.8	61.9
	Unsupervised	72.1	71.8	71.2	70.6	70.0	69.3	63.4	61.3
MUSE	8K	72.0	72.6	70.7	70.9	68.8	68.9	59.2	58.7
	1K	71.9	72.4	70.6	70.7	68.6	68.7	58.7	58.4
	100	65.1	66.3	63.0	63.6	44.7	49.9	47.7	52.1
	Identical	71.0	71.9	69.9	70.5	68.1	68.4	47.9	51.3
	Unsupervised	72.2	72.6	71.0	71.2	68.7	68.9	8.0	6.5
Meemi$_{VM}$	8K	72.5	71.9	71.8	70.9	**70.9**	**70.3**	**65.1**	**63.4**
	1K	70.1	69.6	69.7	69.0	67.6	66.8	5.5	5.8
	100	0.0	0.0	5.1	5.0	4.1	3.3	6.8	6.5
	Identical	71.0	70.4	69.4	68.7	69.3	68.6	56.1	54.1
Meemi$_{MS}$	8K	**73.1**	**72.9**	**72.4**	**71.9**	70.7	70.1	64.1	62.0
	1K	70.4	70.3	69.6	69.6	66.7	66.5	5.0	4.2
	100	2.7	1.4	0.0	0.2	6.3	6.0	0.0	0.0
	Identical	70.6	70.9	68.7	68.8	68.0	67.7	46.6	48.2

Web corpora

Model	Dictionary	English-Spanish		English-Italian		English-German		English-Farsi	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
VecMap	8K	71.0	70.6	69.2	68.8	70.9	70.4	35.9	33.5
	1K	71.0	70.6	69.3	68.8	70.9	70.4	35.9	33.5
	100	71.0	70.4	69.2	68.8	71.0	70.5	35.9	33.5
	Identical	71.0	70.6	69.3	68.8	70.9	70.4	35.9	33.0
	Unsupervised	71.1	70.5	69.2	68.8	70.9	70.4	35.7	33.4
MUSE	8K	71.9	71.9	70.4	70.4	70.5	70.2	29.7	23.9
	1K	71.6	71.5	69.5	69.4	70.3	70.0	28.3	22.3
	100	71.7	71.6	67.4	67.4	68.5	68.8	6.3	7.4
	Identical	69.9	70.1	67.3	67.5	70.1	69.7	17.5	14.5
	Unsupervised	71.7	71.6	69.4	69.4	70.3	70.0	29.6	23.8
Meemi$_{VM}$	8K	71.5	70.9	70.4	70.0	72.3	71.8	**40.2**	**39.0**
	1K	69.1	68.6	68.2	67.8	70.9	70.2	1.2	0.7
	100	0.0	0.0	0.0	0.0	0.0	0.0	3.2	3.0
	Identical	70.1	69.5	69.2	68.4	71.2	70.6	31.6	28.5
Meemi$_{MS}$	8K	**72.5**	**72.3**	**71.5**	**71.1**	**72.5**	**72.1**	36.4	33.0
	1K	70.0	69.7	68.9	68.6	70.3	69.9	0.0	0.0
	100	1.7	0.1	1.6	2.3	0.0	0.0	0.3	0.0
	Identical	69.2	69.1	67.4	66.9	70.1	69.4	17.3	14.5

Social media

Model	Dictionary	English-Spanish		English-Italian		English-German		English-Farsi	
		Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
VecMap	8K	48.5	47.4	53.9	51.8	51.3	49.5	31.1	30.3
	1K	48.7	47.7	53.7	51.6	51.7	50.2	30.3	29.6
	100	49.8	48.9	54.0	51.7	51.0	49.5	25.7	25.4
	Identical	48.4	47.1	54.2	51.9	51.8	50.3	27.8	26.5
	Unsupervised	48.0	46.9	53.8	51.5	30.1	31.2	4.2	2.4
MUSE	8K	48.8	47.6	51.0	49.3	48.5	48.6	43.3	42.2
	1K	46.6	45.5	49.7	47.8	44.8	45.7	38.7	38.9
	100	35.8	36.9	29.6	31.3	30.7	34.0	20.8	21.3
	Identical	48.1	47.7	50.1	49.8	45.6	46.8	30.5	32.4
	Unsupervised	9.9	10.9	50.7	49.7	12.4	13.0	6.9	4.7
Meemi$_{VM}$	8K	51.2	50.1	**56.1**	**53.6**	**55.0**	**53.8**	45.2	43.1
	1K	49.7	48.6	55.4	52.8	52.8	51.3	3.8	3.8
	100	2.4	2.0	2.1	2.8	0.0	0.0	5.6	5.1
	Identical	51.7	50.1	56.2	53.4	52.7	51.3	29.9	28.6
Meemi$_{MS}$	8K	**51.8**	**50.4**	54.8	52.5	53.1	52.0	**48.9**	**46.6**
	1K	49.5	47.9	53.1	50.7	48.4	47.2	0.0	0.0
	100	5.2	5.4	0.0	0.0	6.2	6.8	0.0	0.0
	Identical	49.6	48.2	53.1	51.4	48.4	47.7	30.5	30.6

Table 7: Cross-lingual word similarity results in the SemEval-17 dataset (Camacho-Collados et al., 2017).

9. Bibliographical References

Ahmad, W. U., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., and Peng, N. (2019). On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of NAACL*.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria.

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., and Oroumchian, F. (2009). Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5):382–387.

Artetxe, M., Labaka, G., and Agirre, E. (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, July. Association for Computational Linguistics.

Artetxe, M., Labaka, G., and Agirre, E. (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)*.

Artetxe, M., Labaka, G., and Agirre, E. (2018b). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

Bakarov, A., Suvorov, R., and Sochenkov, I. (2018). The limitations of cross-language word embeddings evaluation. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 94–100.

Barone, A. V. M. (2016). Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126.

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

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5(1):135–146.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (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.

Camacho-Collados, J., Doval, Y., Martínez-Cámara, E., Espinosa-Anke, L., Barbieri, F., and Schockaert, S. (2020). Learning cross-lingual embeddings from Twitter via distant supervision. In *Proceedings of ICWSM*.

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings. *http://crscardellino.me/SBWCE/*, March.

Chandar, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., and Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861.

Chen, X. and Cardie, C. (2018). Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270.

Conneau, A. and Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2018a). Word translation without parallel data. In *Proceedings of ICLR*.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018b). Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Doval, Y., Camacho-Collados, J., Espinosa-Anke, L., and Schockaert, S. (2018). Improving cross-lingual word embeddings by meeting in the middle. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 294–304, Brussels, Belgium. Association for Computational Linguistics.

Doval, Y., Camacho-Collados, J., Espinosa-Anke, L., and Schockaert, S. (2019). Meemi: A simple method for post-processing and integrating cross-lingual word embeddings. *arXiv preprint arXiv:1910.07221*.

Glavas, G., Litschko, R., Ruder, S., and Vulic, I. (2019). How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. *CoRR*, abs/1902.00508.

Godin, F., Vandersmissen, B., De Neve, W., and Van de Walle, R. (2015). Named Entity Recognition for Twitter Microposts using Distributed Word Representations. In *Proceedings of ACL Workshop on Noisy User-generated Text*, pages 146–153.

Han, L., Kashyap, A., Finin, T., Mayfield, J., and Weese, J. (2013). UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages 44–52.

Hartmann, M., Kementchedjhieva, Y., and Søgaard, A. (2018). Why is unsupervised alignment of english embeddings from different algorithms so hard? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 582–586.

Heyman, G., Verreet, B., Vulić, I., and Moens, M.-F. (2019). Learning unsupervised multilingual word embeddings with incremental multilingual hubs. 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)*, pages 1890–1902, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hoshen, Y. and Wolf, L. (2018). Non-adversarial unsupervised word translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 469–478, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kementchedjhieva, Y., Ruder, S., Cotterell, R., and Søgaard, A. (2018). Generalizing procrustes analysis for better bilingual dictionary induction. In *Proceedings of the Conference on Computational Natural Language Learning*.

Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474.

Luong, T., Pham, H., and Manning, C. D. (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.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Ruder, S., Vulić, I., and Søgaard, A. (2018). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*.

Smith, S. L., Turban, D. H., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *Proceedings of ICLR*.

Søgaard, A., Ruder, S., and Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788. Association for Computational Linguistics.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of ACL*, pages 1555–1565.

Upadhyay, S., Faruqui, M., Dyer, C., and Roth, D. (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.

Vulić, I. and Korhonen, A. (2016). On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 247–257.

Vulić, I. and Moens, M.-F. (2015). 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.

Vulić, I., Glavaš, G., Reichart, R., and Korhonen, A. (2019). Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4398–4409.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Xu, R., Yang, Y., Otani, N., and Wu, Y. (2018). Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474. Association for Computational Linguistics.

Yang, X., Macdonald, C., and Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2):183–207, Jun.