Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Learning Cross-Lingual Word Embeddings from Twitter via Distant Supervision

Jose Camacho-Collados,[◇*] **Yerai Doval,**[♡*] **Eugenio Martínez-Cámara,**[♠]
Luis Espinosa-Anke,[◇] **Francesco Barbieri,**[♣] **Steven Schockaert**[◇]
[◇]School of Computer Science and Informatics, Cardiff University, United Kingdom [♣]Snap Inc., USA
[♡]Escola Superior de Enxeñaría Informática, Universidade de Vigo, Spain [♠]DaSCI Institute, University of Granada, Spain
[◇]{camachocolladosj, espinosa-ankel, schockaerts1}@cardiff.ac.uk, [♣]fbarbieri@snap.com
[♡]yerai.doval@uvigo.es, [♠]emcamara@decsai.ugr.es

Abstract

Cross-lingual embeddings represent the meaning of words from different languages in the same vector space. Recent work has shown that it is possible to construct such representations by aligning independently learned monolingual embedding spaces, and that accurate alignments can be obtained even without external bilingual data. In this paper we explore a research direction that has been surprisingly neglected in the literature: leveraging noisy user-generated text to learn cross-lingual embeddings particularly tailored towards social media applications. While the noisiness and informal nature of the social media genre poses additional challenges to cross-lingual embedding methods, we find that it also provides key opportunities due to the abundance of code-switching and the existence of a shared vocabulary of emoji and named entities. Our contribution consists of a very simple post-processing step that exploits these phenomena to significantly improve the performance of state-of-the-art alignment methods.

1 Introduction

Twitter provides a wealth of uncurated text (Derczynski et al. 2013) and has been found to constitute a valuable source for developing natural language processing (NLP) systems in, for example, sentiment analysis (Martínez-Cámara et al. 2014), sarcasm detection (Felbo et al. 2017) or humour and irony modeling (Reyes, Rosso, and Buscaldi 2012). Given their abundance and multilingual nature, we argue that tweets are a powerful but surprisingly neglected source for learning cross-lingual vector representations of words (henceforth, cross-lingual embeddings).

Cross-lingual embeddings are the result of mapping two or more monolingual word embedding spaces into a shared vector space in which words and their translations are represented by similar vectors. Along with obvious applications in, for example, machine translation (Artetxe et al. 2018; Lample, Denoyer, and Ranzato 2018; Lample et al. 2018), cross-lingual embeddings also constitute a major step forward towards knowledge transfer between languages (Ruder, Vulić, and Søgaard 2019), usually having English as source or pivot. Several recent approaches have shown that accurate mappings can indeed be learned with minimal amounts of supervision, to the point that external bilingual data may no longer be needed (Conneau et al. 2018; Artetxe, Labaka, and Agirre 2018b; Xu et al. 2018). However, previous work has mostly focused on controlled or noise-free environments, reporting results from using clean and comparable corpora as source. In this paper we make a case for the potential (and discuss the limitations) of social media data for learning cross-lingual embeddings, thus parting ways with the traditional 'noise-free' setting explored in most recent literature.

In monolingual settings, it has already been shown that word embeddings trained on Twitter lead to increased performance in social media NLP tasks (Tang et al. 2014; Godin et al. 2015; Yang, Macdonald, and Ounis 2018). One of the main reasons is that such embeddings cover a much wider range of slang terms and neologisms, and therefore provide a more faithful snapshot of the particularities of the language used in social media. Twitter-specific cross-lingual embeddings can thus also be expected to provide solid grounds for cross-lingual social media NLP tasks. In this paper, we demonstrate that this is indeed the case for, specifically, word translation and cross-lingual sentiment analysis, where we use data for English to train classifiers for other languages.

Another crucial advantage of Twitter is that it is peppered with a significant number of tokens[1] that are shared across different languages. This is relevant, as previous work has demonstrated that the shared meaning of numerals can be exploited for effectively learning cross-lingual embeddings in a self-supervised fashion (Artetxe, Labaka, and Agirre 2017). For instance, we can assume that the embedding for '5' will embody similar properties in, e.g., English and Spanish. We can also assume that embeddings for emoji obtained from tweets in different languages will generally represent the same or a very similar meaning (Barbieri et al. 2016b). Finally, and most importantly, we also take advantage of the fact that many words in non-English tweets have an exact counterpart in English, which can be attributed to

[*]Authors marked with an asterisk contributed equally.

[1]We use *token* as umbrella term covering anything from a word to an emoji, or any other social media textual artifact.

code-switching and to the presence of *interlingual homographs*[2], including many named entities.

We exploit this vocabulary of shared tokens across tweets from different languages to implement a very simple post-processing technique, which maps identical tokens from different languages to the same vector in the cross-lingual embedding space. Clearly, it is overly simplistic to assume that two words from different languages have the same meaning simply because they are spelled in the same way, and even emoji sometimes have language-specific meanings. Surprisingly, however, we find that such a simple post-processing strategy nonetheless leads to substantial performance gains in the tasks of word translation and cross-lingual sentiment analysis.

Pre-trained monolingual and cross-lingual embedding models for all languages explored in this paper (i.e., English, Spanish, Italian, German and Farsi) are available at https://github.com/pedrada88/crossembeddings-twitter.[3]

2 Motivation

Learning unsupervised cross-lingual embeddings (i.e. without the need for any kind of external supervision) has become one of the most prominent tasks in NLP in recent years (Ruder, Vulić, and Søgaard 2019), both as a research challenge in itself and for its potential for transferring knowledge across languages. Such cross-lingual word embeddings have already proved effective in cross-lingual NLP tasks, typically after being trained on standard corpora such as Wikipedia (Artetxe, Labaka, and Agirre 2017; Conneau et al. 2018; Glavaš et al. 2019). However, cross-lingual resources specifically tailored to the colloquial nature of social media are scarce, which constitutes the main motivation for this paper. We argue that unsupervised cross-lingual embeddings are a highly appealing resource in this regard. They enable the development of multilingual technologies without requiring any form of manual supervision, which usually ranges from parallel data to word translations, both of which are generally lacking in the social media domain and especially in less-resourced languages. Crucially, this technology can have an impact in applications with direct societal implications, e.g., when natural disasters hit areas where people speak a low-resource language (Imran et al. 2018). Despite the success of unsupervised cross-lingual embeddings, their effectiveness in social media have remained untested so far. In this paper, therefore, we compile Twitter corpora in five different languages (English, Spanish, Italian, German and Farsi) and provide an extensive analysis of their resulting cross-lingual word vector mappings.

We found that the effectiveness of existing state-of-the-art cross-lingual embeddings methods in the social media setting is limited, but discovered that using *free* supervision such as identical tokens can be a promising workaround. After analyzing the nature of these identical tokens (see Section 4.1 and the ablation test in Section 6), we put forward a simple post-processing step which causes a sharp performance increase for all languages and tasks we considered, both intrinsic and extrinsic, over state-of-the-art unsupervised and weakly-supervised cross-lingual embedding models. The motivation for this post-processing technique (see Section 4.3) is to exploit some specific features of social media text, namely its code-switching nature and the occurrence of special interlingual tokens such as numerals and, especially, emoji. These features are not usually present in other different types of corpora but, as we will show, they turn out to be powerful bilingual signals in cross-lingual embedding learning.

3 Related Work

3.1 Cross-lingual word embeddings

Cross-lingual embeddings are becoming increasingly popular in NLP (Upadhyay et al. 2016; Ruder, Vulić, and Søgaard 2019), especially since the recent introduction of models requiring almost no supervision (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014; Xing et al. 2015; Smith et al. 2017; Artetxe, Labaka, and Agirre 2017; Doval et al. 2018). These models have shown to be highly competitive compared to fully supervised baselines (which are typically trained on parallel corpora).

Despite their effectiveness, these recent models still need some form of supervision signal, which often takes the form of a bilingual dictionary. This limitation motivated the emergence of fully unsupervised models, based on, among others, adversarial training (Zhang et al. 2017; Conneau et al. 2018; Xu et al. 2018; Chen and Cardie 2018). However, as shown by Søgaard, Ruder, and Vulić (2018), some of these fully unsupervised methods (e.g., Conneau et al. (2018)) may be brittle when dealing with different types of languages and corpora. In a parallel direction, Artetxe, Labaka, and Agirre (2018b) proposed an alternative unsupervised model for learning cross-lingual embeddings, based on a similarity-based dictionary initialization and a linear transformation. While this approach proved to be more robust, and can even surpass supervised models exploiting synthetic or external bilingual dictionaries (Mikolov, Le, and Sutskever 2013; Xing et al. 2015; Smith et al. 2017), they only considered standard corpora.

In this paper we evaluate some of the most prominent cross-lingual embedding models in the more challenging setting of social media. Our evaluation shows that unsupervised models often struggle with noisy user-generated text, and the resulting aligned spaces seem to perform poorly in standard evaluation benchmarks (both intrinsic and extrinsic).

3.2 Cross-lingual sentiment analysis

As with most NLP tasks, the availability of training data and linguistic resources for sentiment analysis (SA) is generally skewed towards English, which motivates the creation of cross-lingual SA systems. However, most existing work in cross-lingual SA is built upon (1) machine translation systems (Salameh, Mohammad, and Kiritchenko 2015; Zhou, Wan, and Xiao 2016); (2) parallel (Meng et al. 2012;

[2] Interlingual homographs can be defined as words written identically in two or more different languages.

[3] This repository will be updated with embeddings for additional languages (Finnish and Japanese are already available).

Chen et al. 2018) or comparable corpora (Rasooli et al. 2018); (3) synthetic corpora developed with documents written in the source and the target language (Vilares, Alonso, and Gómez-Rodríguez 2017); or (4) bilingual lexicons (Barnes, Klinger, and Schulte im Walde 2018). Consequently, all these works depend on the availability of annotated data or the quality of off-the-shelf machine translation systems, which are generally ill-suited for social media text. In contrast, the approach we consider in this paper effectively enables zero-shot cross-lingual transfer in sentiment analysis without the need for external bilingual resources.

4 Learning Cross-lingual Embeddings

Most approaches for learning cross-lingual embeddings without parallel corpora make use of standard pre-trained monolingual vectors. These are mapped onto a shared cross-lingual space, usually with the help of external bilingual dictionaries. As an alternative, in this paper we consider automatically acquired dictionaries. In Section 4.1, we discuss how these dictionaries can be constructed from Twitter data. These dictionaries will then be used as the supervision signal for well-known state-of-the-art methods, which are briefly recalled in Section 4.2. Finally, we introduce a simple post-processing step which drastically improves performance in different benchmarks (Section 4.3).

4.1 Automatic creation of a bilingual dictionary

There are two main approaches to automatic dictionary construction from monolingual corpora: by distant supervision or by relying on the distribution of monolingual embeddings. In our method, we will rely on distant supervision signals from Twitter. However, let us first briefly introduce the latter "fully unsupervised" methods.

Unsupervised (distributional). Approaches from this class construct a dictionary by exploiting the distribution of monolingual embeddings. There are two prominent methods that rely on this intuition: Artetxe, Labaka, and Agirre (2018b) exploit the structural similarity of monolingual embeddings, specifically, the fact that cross-lingual synonyms have close similarity distributions across different languages. Conneau et al. (2018), on the other hand, learn this initial bilingual dictionary through adversarial training.

Distant supervision (identical tokens). To construct a synthetic bilingual dictionary in an automatic fashion, we rely on the following intuition: whenever a token appears in both monolingual corpora, we assume it has the same meaning. In other words, our dictionary only contains trivial entries, where a word is equal to its (presumed) translation. These identical tokens can be split into the following three types:

(i) Numerals: Given their extensive usage, Arabic numerals constitute a ubiquitous cross-lingual distant supervision signal. They were first leveraged by Artetxe, Labaka, and Agirre (2017).

(ii) Emoji: Emoji are ideograms depicting people, objects and scenes (Cappallo, Mensink, and Snoek 2015), which co-exist with words in social media communication. While some emoji preserve cultural differences, they have been shown to share similar meaning across languages and countries (Barbieri et al. 2016b). One of their potential advantages with respect to numerals, in addition to their prevalence in social media, is their diversity, as there are emoji for a wide range of domains such as medicine (🏥), sports (🏀), business (💼) or geography (⛰). Emoticons such as smileys, e.g., `:-)`, provide a similar bilingual signal.

(iii) Shared words : English words are often used by non-English speakers in spontaneous communication in social media. This phenomenon is particularly common in languages that are related to or which share their alphabet with English, where vocabularies of shared words may arise due to the existence of interlingual homographs[4] or code-switching environments. Even in more distantly related languages, English words are used in the form of many borrowed and loan words, especially in digital communication.

4.2 Alignment strategies

Various methods have been proposed for aligning two monolingual embedding spaces. Two recent methods in particular have obtained outstanding results in both unsupervised and semi-supervised settings: MUSE (Conneau et al. 2018) and VecMap (Artetxe, Labaka, and Agirre 2018a). Recall that the seed supervision signal required for these methods comes in the form of a bilingual dictionary, which may be external or automatically generated. These two methods are similar in that they learn an orthogonal linear transformation which maps one monolingual embedding space into the other. In VecMap this is done using SVD, while MUSE uses Procrustes analysis. VecMap applies this approach in an iterative fashion, where at each step the previously used bilingual dictionary is extended based on the current alignment. It is also worth noting that after the initial orthogonal transformation, VecMap fine-tunes the resulting embeddings by giving more weight to highly correlated embedding components, improving its performance in word translation.

Finally, let us refer to Doval et al. (2018), who recently proposed a method which extends VecMap and MUSE with a post-processing step. This method consists in applying an additional linear transformation, learned by linear regression on the translation pairs from external bilingual dictionaries. In this way, cross-lingual synonyms are mapped to their corresponding average embedding. Note that this dictionary can again be obtained through distant supervision, although this was not explored in Doval et al. (2018) or its extension (Doval et al. 2019).

4.3 Averaging cross-lingual embeddings

We put forward a simple post-processing step inspired by Doval et al. (2018). However, in contrast to the latter

[4]Clearly, there are examples of words which are written in the same way in two languages, but which have a different meaning. For instance, the correct English translation of the Spanish word *sensible* is *sensitive*, not *sensible*. Nonetheless, such a naïve assumption proves to be indisputably helpful.

method, which modifies the vector representations of all words, we simply replace the representations of the words in our synthetic dictionary by the average of their initial vector and the initial vector of their presumed translation, leaving all other vectors unchanged. In our experimental results, we show that, surprisingly, this simple approach leads to substantially better results than those obtained by competing baselines. Our method crucially relies on the availability of a sufficiently large bilingual dictionary. In this regard, one of the main contributions of this paper is showing that suitable dictionaries can be obtained automatically from Twitter corpora.

In addition to this vanilla averaging method, we also consider a variant in which the average is weighted by frequency:

$$\vec{\mu}_{w_1,w_2} = \frac{f_1\vec{v}_1 + f_2\vec{v}_2}{f_1 + f_2} \tag{1}$$

where f_1 and f_2 are the number of occurrences of the tokens w_1 and w_2 in their corresponding monolingual corpora, and $\vec{v}_1$ and $\vec{v}_2$ represent the embeddings of w_1 and w_2 in the cross-lingual vector space.[5] The main intuition behind this alternative is that even when a word occurs in tweets from both languages, it may still be underrepresented in one of them. This would be the case, for instance, if in one of the languages the word were only used in a code-switching context, or simply because of it being less prominent due to cultural or geographical differences. For instance, the word *NFL*, which stands for *National Football League* in the United States is also used in Spain, but much less frequently. We can thus expect that its Spanish embedding will be less accurate than the English one. Therefore, in this case it would make sense to give more prominence to the English vector. We will use *Plain* and *Weighted* to refer to our standard and weighted averaging strategies respectively.

5 Evaluation

We analyze the performance of cross-lingual word embeddings in the context of Twitter corpora, focusing in particular on the effectiveness of our post-processing method. First, however, let us describe the setting for cross-lingual embedding training.

Corpus compilation We collected five monolingual Twitter corpora between October 2015 and July 2018.[6] These corpora were independently gathered using geolocalized tweets which were tagged with specific languages:[7] United States (English), Spain (Spanish), Italy (Italian), Germany (German) and Iran (Farsi). To encourage more tweet diversity, only a maximum of twenty tweets per user were retained. After preprocessing (tokenization and duplicate removal), the final corpora consisted of 21,461,242 tweets for English, 10,122,550 for Spanish, 4,546,508 for Italian, 7,905,827 for German and 3,724,602 for Farsi. All languages contained more than 60M tokens overall and more than 1M unique tokens, with English being the largest among the five languages considered. Table 1 summarizes the main statistics (overall number of tweets and tokens, and number of unique tokens) of all Twitter language-specific corpora used in our experiments.

Language	# Tweets	# Tokens	# Unique
English	21,461,241	294,276,603	5,499,846
Spanish	10,122,550	144,394,815	3,312,603
Italian	4,546,508	63,076,614	1,601,218
German	7,905,827	114,545,634	2,301,059
Farsi	3,724,602	90,288,567	1,038,666

Table 1: Number of tweets and tokens (overall and unique) per corpus.

Monolingual embeddings All comparison systems use the same monolingual embeddings as input. These embeddings were trained on the Twitter corpora described above using FastText (Bojanowski et al. 2017). FastText was chosen due to its handling of subword units, making it more robust to misspellings as compared to alternatives like Word2Vec (Mikolov et al. 2013) or GloVe (Pennington, Socher, and Manning 2014). The monolingual embeddings were trained with FastText's default hyperparameters, fixing the dimension size to 100.

Distant supervision As explained in Section 4.1, we automatically extracted bilingual dictionaries of identical tokens to be used as supervision for the cross-lingual models. This resulted in dictionaries of 122,469 word pairs for English-Spanish, 66,037 for English-Italian, 93,695 for English-German and 6,142 for English-Farsi.

Comparison systems We used VecMap (Artetxe, Labaka, and Agirre 2018b) and MUSE (Conneau et al. 2018) to obtain the initial cross-lingual word embeddings, experimenting with their (semi-)supervised and unsupervised settings. In the former case, the supervision came from our synthetic dictionaries of identical tokens. The semi-supervised version of VecMap (Artetxe, Labaka, and Agirre 2018a) is used as our base model on which we evaluate two post-processing techniques: Meemi (Doval et al. 2018) and our proposed averaging strategy. Finally, for the sake of completeness, as baseline we also include a version of VecMap which uses external bilingual data in the form of a bilingual dictionary as external supervision. Following Artetxe, Labaka, and Agirre (2017), who showed how their semi-supervised model could work with as few as 25 word pairs as supervision, we perform an experiment with 100 word pairs from an external dictionary as supervision.[8] For all the baseline systems we

[5] f_1 and f_2 may be either absolute or relative frequencies. In our case we did not find noticeable differences given that all monolingual corpora were of comparable size.

[6] While all tweets were downloaded between these two months for all corpora, the Farsi set contains tweets from a shorter period.

[7] We relied on Twitter language identification procedure for gathering the language-specific tweets.

[8] These 100 pairs sampled from the Europarl training dictionaries provided by Dinu, Lazaridou, and Baroni (2015) and Artetxe, Labaka, and Agirre (2017).

Supervision	Model	EN-ES						EN-IT						EN-DE						EN-FA		
		Europarl			Facebook			Europarl			Facebook			Europarl			Facebook			Facebook		
		P1	P5	P10	P1	P5	P10	P1	P5	P10	P1	P5	P10	P1	P5	P10	P1	P5	P10	P1	P5	P10
Unsupervised	MUSE	8.3	14.1	17.8	6.8	15.3	19.0	8.7	13.7	17.1	6.7	14.4	18.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	VecMap	9.5	17.0	19.4	8.1	16.4	20.4	9.2	16.9	20.9	8.8	17.0	22.3	0.2	0.4	0.9	0.1	0.4	0.5	0.0	0.0	0.0
Distant	MUSE	2.7	5.4	7.2	2.6	5.3	7.0	3.6	9.1	12.4	4.0	10.0	13.6	1.3	2.7	3.5	1.4	3.0	4.1	0.1	0.4	0.9
	VecMap	10.1	17.8	21.2	8.5	16.9	21.6	9.6	17.0	21.0	9.1	16.8	21.8	3.4	6.9	9.7	2.6	6.7	9.6	0.2	0.5	1.1
	Meemi	3.7	9.7	12.4	3.9	9.1	12.0	7.9	16.2	19.1	6.7	14.0	18.2	1.9	4.5	6.8	2.0	4.1	5.9	0.1	0.3	0.4
	Plain	**21.1**	25.9	29.6	**16.7**	20.2	23.2	**20.3**	28.3	33.1	**22.4**	**31.3**	**35.7**	24.0	**27.3**	**29.4**	**16.2**	**19.4**	**21.3**	**1.3**	**1.7**	**2.0**
	Weighted	20.8	**28.6**	**33.7**	**16.7**	**22.8**	**28.5**	19.4	**28.8**	**33.7**	21.2	28.9	33.5	**24.3**	26.2	28.6	**16.2**	18.0	19.9	1.2	1.5	1.8
100 pairs	VecMap	10.0	17.0	20.5	7.9	15.9	20.2	9.2	17.1	20.9	8.8	17.6	22.3	3.2	7.6	9.7	0.0	0.0	0.1	0.1	0.3	0.4
	MUSE	0.6	1.7	2.5	0.4	1.1	1.9	0.6	1.6	2.3	0.3	1.1	1.8	0.3	0.5	0.5	0.1	0.3	0.6	0.1	0.3	0.4

Table 2: Word translation results on 4 target languages: Spanish (ES), Italian (IT), German (DE) and Farsi (FA). The *Plain* and *Weighted* models are built with our proposed post-processing technique over the base VecMap model.

followed their official implementations on GitHub.[9]

5.1 Intrinsic evaluation: Word translation

The task of word translation, or bilingual dictionary induction, consists in retrieving the correct translation in a target language given a word in a source language.

Experimental setting To predict the translation of a word, we return its nearest neighbor from the other language in the cross-lingual embedding space, using cosine similarity. The performance is evaluated with the precision at k metric (P-k, where $k \in \{1, 5, 10\}$), which is defined in this context as the percentage of test instances for which the correct answer is among the k highest ranked candidates. For this task, we used the standard test sets released by Conneau et al. and those extracted from Europarl (Dinu, Lazaridou, and Baroni 2015; Artetxe, Labaka, and Agirre 2017).

Results As can be observed in Table 2, the plain and weighted averaging methods yield the best overall results in Spanish, Italian and German. A similar pattern can be observed for Farsi, although in this case the results are poor overall. The base VecMap embeddings perform better overall in the distantly supervised setting than in the unsupervised setting and even the weakly-supervised setting of 100 pairs. This lends support to the usefulness of synthetically constructed dictionaries with identical tokens in the social media context. The trend contrasts with previous analyses in more standard corpora (Vulić and Korhonen 2016), where this seeding was proved inferior to other strategies. However, this behaviour is not consistent in the case of MUSE, which differs from what was found by Søgaard, Ruder, and Vulić (2018) on more standard corpora. With the exception of English-Farsi, where it fails to generalize[10], VecMap outperforms MUSE in both supervised and unsupervised settings.

Table 2 also shows that going from English to Farsi is challenging for all the tested models. This may be attributed to the structural differences between this language and English, and a reflection of cultural differences, which in turn causes a lower prevalence of English words. Indeed, the bilingual dictionary of identical tokens for Farsi is notably the smallest one: 6,142 word pairs against 66,037 for the second-smallest dictionary.

Finally, it is important to highlight that the test dictionaries contain a large number of words whose translation is identical to the word itself. Therefore, the fact that the tested methods use synthetic dictionaries which are based on identical tokens might be regarded as giving them an unfair advantage in this task. In particular, this means that the scores obtained for $P@1$ may be artificially high, especially for Spanish and Italian, where the number of words with identical translations is considerable.[11] However, we should stress that the training dictionaries are obtained automatically from the training corpora, and they were used by all comparison systems in the distantly supervised mode. Furthermore note that they share no connection with the test corpora other than being in the same language.[12] In what follows we discuss the capability of our post-processing technique by means of a qualitative analysis.

Analysis We performed an error analysis on our model, examining wrong translations, and found that in many cases, the mistranslated word was very similar to the correct translation. For example, in our *weighted* model the English verb *requested* is mapped to the Spanish verb *mandado* (*ordered*), and is also near its gold translation, *pedido*. As far as the baseline post-processing technique is concerned (i.e., Meemi), we can observe substantial drops in the quantitative scores with respect to the base model VecMap. A quick

[9]VecMap: https://github.com/artetxem/vecmap;
MUSE: https://github.com/facebookresearch/MUSE;
Meemi: https://github.com/yeraidm/meemi

[10]In this case, VecMap gets stuck in poor local optima, probably due to the non-optimal initialization in this language pair, an issue that was discussed in Artetxe, Labaka, and Agirre (2018b).

[11]The percentage of identical word pairs in the Facebook test sets are 16.5% for Spanish, 21.1% for Italian, 16.0% for German, and 4.3% for Farsi. Accordingly, a significant percentage of test pairs are included in our synthetic dictionary: 16.1% for Spanish, 19.0% for Italian, 15.7% for German, and 0.4% for Farsi.

[12]In the Europarl test sets word translations were obtained from alignments of the European Parliament proceedings, and therefore reflect a realistic distribution of the languages in that domain.

review of the output reveals a general trend of translating source words to target words in the same language (i.e., an English word in the source domain is often translated to some English word which also exists in the target domain). This phenomenon can also be observed for our model. For example, the five nearest neighbors of the English word *recognize* are also English words from the induced Spanish dictionary: *recognize*, *recognizes*, *acknowledge*, *acknowledged* and *acknowledgement*. While this is not the intended result for the bilingual dictionary induction task, this reveals a seamless integration of both languages which may partially explain the success of these embeddings in cross-lingual sentiment analysis (Section 5.2).

Finally, we performed a more qualitative analysis on the types of translations that cannot be found in standard dictionaries, for which cross-lingual embeddings trained on Twitter are particularly well-suited. Table 3 shows some examples for translations of selected English slang words and neologisms found in our *weighted* model (top three nearest neighbours according to cosine similarity). From the examples presented, we may draw special attention to *chillax*, a neologism composed of the verbs *chill* and *relax*, which translates also to colloquial ways of referring to the same idea across languages (*relajadito* in Spanish), and perhaps evoking more the notion of coziness in German (*gemütlich*). Let us also highlight acronyms like *wth* and *omfg*, whose translations denote surprise, but in an informal register (going as far as translating into swearwords in Farsi, for example). In line with the quantitative results, for Farsi we find mostly noise, but also interesting translations like بیدارشب (*sleep*), بیدارشب (*wake up*) and بیدارشب (*night awake*) for 'chillax'.

5.2 Extrinsic evaluation: Cross-lingual sentiment analysis

In this section we test the performance of our cross-lingual embeddings in the sentiment analysis (SA) task. We focus in particular on polarity classification (Pang and Lee 2008).

Experimental setting We selected an annotated dataset of English tweets as training data, and annotated datasets of Spanish, Italian and German tweets as test data. Since our main aim is the comparison of the cross-lingual embeddings, we used a standard Bidirectional Long Short-Term Memory (BiLSTM) architecture as classification system, with the same configuration across all experiments.[13] We used the cross-lingual embeddings for initializing the embedding layer. Given our cross-lingual evaluation setting, the weights of this embedding layer were not updated during training.

Datasets As training data we used the English dataset of the Sentiment Analysis in Twitter task of SemEval 2016 (Nakov et al. 2016). For evaluation we used the General Corpus of TASS (GCTASS) (Villena-Román et al. 2013), COST (Martínez-Cámara et al. 2015) and InterTASS (Díaz-Galiano et al. 2018) for Spanish, Sentipolc (Barbieri et al. 2016a) for Italian and SB-10K (Cieliebak et al. 2017) for German. Table 4 lists statistics of these datasets. We carried out both two-class (positive and negative) and three-class (positive, neutral and negative) evaluations with the GCTASS, InterTASS, Sentipolc and SB-10K datasets, and a two-class evaluation with COST.

Lower and upper bounds In addition to the comparison systems, in this experiment we also considered two lower bound systems and one upper bound, aimed at providing a broader context for our experimental results. As lower bound systems we included: (1) always predicting the majority class from the SemEval 2016 training corpus; and (2) training and testing the neural network with a set of monolingual English embeddings (FastText EN). This latter baseline is introduced as a sanity check, as its only source for cross-lingual transfer comes from the fact that the vocabularies of different languages may overlap. The upper bound is a monolingual BiLSTM classification system which is trained for each test dataset using the associated training data.

Results Table 5 summarizes the results for this cross-lingual SA evaluation. Our main findings, which are consistent for the three languages, are as follows: (a) there are no large differences between the unsupervised and distantly supervised variants of MUSE and VecMap, which in general behave similarly to the two lower bound baselines; (b) the results of the Meemi post-processing technique are also in line with the base VecMap model; (c) our two post-processing techniques lead to substantial improvements over the base VecMap model, including its weakly-supervised variant; and (d) using frequency weighting clearly outperforms the unweighted variant of our model, with peak performances on COST and InterTASS.

In general, the results provided by our simple post-processing technique are encouraging, especially taking into account that (1) these embeddings were learned without making use of any external resources or bilingual data, (2) no data in the target language was used for training, and (3) the distribution of the English dataset used for training clearly differs from all these datasets (see Table 4). What is particularly surprising is the performance gap of our proposed technique with respect to the state-of-the-art cross-lingual embeddings of VecMap and MUSE. In fact, our weighted postprocessing technique leads to improvements of over 40% over the base models in most cases.

Analysis The main difference of our proposed averaging methods compared to VecMap and MUSE lies in the fact that they are creating *anchor* points between languages. This turns out to be essential in a zero-shot cross-lingual transfer task. As argued in Section 4.1, identical tokens such as emoji, numerals or homographs provide a reliable bilingual signal, and anchoring them to a middle point in the vector space facilitates the learning process. For example, the following Spanish tweet *Buenos Dias a todos, menos a mi :(* (*Good Morning everyone, except for me*) was tagged as positive for both VecMap and MUSE, irrespective of their su-

[13] More details about the model and configuration (hyperparameters, etc.) are provided in the appendix.

wth				supernerd			
ES	IT	DE	FA	ES	IT	DE	FA
pufff	schifo	hä	نمیایی	frikifan	ratman	lovecrafts	باییم
aggg	chissene	hääh	نصیز	friky	cinecomic	trilogie	ماتریکس
madremia	chifo	näää	نمیکت	frikie	fumetti	gamestar	فیلماییه
chillax				**omfg**			
ES	IT	DE	FA	ES	IT	DE	FA
relajadito	rilassando	entspan	خواب	diooo	mioddio	maaah	نمیفهمیدی
relaxx	rilasso	gemütlich	بیدارشو	wtfff	ommiodio	njaahah	نمیبخشم
relajaito	rilassa	relaxte	بیدارشب	diooo	oddiooo	hahaha	***ک

Table 3: Translations of slang words and neologisms.

pervision. Similarly, VecMap and MUSE tagged the Italian tweet *Alla ricerca del nirvana* (*Looking for nirvana*) as negative. These systems thus overlooked a key emotion feature, i.e., :(, and a critical loanword, i.e., *nirvana*. In contrast, the same sentiment analysis model trained with our *weighted* cross-lingual embeddings correctly classified these two examples.

6 Ablation analysis

As shown throughout all the experiments, using identical tokens as supervision proved more robust than fully-unsupervised methods. In order to get more insights from the results achieved in both evaluation tasks, we performed an ablation test on the different types of identical tokens in the synthetic dictionaries (see Section 4.1). For this analysis, we focus on the base VecMap model and our proposed *weighted* post-processing strategy.

Table 6 shows the results of this ablation test on the two considered tasks: word translation and cross-lingual sentiment analysis (SA).[14] Unsurprisingly, the dictionaries of shared words (i.e. identical tokens that are neither numerals nor emoji) provide the best results among the individual features, often being close to the full dictionary of identical tokens. This type of dictionary is the largest in size, comprising over 95% of all identical tokens in the cases of Spanish and Italian. However, for the base VecMap model dictionaries consisting of either numerals or emoji seem to be enough to achieve similar results in most tasks and languages. This is not the case when using our weighted post-processing, which highlights its potential for taking advantage of the heterogeneity of all identical tokens. In fact, using the dictionary of all identical tokens consistently provides the best results in all tasks and languages (including Farsi) except in one single measure in German SA.

[14]Due to space constraints and for the sake of clarity, for this ablation test, Table 6 shows the results on the Facebook datasets on word translation and on the 3-class configuration on SA, using the most recent InterTASS dataset for Spanish.

Dataset	Positive	Neutral	Negative	Total
$\text{SemEval}_{\text{EN}}$	3,094	2,043	863	5,999
$\text{GCTASS}_{\text{ES}}$	22,233	1,305	15,845	39,382
$\text{InterTASS}_{\text{ES}}$	642	216	768	1,625
COST_{ES}	5,637	-	5,789	11,426
$\text{Sentipolc}_{\text{IT}}$	316	255	734	1,305
$\text{SB-10K}_{\text{DE}}$	533	351	216	1,426

Table 4: Size of the sentiment analysis datasets.

7 Conclusion

The main contribution of this paper is two-fold. On the one hand, we have presented a comprehensive study on the performance of state-of-the-art methods for learning cross-lingual embeddings without external bilingual data in the domain of social media communication. The overall results are encouraging, as they show that high-quality cross-lingual embeddings can be obtained directly from noisy user-generated corpora without external resources via distant supervision. These embeddings can be leveraged for cross-lingual downstream applications where training data may be scarce, as shown in our sentiment analysis experiments. However, our evaluation suggests there is significant room for improvement overall. Our results show that, especially in the case of distant languages such as English-Farsi, state-of-the-art cross-lingual mappings fail to learn an accurate mapping between the languages.

On the other hand, we have also introduced a simple post-processing technique which alters the embeddings of tokens that appear in both languages by simply averaging their initial embeddings. Despite its simplicity, our proposed technique clearly improves the quality of state-of-the-art cross-lingual word embedding approaches. In fact, we showed how a standard sentiment analysis system can achieve results of up to 80% in accuracy without the need of any training data in the test language by using our proposed method, improving the state of the art by more than 40% in several cases. The results also suggest that our method can be further improved by tuning it to specific applications or by exploiting the underlying idea to local neighbours in the vector space, amplifying its impact. In general, the simplicity of our approach opens up exciting avenues of research on cross-lingual applications in social media, where annotated data in English can be exploited for other languages with few resources. The construction of cross-lingual embedding

Supervision	Model	COST (ES) 2-class F1	COST (ES) 2-class Acc	GCTASS (ES) 3-class F1	GCTASS (ES) 3-class Acc	GCTASS (ES) 2-class F1	GCTASS (ES) 2-class Acc	InterTASS (ES) 3-class F1	InterTASS (ES) 3-class Acc	InterTASS (ES) 2-class F1	InterTASS (ES) 2-class Acc	Sentipolc (IT) 3-class F1	Sentipolc (IT) 3-class Acc	Sentipolc (IT) 2-class F1	Sentipolc (IT) 2-class Acc	SB-10K (DE) 3-class F1	SB-10K (DE) 3-class Acc	SB-10K (DE) 2-class F1	SB-10K (DE) 2-class Acc
Unsupervised	MUSE	50.0	58.0	5.2	26.1	37.0	58.4	24.0	28.2	32.0	45.8	22.1	25.9	25.3	31.1	27.8	33.4	**60.0**	62.5
	VecMap	57.9	61.5	22.8	35.6	37.8	57.6	22.9	27.4	33.7	45.3	21.2	24.5	26.3	32.0	36.6	45.8	56.1	56.2
Distant	MUSE	49.4	57.7	25.1	52.3	37.9	58.3	22.9	33.8	35.1	45.9	23.3	25.4	28.1	32.8	41.7	50.8	56.8	57.1
	VecMap	46.8	56.2	24.1	43.4	37.0	58.3	25.5	36.4	33.1	46.2	22.3	25.7	27.0	32.3	40.0	50.3	55.3	60.0
	Meemi	45.1	54.9	25.5	41.0	37.2	58.0	26.3	36.9	33.0	45.1	24.8	25.7	25.0	31.0	41.4	58.6	58.1	58.6
	Plain	77.4	77.5	33.0	46.2	50.7	62.4	33.4	33.4	63.4	63.4	26.7	28.7	36.7	38.1	42.2	47.5	57.9	63.8
	Weighted	**80.4**	**80.5**	**42.6**	**53.5**	**64.7**	**66.2**	**45.3**	**51.7**	**65.9**	**67.2**	**30.7**	**32.0**	**51.3**	**51.5**	**44.8**	57.3	57.7	**65.4**
100 pairs	VecMap	58.0	63.1	25.8	50.6	37.1	58.3	24.2	39.5	32.0	45.8	24.3	27.2	25.6	31.3	36.1	56.6	57.2	62.3
	MUSE	56.8	62.4	26.0	43.9	37.6	58.4	23.7	35.4	32.1	45.9	23.6	25.1	23.9	30.4	42.0	**60.4**	56.8	63.3
Lower bounds	Majority	33.0	49.3	24.1	56.5	36.9	58.4	18.9	39.5	31.3	45.6	13.0	24.2	23.1	30.1	12.4	23.0	37.6	60.2
	FT (EN)	50.2	57.1	23.2	36.0	37.5	58.0	21.9	37.2	32.3	43.9	25.6	26.3	36.6	37.6	40.2	55.2	58.0	58.0
Upper bound	FT (ES/IT/DE)	87.9	87.9	56.1	78.4	80.5	81.0	49.7	59.2	71.1	71.8	49.4	53.6	73.3	75.8	62.1	72.3	76.1	77.4

Table 5: Macro-average F1 and accuracy (%) results in the cross-lingual SA evaluation, using different embeddings as features. The *Plain* and *Weighted* models are built with our proposed post-processing technique over the base VecMap model.

Model	Supervision	EN-ES Word trans. P1	EN-ES Word trans. P5	EN-ES Word trans. P10	EN-ES SA F1	EN-ES SA Acc	EN-IT Word trans. P1	EN-IT Word trans. P5	EN-IT Word trans. P10	EN-IT SA F1	EN-IT SA Acc	EN-DE Word trans. P1	EN-DE Word trans. P5	EN-DE Word trans. P10	EN-DE SA F1	EN-DE SA Acc	EN-FA Word trans. P1	EN-FA Word trans. P5	EN-FA Word trans. P10
VecMap	*All*	**8.5**	16.9	**21.6**	25.5	**36.4**	**9.1**	16.8	21.8	22.3	**25.7**	2.6	**6.7**	**9.6**	40.0	50.3	0.2	0.5	1.1
	Numerals	7.6	15.7	20.2	23.1	36.1	8.6	17.2	21.9	**23.6**	24.8	2.7	6.4	9.3	38.9	43.4	0.0	0.0	0.0
	Emoji	7.8	16.9	21.2	**27.3**	32.6	8.6	16.8	21.8	22.6	24.7	**3.1**	6.2	8.5	44.3	55.5	**0.5**	1.3	1.7
	Words	8.1	**17.0**	**21.6**	23.5	31.6	8.8	**17.5**	22.0	20.7	24.6	2.8	6.5	8.7	**44.8**	**57.7**	0.3	**1.4**	**2.0**
	Unsup.	8.1	16.4	20.4	22.9	27.4	8.8	17.0	**22.3**	21.2	24.5	0.1	0.4	0.5	36.6	45.8	0.0	0.0	0.0
Weighted	*All*	**16.7**	**22.8**	**28.5**	**45.3**	**51.7**	**21.2**	**28.9**	**33.5**	**30.7**	**32.0**	**16.2**	**18.0**	**19.9**	**44.8**	57.3	**1.2**	**1.5**	**1.8**
	Numerals	7.6	15.7	20.2	23.0	25.4	8.5	17.1	21.9	23.7	24.8	2.7	6.4	9.3	35.9	43.9	0.0	0.0	0.0
	Emoji	7.7	16.9	21.2	24.5	34.3	8.5	16.8	21.8	22.4	24.0	3.1	6.1	8.5	36.7	43.7	0.5	1.3	1.7
	Words	**16.7**	22.7	28.3	37.2	37.4	**21.2**	**28.9**	33.4	21.6	25.4	**16.2**	17.8	19.8	**44.8**	**57.7**	**1.2**	1.4	**1.8**

Table 6: Ablation test. Tasks: word translation (Word trans.) and cross-lingual sentiment analysis (SA).

models also paves the way for the development of unsupervised machine translation systems (Artetxe et al. 2018; Lample et al. 2018), in this case specifically targeting noisy user-generated text for which parallel data is extremely scarce, and not even available at all for widely spoken language pairs. Indeed, standard machine translation tools are generally not suited for the kind of noisy text that is found in social media, where the language used is very dynamic and new terms are constantly being introduced.

Acknowledgements

Steven Schockaert was supported by ERC Starting Grant 637277. Yerai Doval is supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through projects FFI2014-51978-C2-2-R, TIN2017-85160-C2-1-R and TIN2017-85160-C2-2-R; the Spanish State Secretariat for Research, Development and Innovation and the European Social Fund (ESF) under an FPI fellowship (BES-2015-073768) associated to project FFI2014-51978-C2-1-R; and by the Galician Regional Government under project ED431D 2017/12. Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Formación (FJCI-2016-28353). Finally, we would like to thank Taher Pilehvar for his help with the Farsi qualitative analysis.

Appendix: Sentiment Analysis Classification System

We provide specific details of the classification system used in our sentiment analysis experiments (Section 4.2 of the paper). As classification system, we made use of a standard Bidirectional Long Short-Term Memory (BiLSTM) recurrent neural network architecture (Hochreiter and Schmidhuber 1997), developed in Keras (Chollet 2015). In the following we describe the details and hyperparameters of the specific architecture which is used across all experiments. The goal of the neural network is the classification of the opinion of tweets, hence the input is composed of a sequence of tokens of a tweet, and the output is the sentiment meaning of the tweet (t). Specifically, the input of the neural network is the sequence of tokens $t_{1:l}$.

The first layer is the embeddings lookup layer, which re-

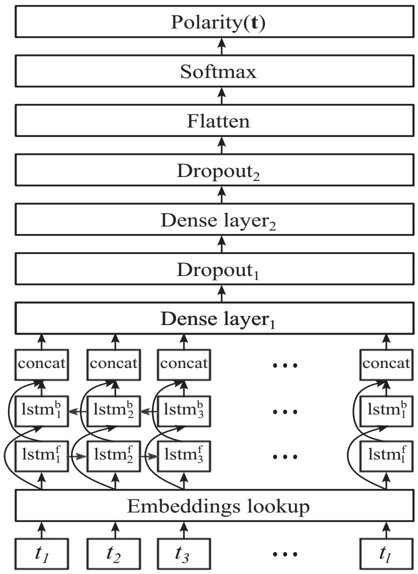

Figure 1: Architecture of the neural network developed for the cross-lingual sentiment analysis evaluation.

turns the sequence $s \in \mathbb{R}^{l \times 100}$. Since our aim is to test our cross-lingual embeddings, which were used to initialize the embedding layer, in a cross-lingual setting, the embedding weights are not updated during the training of the network. The output of the embedding layer is encoded by a BiLSTM, which is an elaboration of two Long Short-Term Memory (LSTM) layers. One of the LSTM layers processes the sequence $s_{1:l}$ ($LSTM^f$), and the second one processes the sequence $s_{l:1}$ ($LSTM^b$). We concatenated the output of the two LSTM layers, which are the state vectors of each token of the sequence s. Since the number of internal units of each LSTM layer is 128, the output of the BiLSTM layer is the sequence $c \in \mathbb{R}^{l \times 256}$.

Two fully connected layers activated by the ReLU function (Nair and Hinton 2010) process the sequence to the output of the BiLSTM layer. The output dimensions of the two fully connected layers are 64 and 32, respectively. A dropout layer is added after each fully connected layer, with a rate value of 0.5. L_2 regularization is applied to the weights of the fully connected layers with a value of 0.001, and to the output of the fully connected layers with a value of 0.0001. The output of the last fully connected layer is flattened, hence the dimension of the sequence c after the processing of the two dense layers is $\mathbb{R}^{l \times 32}$. The last layer is a softmax classification function. The output dimension of the softmax layer depends on the number of opinion labels (o), which in our case is 2 or 3 ($o \in \{2, 3\}$). Finally, the training is performed by a cross-entropy loss function, and optimized using Adam. For the sake of clarity, Figure 1 depicts the architecture of the neural network architecture.

References

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2018. Unsupervised neural machine translation. In *Proceedings of ICLR*.

Artetxe, M.; Labaka, G.; and Agirre, E. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL*, 451–462.

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 AAAI*.

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*, 789–798.

Barbieri, F.; Basile, V.; Croce, D.; Nissim, M.; Novielli, N.; and Patti, V. 2016a. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of CLiC-it*.

Barbieri, F.; Kruszewski, G.; Ronzano, F.; and Saggion, H. 2016b. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of ACMM*, 531–535.

Barnes, J.; Klinger, R.; and Schulte im Walde, S. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2483–2493. Melbourne, Australia: Association for Computational Linguistics.

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.

Cappallo, S.; Mensink, T.; and Snoek, C. G. 2015. Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of ACMM*, 1311–1314.

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

Chen, X.; Sun, Y.; Athiwaratkun, B.; Cardie, C.; and Weinberger, K. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics* 6:557–570.

Chollet, F. 2015. Keras. https://github.com/fchollet/keras.

Cieliebak, M.; Deriu, J. M.; Egger, D.; and Uzdilli, F. 2017. A twitter corpus and benchmark resources for german sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 45–51. Association for Computational Linguistics.

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

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of RANLP*, 198–206.

Díaz-Galiano, M. C.; Martínez-Cámara, E.; Ángel García Cumbreras, M.; Vega, M. G.; and Román, J. V.

2018. The democratization of deep learning in TASS 2017. *Procesamiento del Lenguaje Natural* 60:37–44.

Dinu, G.; Lazaridou, A.; and Baroni, M. 2015. Improving zero-shot learning by mitigating the hubness problem. In *Proceedings of ICLR, Workshop track*.

Doval, Y.; Camacho-Collados, J.; Anke, L. E.; and Schockaert, S. 2019. Meemi: A Simple Method for Post-processing Cross-lingual Word Embeddings. *arXiv preprint*. arXiv: 1910.07221.

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 EMNLP*.

Faruqui, M., and Dyer, C. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of EACL*, 462–471.

Felbo, B.; Mislove, A.; Søgaard, A.; Rahwan, I.; and Lehmann, S. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of EMNLP*, 1615–1625.

Glavaš, G.; Litschko, R.; Ruder, S.; and Vulić, I. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 710–721. Florence, Italy: Association for Computational Linguistics.

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*, 146–153.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2018. Processing social media messages in mass emergency: Survey summary. In *Companion Proceedings of the The Web Conference 2018*, 507–511. International World Wide Web Conferences Steering Committee.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5039–5049.

Lample, G.; Denoyer, L.; and Ranzato, M. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of ICLR*.

Martínez-Cámara, E.; Martín-Valdivia, M. T.; Ureña López, L. A.; and Montejo-Ráez, A. 2014. Sentiment analysis in Twitter. *Natural Language Engineering* 20(1):1–28.

Martínez-Cámara, E.; Martín-Valdivia, M. T.; Ureña-López, L. A.; and Mitkov, R. 2015. Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science* 41(3):263–272.

Meng, X.; Wei, F.; Liu, X.; Zhou, M.; Xu, G.; and Wang, H. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 572–581. Association for Computational Linguistics.

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

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, 807–814.

Nakov, P.; Ritter, A.; Rosenthal, S.; Sebastiani, F.; and Stoyanov, V. 2016. Semeval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of SemEval*, 1–18.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2:1–135.

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

Rasooli, M. S.; Farra, N.; Radeva, A.; Yu, T.; and McKeown, K. 2018. Cross-lingual sentiment transfer with limited resources. *Machine Translation* 32(1):143–165.

Reyes, A.; Rosso, P.; and Buscaldi, D. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering* 74:1–12.

Ruder, S.; Vulić, I.; and Søgaard, A. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research* 65:569–631.

Salameh, M.; Mohammad, S.; and Kiritchenko, S. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 767–777. Denver, Colorado: Association for Computational Linguistics.

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 ACL*, 778–788.

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*, 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, 1661–1670.

Vilares, D.; Alonso, M. A.; and Gómez-Rodríguez, C. 2017.

Supervised sentiment analysis in multilingual environments. *Information Processing & Management* 53(3):595 – 607.

Villena-Román, J.; Lana-Serrano, S.; Martínez-Cámara, E.; and González-Cristóbal, J. C. 2013. TASS - workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural* 50(0):37–44.

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, 247–257.

Xing, C.; Wang, D.; Liu, C.; and Lin, Y. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of NAACL*, 1006–1011.

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*, 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.

Zhang, M.; Liu, Y.; Luan, H.; and Sun, M. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of EMNLP*, 1934–1945.

Zhou, X.; Wan, X.; and Xiao, J. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of ACL*, 1403–1412.