Two Methods for Domain Adaptation of Bilingual Tasks: Delightfully Simple and Broadly Applicable

Viktor Hangya[1], **Fabienne Braune**[1,2], **Alexander Fraser**[1], **Hinrich Schütze**[1]
[1]Center for Information and Language Processing
LMU Munich, Germany
[2]Volkswagen Data Lab Munich, Germany
{hangyav, fraser}@cis.uni-muenchen.de
fabienne.braune@volkswagen.de

Abstract

Bilingual tasks, such as bilingual lexicon induction and cross-lingual classification, are crucial for overcoming data sparsity in the target language. Resources required for such tasks are often out-of-domain, thus domain adaptation is an important problem here. We make two contributions. First, we test a delightfully simple method for *domain adaptation of bilingual word embeddings*. We evaluate these embeddings on two bilingual tasks involving different domains: cross-lingual twitter sentiment classification and medical bilingual lexicon induction. Second, we tailor a *broadly applicable semi-supervised classification* method from computer vision to these tasks. We show that this method also helps in low-resource setups. Using both methods together we achieve large improvements over our baselines, by using only additional unlabeled data.

1 Introduction

In this paper we study two bilingual tasks that strongly depend on bilingual word embeddings (BWEs). Previously, specialized domain adaptation approaches to such tasks were proposed. We instead show experimentally that a simple adaptation process involving only unlabeled text is highly effective. We then show that a semi-supervised classification method from computer vision can be applied successfully for further gains in cross-lingual classification.

Our BWE adaptation method is delightfully simple. We begin by adapting monolingual word embeddings to the target domain for source and target languages by simply building them using both general and target-domain unlabeled data. As a second step we use post-hoc mapping (Mikolov et al., 2013b), i.e., we use a seed lexicon to transform the word embeddings of the two languages into the same vector space. We show experimentally for the first time that the domain-adapted bilingual word embeddings we produce using this extremely simple technique are highly effective. We study two quite different tasks and domains, where resources are lacking, showing that our simple technique performs well for both of them: cross-lingual twitter sentiment classification and medical bilingual lexicon induction. In previous work, task-dependent approaches were used for this type of domain adaptation. Our approach is simple and task independent.

Second, we adapt the semi-supervised image classification system of Häusser et al. (2017) for NLP problems for the first time. This approach is broadly applicable to many NLP classification tasks where unlabeled data is available. We tailor it to both of our cross-lingual tasks. The system exploits unlabeled data during the training of classifiers by learning similar features for similar labeled and unlabeled training examples, thereby extracting information from unlabeled examples as well. As we show experimentally, the system further improves cross-lingual knowledge transfer for both of our tasks.

After combining both techniques, the results of sentiment analysis are competitive with systems that use annotated data in the target language, an impressive result considering that we require no target-language annotated data. The method also yields impressive improvements for bilingual lexicon induction compared with baselines trained on in-domain data. We show that this system requires the high-quality domain-adapted bilingual word embeddings we previously created to use unlabeled data well.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 810–820
Melbourne, Australia, July 15 - 20, 2018.

2 Previous Work

2.1 Bilingual Word Embeddings

Many approaches have been proposed for creating high quality BWEs using different bilingual signals. Following Mikolov et al. (2013b), many authors (Faruqui and Dyer, 2014; Xing et al., 2015; Lazaridou et al., 2015; Vulić and Korhonen, 2016) map monolingual word embeddings (MWEs) into the same bilingual space. Others leverage parallel texts (Hermann and Blunsom, 2014; Gouws et al., 2015) or create artificial cross-lingual corpora using seed lexicons or document alignments (Vulić and Moens, 2015; Duong et al., 2016) to train BWEs.

In contrast, our aim is not to improve the intrinsic quality of BWEs, but to adapt BWEs to specific domains to enhance their performance on bilingual tasks in these domains. Faruqui et al. (2015), Gouws and Søgaard (2015), Rothe et al. (2016) have previously studied domain adaptation of bilingual word embeddings, showing it to be highly effective for improving downstream tasks. However, importantly, their proposed methods are based on specialized domain lexicons (such as, e.g., sentiment lexicons) which contain task specific word relations. Our delightfully simple approach is, in contrast, effectively task independent (in that it only requires unlabeled in-domain text), which is an important strength.

2.2 Cross-Lingual Sentiment Analysis

Sentiment analysis is widely applied, and thus ideally we would have access to high quality supervised models in all human languages. Unfortunately, good quality labeled datasets are missing for many languages. Training models on resource rich languages and applying them to resource poor languages is therefore highly desirable. Cross-lingual sentiment classification (CLSC) tackles this problem (Mihalcea et al., 2007; Banea et al., 2010; Wan, 2009; Lu et al., 2011; Balamurali and Joshi, 2012; Gui et al., 2013). Recent CLSC approaches use BWEs as features of deep learning architectures which allows us to use a model for target-language sentiment classification, even when the model was trained only using source-language supervised training data. Following this approach we perform CLSC on Spanish tweets using English training data. Even though Spanish is not resource-poor we simulate this by using only English annotated data.

Xiao and Guo (2013) proposed a cross-lingual log-bilinear document model to learn distributed representations of words, which can capture both the semantic similarities of words across languages and the predictive information with respect to the classification task. Similarly, Tang and Wan (2014) jointly embedded texts in different languages into a joint semantic space representing sentiment. Zhou et al. (2014) employed aligned sentences in the BWE learning process, but in the sentiment classification process only representations in the source language are used for training, and representations in the target language are used for predicting labels. An important weakness of these three works was that aligned sentences were required.

Some work has trained sentiment-specific BWEs using annotated sentiment information in both languages (Zhou et al., 2015, 2016), which is desirable, but this is not applicable to our scenario. Our goal is to adapt BWEs to a specific domain without requiring additional task-specific engineering or knowledge sources beyond having access to plentiful target-language in-domain unlabeled text. Both of the approaches we study in this work fit this criterion, the delightfully simple method for adapting BWEs can improve the performance of any off-the-shelf classifier that is based on BWEs, while the broadly applicable semi-supervised approach of Häusser et al. (2017) can improve the performance of any off-the-shelf classifier.

2.3 Bilingual Lexicon Induction (BLI)

BLI is an important task that has been addressed by a large amount of previous work. The goal of BLI is to automatically extract word translation pairs using BWEs. While BLI is often used to provide an intrinsic evaluation of BWEs (Lazaridou et al., 2015; Vulić and Moens, 2015; Vulić and Korhonen, 2016) it is also useful for tasks such as machine translation (Madhyastha and España Bohnet, 2017). Most work on BLI using BWEs focuses on frequent words in high-resource domains such as parliament proceedings or news texts. Recently Heyman et al. (2017) tackled BLI of words in the medical domain. This task is useful for many applications such as terminology extraction or OOV mining for machine translation of medical texts. Heyman et al. (2017) show that when only a small amount of medical data is available, BLI using BWEs tends to perform poorly. Especially BWEs obtained using post-hoc mapping (Mikolov et al., 2013b; Lazaridou et al., 2015) fail on this task. Consequently, Heyman et al. (2017) build BWEs using aligned documents and then engineer a specialized classification-based approach to BLI. In contrast, our delightfully simple approach to create high-quality BWEs for the medical domain requires only monolingual data. We show that our adapted BWEs yield impressive improvements over non-adapted BWEs in this task with both cosine similarity and with the classifier of Heyman et al. (2017). In addition, we show that the broadly applicable method can push performance further using easily accessible unlabeled data.

3 Adaptation of BWEs

BWEs trained on *general domain* texts usually result in lower performance when used in a system for a *specific domain*. There are two reasons for this. (i) Vocabularies of specific domains contain words that are not used in the general case, e.g., names of medicines or diseases. (ii) The meaning of a word varies across domains; e.g., “apple” mostly refers to a fruit in general domains, but is an electronic device in many product reviews.

The delightfully simple method adapts general domain BWEs in a way that preserves the semantic knowledge from general domain data and leverages *monolingual* domain specific data to create domain-specific BWEs. Our domain-adaptation approach is applicable to any language-pair in which monolingual data is available. Unlike other methods, our approach is task independent: it only requires unlabeled in-domain target language text.

3.1 Approach

To create domain adapted BWEs, we first train MWEs (monolingual word embeddings) in both languages and then map those into the same space using post-hoc mapping (Mikolov et al., 2013b). We train MWEs for both languages by concatenating monolingual out-of-domain and in-domain data. The out-of-domain data allows us to create accurate distributed representations of common vocabulary while the in-domain data embeds domain specific words. We then map the two MWEs using a small seed lexicon to create the adapted BWEs. Because post-hoc mapping only requires a seed lexicon as bilingual signal it can easily be used with (cheap) monolingual data.

For **post-hoc mapping**, we use Mikolov et al. (2013b)'s approach. This model assumes a $W \in \mathbb{R}^{d_1 \times d_2}$ matrix which maps vectors from the source to the target MWEs where d_1 and d_2 are the embedding space dimensions. A seed lexicon of $(x_i, y_i) \in L \subseteq \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ pairs is needed where x_i and y_i are source and target MWEs. W can be learned using ridge regression by minimizing the L_2-regularized mapping error between the source x_i and the target y_i vectors:

$$\min_W \sum_i ||Wx_i - y_i||_2^2 + \lambda ||W||_2^2 \qquad (1)$$

where λ is the regularization weight. Based on the source embedding x, we then compute a target embedding as Wx.

We create MWEs with word2vec skipgram (Mikolov et al., 2013a)[1] and estimate W with *scikit-learn* (Pedregosa et al., 2011). We use default parameters.

4 Cross-Lingual Sentiment Classification

In CLSC, an important application of BWEs, we train a supervised sentiment model on training data available in the source (a resource rich language) and apply it to the target (a resource poor language, for which there is typically no training data available). Because BWEs embed source and target words in the same space, annotations in the source (represented as BWEs) enable transfer learning. For CLSC of tweets, a drawback of BWEs trained on non-twitter data is that they do not produce embeddings for twitter-specific vocabulary, e.g., slang words like English *coool* and (Mexican) Spanish *chido*, resulting in lost information when a sentiment classifier uses them.

4.1 Training Data for Twitter Specific BWEs

As comparable non-twitter data we use OpenSubtitles (Lison and Tiedemann, 2016) which contains 49.2M English and Spanish subtitle sentences respectively (**Subtitle**). The reason behind choosing Subtitles is that although it is out-of-domain it contains slang words similar to tweets thus serving as a strong baseline in our setup. We experiment with two monolingual twitter data sets:

(i) **22M_tweets**: Downloaded[2] English (17.2M) and Spanish (4.8M) tweets using the public

[1] https://github.com/dav/word2vec

[2] We downloaded for a month starting on 2016-10-15.

Twitter Streaming API[3] with language filters *en* and *es*

(ii) a **BACKGROUND** corpus of 296K English and 150K Spanish (non-annotated) tweets released with the test data of the RepLab task (Amigó et al., 2013) described below

All twitter data was tokenized using Bird et al. (2009) and lowercased. User names, URLs, numbers, emoticons and punctuation were removed.

As lexicon for the mapping, we use the BNC word frequency list (Kilgarriff, 1997), a list of 6,318 frequent English lemmas and their Spanish translations, obtained from Google Translate. Note that we do not need a domain-specific lexicon in order to get good quality adapted BWEs.

4.2 Training Data for Sentiment Classifiers

For sentiment classification, we use data from the *RepLab 2013* shared task (Amigó et al., 2013). The data is annotated with positive, neutral and negative labels and contains English and Spanish tweets. We used the official English training set (26.6K tweets) and the Spanish test set (14.9K) in the resource-poor setup. We only use the 7.2K Spanish labeled training data for comparison reasons in §6.2, which we will discuss later.

The shared task was on target-level sentiment analysis, i.e., given a pair (document, target entity), the gold annotation is based on whether the sentiment expressed by the document is about the target. For example: *I cried on the back seat of my BMW!* where *BMW* is the target would be negative in the sentence-level scenario. However, it is neutral in the target-level case because the negative sentiment is not related to BMW. The reason for using this dataset is that it contains comparable English and Spanish tweets annotated for sentiment. There are other twitter datasets for English (Nakov et al., 2016) and Spanish (Garcıa-Cumbreras et al., 2016), but they were downloaded at different times and were annotated using different annotation methodologies, thus impeding a clean and consistent evaluation.

4.3 Sentiment Systems

For evaluating our adapted BWEs on the *RepLab* dataset we used a **target-aware** sentiment classifier introduced by Zhang et al. (2016). The network first embeds input words using pre-trained BWEs and feeds them to a bi-directional gated neural network. Pooling is applied on the hidden representations of the left and right context of the target mention respectively. Finally, gated neurons are used to model the interaction between the target mention and its surrounding context. During training we hold our pre-trained BWEs fixed and keep the default parameters of the model.

We also implement Kim (2014)'s *CNN-non-static* system, which does not use the target information in a given document (**target-ignorant**). The network first embeds input words using pre-trained BWEs and feeds them to a convolutional layer with multiple window sizes. Max pooling is applied on top of convolution followed by a fully connected network with one hidden layer. We used this system as well because it performed comparably to the target-aware system. The reason for this is that only 1% of the used data contains more than one target and out of these rare cases only 14% have differing sentiment labels in the same sentence, which are the difficult cases of target-level sentiment analysis. We used the default parameters as described in (Kim, 2014) with the exception of using 1000 feature maps and 30 epochs, based on our initial experiments. Word embeddings are fixed during the training just as for the target-aware classifier.

4.4 Results

As we previously explained we evaluate our adaptation method on the task of target-level sentiment classification using both target-aware and target-ignorant classifiers. For all experiments, our two baselines are off-the-shelf classifiers using non-adapted BWEs, i.e., BWEs trained only using Subtitles. Our goal is to show that our BWE adaptation method can improve the performance of such classifiers. We train our adapted BWEs on the concatenation of Subtitle and 22M_tweets or BACKGROUND respectively. In addition, we also report results with BWEs trained only on tweets.

To train the sentiment classifiers we use the English Replab training set and we evaluate on the Spanish test set. To show the performance that can be reached in a monolingual setup, we report results obtained by using annotated Spanish sentiment data instead of English (oracle). We train two oracle sentiment classifiers using (i) MWEs trained on only the Spanish part of Subtitle and (ii)

[3] dev.twitter.com/streaming/overview

			target-	
			aware	ignorant
oracle		MWE_Subtitle	62.17%	63.27%
		BWE_Subtitle	62.46%	63.50%
domain adaptation		Baseline	55.14%	59.05%
	del. simple	BACKGROUND	56.79%	58.50%
		22M_tweets	59.44%	61.14%
		Subtitle+BACKGROUND	58.64%	59.34%
		Subtitle+22M_tweets	60.99%	61.06%

Table 1: Accuracy of the BWE adaptation approach on the target-level sentiment classification task. The oracle systems used Spanish sentiment training data instead of English.

BWEs trained on Subtitle using posthoc mapping. The difference between the two is that the embeddings of (ii) are enriched with English words which can be beneficial for the classification of Spanish tweets because they often contain a few English words.

We do not compare with word embedding adaptation methods relying on specialized resources. The point of our work is to study task-independent methods and to the best of our knowledge ours is the first such attempt. Similarly, we do not compare against machine translation based sentiment classifiers (e.g., (Zhou et al., 2016)) because for their adaptation in-domain parallel data would be needed.

Table 1 gives results for both classifiers. It shows that the adaptation of Subtitle based BWEs with data from Twitter (22M_tweets and BACKGROUND) clearly outperforms the Baseline in all cases. The target-aware system performed poorly with the baseline BWEs and could benefit significantly from the adaptation approach. The target-ignorant performed better with the baseline BWEs but could also benefit from the adaptation. Comparing results with the Twitter-dataset-only based BWEs, the 22M_tweets performed better even though the BACKGROUND dataset is from the same topic as the RepLab train and test sets. Our conjecture is that the latter is too small to create good BWEs. In combination with Subtitles, 22M_tweets also yields better results than when combined with BACKGROUND. Although the best accuracy was reached using the 22M_tweets-only based BWEs, it is only slightly better then the adapted Subtitles+22M_tweets based BWEs. In §6 we show that both the semantic knowledge from Subtitles and the domain-specific information from tweets are needed to further improve results.

Comparing the two classifiers we can say that they performed similarly in terms of their best results. On the other hand, the target-ignorant system had better results on average. This might seem surprising at first because the system does not use the target as information. But considering the characteristics of RepLab, i.e., that the number of tweets that contains multiple targets is negligible, using the target offers no real advantage.

Although we did not focus on the impact of the seed lexicon size, we ran post-hoc mapping with different sizes during our preliminary experiments. With 1,000 and 100 word pairs in the lexicon the target-ignorant system suffered 0.5% and 4.0% drop in average of our setups respectively.

To summarize the result: using adapted BWEs for the Twitter CLSC task improves the performance of off-the-shelf classifiers.

5 Medical Bilingual Lexicon Induction

Another interesting downstream task for BWEs is bilingual lexicon induction. Given a list of words in a source language, the goal of BLI is to mine translations for each word in a chosen target language. The medical bilingual lexicon induction task proposed in (Heyman et al., 2017) aims to mine medical words using BWEs trained on a very small amount of English and Dutch monolingual medical data. Due to the lack of resources in this domain, good quality BWEs are hard to build using in-domain data only. We show that by enriching BWEs with general domain knowledge (in the form of general domain monolingual corpora) better results can be achieved on this medical domain task.

5.1 Experimental Setup

We evaluate our improved BWEs on the dataset provided by Heyman et al. (2017). The monolingual medical data consists of English and Dutch medical articles from Wikipedia. The English (resp. Dutch) articles contain 52,336 (resp. 21,374) sentences. A total of 7,368 manually annotated word translation pairs occurring in the English (source) and Dutch (target) monolingual corpora are provided as gold data. This set is split 64%/16%/20% into trn/dev/test. 20% of the English words have multiple translations. Given an English word, the task is to find the correct Dutch translation.

As monolingual general-domain data we use

	cosine similarity		classifier	
	F_1 (top)	F_1 (all)	F_1 (top)	F_1 (all)
Baseline	13.43	9.84	37.73	36.61
Baseline BNC lexicon	-	-	20.73	21.78
Adapted medical lexicon	14.18	14.15	40.71	38.09
Adapted BNC lexicon	16.29	16.71	22.10	21.50

Table 2: We report F_1 results for medical BLI with the cosine similarity and the classifier based systems. We present baseline and our proposed domain adaptation method using both general and medical lexicons.

the English and Dutch data from Europarl (v7) (Koehn, 2005), a corpus of 2 million sentence pairs. Although Europarl is a parallel corpus, we use it in a monolingual way and shuffle each side of the corpus before training. By using massive cheap data we create high-quality MWEs in each language which are still domain-specific (due to inclusion of medical data). To obtain an out-of-domain seed lexicon, we translated the English words in BNC to Dutch using Google Translate (just as we did before for the Twitter CLSC task). We then use the out-of-domain BNC and the in-domain medical seed lexicons in separate experiments to create BWEs with post-hoc mapping. Note, we did not concatenate the two lexicons because (i) they have a small common subset of source words which have different target words, thus having a negative effect on the mapping and (ii) we did not want to modify the medical seed lexicon because it was taken from previous work.

5.2 BLI Systems

To perform BLI we use two methods. Because BWEs represent words from different languages in a shared space, BLI can be performed via *cosine similarity* in this space. In other words, given a BWE representing two languages V_s and V_t, the translation of each word $s \in V_s$ can be induced by taking the word $t \in V_t$ whose representation $\vec{x_t}$ in the BWE is closest to the representation $\vec{x_s}$.

As the second approach we use a *classifier based system* proposed by Heyman et al. (2017). This neural network based system is comprised of two main modules. The first is a character-level LSTM which aims to learn orthographic similarity of word pairs. The other is the concatenation of the embeddings of the two words using embedding layers with the aim of learning the similarity among semantic representations of the words. Dense layers are applied on top of the two modules before the output soft-max layer. The classifier is trained using positive and negative word pair examples and a pre-trained word embedding model. Negative examples are randomly generated for each positive one in the training lexicon. We used default parameters as reported by Heyman et al. (2017) except for the t classification thresholds (used at prediction time). We fine-tuned these on dev. We note that the system works with pre-trained MWEs as well (and report these as official baseline results) but it requires BWEs for candidate generation at prediction time, thus we use BWEs for the system's input for all experiments. In preliminary work, we had found that MWE and BWE results are similar.

5.3 Results

Heyman et al. (2017)'s results are our *baseline*. Table 2 compares its performance with our adapted BWEs, with both cosine similarity and classification based systems. "top" F_1 scores are based on the most probable word as prediction only; "all" F_1 scores use all words as prediction whose probability is above the threshold. It can be seen that the cosine similarity based system using adapted BWEs clearly outperforms the non-adapted BWEs which were trained in a resource poor setup.[4] Moreover, the best performance was reached using the general seed lexicon for the mapping which is due to the fact that general domain words have better quality embeddings in the MWE models, which in turn gives a better quality mapping.

The classification based system performs significantly better comparing to cosine similarity by exploiting the seed lexicon better. Using adapted BWEs as input word embeddings for the system further improvements were achieved which shows the better quality of our BWEs. Simulating an even poorer setup by using a general lexicon, the

[4] The results for cosine similarity in (Heyman et al., 2017) are based on BWESG-based BWEs (Vulić and Moens, 2016) trained on a small document aligned parallel corpus without using a seed lexicon.

performance gain of the classifier is lower. This shows the significance of the medical seed lexicon for this system. On the other hand, adapted BWEs have better performance compared to non-adapted ones using the best translation while they have just slightly lower F_1 using multiple translations. This result shows that while with adapted BWEs the system predicts better "top" translations, it has a harder time when predicting "all" due to the increased vocabulary size.

To summarize: we have shown that adapted BWEs increase performance for this task and domain; and they do so independently of the task-specific system that is used.

6 Semi-Supervised Learning

In addition to the experiments that show our BWE-adaptation method's task and language independence, we investigate ways to further incorporate unlabeled data to overcome data sparsity.

Häusser et al. (2017) introduce a semi-supervised method for neural networks that makes associations from the vector representation of labeled samples to those of unlabeled ones and back. This lets the learning exploit unlabeled samples as well. While Häusser et al. (2017) use their model for image classification, we adapt it to CLSC of tweets and medical BLI. We show that our semi-supervised model requires adapted BWEs to be effective and yields significant improvements. This innovative method is general and can be applied to any classification when unlabeled text is available.

6.1 Model

Häusser et al. (2017)'s basic assumption is that the embeddings of labeled and unlabeled samples – i.e., the representations in the neural network on which the classification layer is applied – are similar within the same class. To achieve this, walking cycles are introduced: a cycle starts from a labeled sample, goes to an unlabeled one and ends at a labeled one. A cycle is correct if the start and end samples are in the same class. The probability of going from sample A to B is proportional to the cosine similarity of their embeddings. To maximize the number of correct cycles, two loss functions are employed: Walker loss and Visit loss.

Walker loss penalizes incorrect walks and encourages a uniform probability distribution of walks to the correct class. It is defined as:

$$\mathcal{L}_{walker} := H(T, P^{aba}) \qquad (2)$$

where H is the cross-entropy function, P^{aba}_{ij} is the probability that a cycle starts from sample i and ends at j and T is the uniform target distribution:

$$T_{ij} := \begin{cases} 1/(\#c(i)) & \text{if } c(i) = c(j) \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

where $c(i)$ is the class of sample i and $\#c(i)$ is the number of occurrences of $c(i)$ in the labeled set.

Visit loss encourages cycles to visit all unlabeled samples, rather than just those which are the most similar to labeled samples. It is defined as:

$$\mathcal{L}_{visit} := H(V, P^{visit})$$

$$P^{visit}_j := \langle P^{ab}_{ij} \rangle_i \qquad (4)$$

$$V_j := \frac{1}{U}$$

where H is cross-entropy, P^{ab}_{ij} is the probability that a cycle starts from sample i and goes to j and U is the number of unlabeled samples.

The total loss during training is the sum of the walker, visit and classification (cross-entropy between predicted and gold labels) losses which is minimized using Adam (Kingma and Ba, 2015).

We adapt this model (including the two losses) to sentiment classification, focusing on the target-ignorant classifier, and the classifier based approach for BLI. We will call these systems **semisup**[5]. Due to the fact that we initialize the embedding layers for both classifiers with BWEs the models are able to make some correct cycles at the beginning of the training and improve them later on. We will describe the labeled and unlabeled datasets used in the subsequent sections below.

We use Häusser et al. (2017)'s implementation of the losses, with 1.0, 0.5 and 1.0 weights for the walker, visit and classification losses, respectively, for CLSC based on preliminary experiments. We fine-tuned the weights for BLI on dev for each experiment.

[5] We publicly release our implementation: https://github.com/hangyav/biadapt

		semisup
	Baseline	58.67% (-0.38%)
domain adaptation	BACKGROUND	57.41% (-1.09%)
	22M_tweets	60.19% (-0.95%)
	Subtitle+BACKGROUND	60.31% (0.97%)
	Subtitle+22M_tweets	63.23% (2.17%)

Table 3: Accuracy on CLSC of the adapted BWE approach with the semisup (target-ignorant with additional loss functions) system comparing to the target-ignorant in brackets.

6.2 Semi-Supervised CLSC

As in §4.4, we use pre-trained BWEs to initialize the classifier and use English sentiment training data as the labeled set. Furthermore, we use the Spanish sentiment training data as the unlabeled set, ignoring its annotation. This setup is very similar to real-word low-resource scenarios: unlabeled target-language tweets are easy to download while labeled English ones are available.

Table 3 gives results for adapted BWEs and shows that semisup helps only when word embeddings are adapted to the Twitter domain. As mentioned earlier, semisup compares labeled and unlabeled samples based on their vector representations. By using BWEs based on only Subtitles, we lose too many embeddings of similar English and Spanish tweets. On the other hand, if we use only tweet-based BWEs we lose good quality semantic knowledge which can be learned from more standard text domains. By combining the two domains we were able to capture both sides. For Subtitle+22M_tweets, we even get very close to the best oracle (BWE_Subtitle) in Table 1 getting only 0.27% less accuracy – an impressive result keeping in mind that we did not use labeled Spanish data.

The RepLab dataset contains tweets from 4 topics: automotive, banking, university, music. We manually analyzed similar tweets from the labeled and unlabeled sets. We found that when using semisup, English and Spanish tweets from the same topics are more similar in the embedding space than occurs without the additional losses. Topics differ in how they express sentiment – this may explain why semisup increases performance for RepLab.

Adding supervision. To show how well semisup can exploit the unlabeled data we used both English and Spanish sentiment training data together to train the sentiment classifiers. Table 4 shows that by using annotated data in both languages we get clearly better results than when using only one language. Tables 3 and 4 show that for Subtitle+22M_tweets based BWEs, the semisup approach achieved high improvement (2.17%) comparing to target-ignorant with English training data only, while it achieved lower improvement (0.97%) with the Subtitle+BACKGROUND based BWEs. On the other hand, adding labeled Spanish data caused just a slight increase comparing to semisup with Subtitle+22M_tweets based BWEs (0.59%), while in case of Subtitle+BACKGROUND we got significant additional improvement (2.61%). This means that with higher quality BWEs, unlabeled target-language data can be exploited better.

It can also be seen that the target-aware system outperformed the target-ignorant system using additional labeled target-language data. The reason could be that it is a more complex network and therefore needs more data to reach high performance.

The results in table 4 are impressive: our target-level system is strongly competitive with the official shared task results. We achieved high accuracy on the Spanish test set by using only English training data. Comparing our best system which used all training data to the official results (Amigó et al., 2013) we would rank 2^{nd} even though our system is not fine-tuned for the RepLab dataset. Furthermore, we also outperformed the oracles when using annotated data from both languages which shows the additional advantage of using BWEs.

6.3 Semi-Supervised BLI

For BLI experiments with semisup we used word pairs from the medical seed lexicon as the labeled set (with negative word pairs generated as described in §5.2). As opposed to CLSC and the work of (Häusser et al., 2017), for this task we do not have an unlabeled set, and therefore we need to generate it. We developed two scenarios. For the first, **BNC**, we generate a general unlabeled set using English words from the **BNC** lexicon and generate 10 pairs out of each word by using the 5 most similar Dutch words based on the corresponding BWEs and 5 random Dutch words. For the second scenario, **medical**, we generate an in-domain unlabeled set by generating for each English word in the **medical** lexicon the 3 most similar Dutch

		lang	target-aware	target-ignorant
oracle	MWE_Subtitle	Es	62.17%	63.27%
	BWE_Subtitle	Es	62.46%	63.50%
domain adaptation	Subtitle+BACKGROUND	En	58.64%	59.34%
	Subtitle+BACKGROUND	En+Es	64.01%	62.92% (2.61%)
	Subtitle+22M_tweets	En	60.99%	61.06%
	Subtitle+22M_tweets	En+Es	64.24%	63.82% (0.59%)

Table 4: Accuracy on CLSC of both target-aware and target-ignorant systems using English or/and Spanish sentiment training data. Column *lang* shows the language of the used training data. Differences comparing to semisup are indicated in brackets.

	F_1 (top)	F_1 (all)
Baseline+BNC	35.04 (-0.66)	34.98 (-1.40)
Baseline+medical	36.20 (0.50)	36.55 (0.16)
Adapted+BNC	41.01 (0.30)	39.04 (0.95)
Adapted+medical	41.44 (0.73)	37.51 (-0.57)

Table 5: Results with the semi-supervised system for BLI. Differences comparing to previous results are indicated in brackets. Baseline results are compared to rerun experiments of Heyman et al. (2017) using BWEs instead of MWEs.

words based on BWEs and for each of these we use the 5 most similar English words (ignoring the words which are in the original medical lexicon) and 5 negative words. The idea behind these methods is to automatically generate an unlabeled set that hopefully has a similar positive and negative word pair distribution to the distribution in the labeled set.

Results in Table 5 show that adding semisup to the classifier further increases performance for BLI as well. For the baseline system, when using only in-domain text for creating BWEs, only the medical unlabeled set was effective, general domain word pairs could not be exploited due to the lack of general semantic knowledge in the BWE model. On the other hand, by using our domain adapted BWEs, which contain both general domain and in-domain semantical knowledge, we can exploit word pairs from both domains. Results for adapted BWEs increased in 3 out of 4 cases, where the only exception is when using multiple translations for a given source word (which may have been caused by the bigger vocabulary size).

These results show that adapted BWEs are needed to exploit unlabeled data well which leads to an impressive overall 3.71 increase compared with the best result in previous work (Heyman et al., 2017), by using only unlabeled data.

7 Conclusion

Bilingual word embeddings trained on general domain data yield poor results in out-of-domain tasks. We presented experiments on two different low-resource task/domain combinations. Our delightfully simple task independent method to adapt BWEs to a specific domain uses unlabeled monolingual data only. We showed that with the support of adapted BWEs the performance of off-the-shelf methods can be increased for both cross-lingual Twitter sentiment classification and medical bilingual lexicon induction. Furthermore, by adapting the broadly applicable semi-supervised approach of Häusser et al. (2017) (which until now has only been applied in computer vision) we were able to effectively exploit unlabeled data to further improve performance. We showed that, when also using high-quality adapted BWEs, the performance of the semi-supervised systems can be significantly increased by using unlabeled data at classifier training time. In addition, CLSC results are competitive with a system that uses target-language labeled data, even when we use no such target-language labeled data.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable input. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 640550).

References

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten de Rijke, Damiano Spina, Enrique Amigo, Jorge Carrillo de Albornoz, Tamara Martin, and Maarten de Rijke. 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Proc. CLEF*.

A.R. Balamurali and Adity Joshi. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proc. COLING*.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proc. COLING*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. In *Proc. EMNLP*.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proc. NAACL-HLT*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. EACL*.

Miguel Ángel Garcıa-Cumbreras, Julio Villena-Román, Eugenio Martınez-Cámara, Manuel Carlos Díaz-Galiano, María-Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2016. Overview of tass 2016. In *Proc. TASS*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proc. ICML*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proc. NAACL-HLT*.

Lin Gui, Ruifeng Xu, Qin Lu, Jun Xu, Jian Xu, Bin Liu, and Wang Xiaolong. 2013. A mixed model for cross lingual opinion analysis. In *Proc. NLPCC*.

Philip Häusser, Alexander Mordvintsev, and Daniel Cremers. 2017. Learning by Association - A versatile semi-supervised training method for neural networks. In *Proc. CVPR*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proc. ACL*.

Geert Heyman, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-level representations. In *Proc. EACL*.

Adam Kilgarriff. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. EMNLP*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. ICLR*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. MT Summit*.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proc. ACL*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proc. LREC*.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proc. ACL*.

Pranava Swaroop Madhyastha and Cristina España Bohnet. 2017. Learning bilingual projections of embeddings for vocabulary expansion in machine translation. In *Proc. RepL4NLP*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proc. ACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proc. ICLR*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Fabrizio Sebastiani. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proc. SemEval*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense Word Embeddings by Orthogonal Transformation. In *Proc. NAACL-HLT*.

Xuewei Tang and Xiaojun Wan. 2014. Learning bilingual embedding model for cross-language sentiment classification. In *Proc. WI-IAT*.

Ivan Vulić and Anna Korhonen. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proc. ACL*.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proc. ACL*.

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

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proc. ACL*.

Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *Proc. EMNLP*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proc. NAACL-HLT*.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated Neural Networks for Targeted Sentiment Analysis. In *Proc. AAAI 2016*.

Guangyou Zhou, Tingting He, and Jun Zhao. 2014. Bridging the Language Gap: Learning Distributed Semantics for Cross-Lingual Sentiment Classification. In *Proc. NLPCC*.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proc. ACL*.

Xinjie Zhou, Xianjun Wan, and Jianguo Xiao. 2016. Cross-lingual sentiment classification with bilingual document representation learning. In *Proc. ACL*.