Learning Bilingual Sentiment-Specific Word Embeddings without Cross-lingual Supervision

Yanlin Feng and **Xiaojun Wan**
Institute of Computer Science and Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{fengyanlin,wanxiaojun}@pku.edu.cn

Abstract

Word embeddings learned in two languages can be mapped to a common space to produce Bilingual Word Embeddings (BWE). Unsupervised BWE methods learn such a mapping without any parallel data. However, these methods are mainly evaluated on tasks of word translation or word similarity. We show that these methods fail to capture the sentiment information and do not perform well enough on cross-lingual sentiment analysis. In this work, we propose UBiSE (Unsupervised Bilingual Sentiment Embeddings), which learns sentiment-specific word representations for two languages in a common space without any cross-lingual supervision. Our method only requires a sentiment corpus in the source language and pretrained monolingual embeddings of both languages. We evaluate our method on three language pairs for cross-lingual sentiment analysis. Experimental results show that our method outperforms previous unsupervised BWE methods and even supervised BWE methods. Our method succeeds for a distant language pair English-Basque.

1 Introduction

Lack of annotated corpora degrades the quality of sentiment analysis in low-resource languages. Cross-lingual sentiment analysis tackles this problem by adapting the sentiment resource in a resource-rich language (the source language) to a resource-poor language (the target language).

Bilingual Word Embeddings (BWE) provide a way to transfer the sentiment information from the source language to the target language. There has been an increasing interest in BWE methods in recent years, including both supervised methods and unsupervised methods. Supervised BWE methods map the word vectors of the two languages in a common space by exploiting either a bilingual seed dictionary or other parallel data, while unsupervised BWE methods do not utilize any form of bilingual supervision. Yet, these methods are mostly evaluated on tasks of word translation or word similarity, and do not perform well enough on cross-lingual sentiment analysis as shown in Section 4.

Consider the case where we want to perform sentiment analysis on the target language with merely an annotated sentiment corpus in the source language. We assume pretrained monolingual embeddings of both languages are available to us. One solution is to first align the embeddings of both languages in a common space using unsupervised BWE methods, then train a classifier based on the source sentiment corpus. In this solution, no sentiment information is utilized to learn the alignment.

In this paper, we propose to exploit the sentiment information and learn sentiment-specific alignment. The sentiment information is gradually incorporated into the BWE through an iterative constraint relaxation procedure. Unlike previous work which performed alignment in a single direction by linearly mapping the source vectors to the target vector space, we propose an alignment model that maps the vectors of the two languages to a new shared space with two non-linear transformations. Our model is able to separate positive vectors from negative vectors in the bilingual space and allow such sentiment information to be transferred to the target language. Our main contributions are as follows:

1. We propose a novel approach to learn bilingual sentiment-specific word embeddings without any cross-lingual supervision and perform cross-lingual sentiment analysis with minimum resource requirement. We propose an iterative constraint relaxation pro-

Proceedings of NAACL-HLT 2019, pages 420–429
Minneapolis, Minnesota, June 2 - June 7, 2019.

cedure that gradually incorporates the sentiment information into the BWE. Our proposed approach achieves state-of-the-art results.

2. We introduce a novel sentiment-specific objective without having to explicitly build a classifier. Our approach is more explainable and better balances sentimental similarity and semantic similarity compared to previous approaches.

3. We introduce an alignment-specific objective and a simple re-normalization trick. Unlike previous BWE methods that learn orthogonal mappings, we introduce non-orthogonal mappings which enable the transfer of sentiment information from the source language to the target language.

2 Related Work

Cross-Lingual Sentiment Analysis Existing approaches for cross-lingual sentiment analysis can be mainly divided into two categories: (i) approaches that rely on machines translation (MT) systems (ii) approaches that rely on cross-lingual word embeddings.

Standard MT-based approaches perform cross-lingual sentiment analysis by translating the sentiment data into a selected language (e.g. English). More sophisticated algorithms including co-training (Wan, 2009; Demirtas and Pechenizkiy, 2013) and multi-view learning (Xiao and Guo, 2012) have been shown to improve performance.

Zhou et al. (2015, 2016b,a) performed cross-lingual sentiment analysis by learning bilingual document representations. These methods translate each document into the other language and enforce a bilingual constraint between the original document and the translated version.

Bilingual Word Embeddings Word embeddings trained separately on two languages can be aligned in a shared space to produce Bilingual Word Embeddings (BWE), which support many NLP tasks including machine translation (Lample et al., 2017), cross-lingual sentiment analysis (Barnes et al., 2018; Zhou et al., 2015) and cross-lingual dependency parsing (Guo et al., 2015). BWE can be obtained in a supervised way using a seed dictionary (Joulin et al., 2018; Artetxe et al., 2016), or in an unsupervised way without any bilingual data. Adversarial training was the first successful attempt to learn unsupervised BWE (Zhang et al., 2017; Conneau et al., 2017). Self-learning was proposed by (Artetxe et al., 2017)to learn BWE with minimum bilingual resources, which was later extended into a fully unsupervised framework by adding an unsupervised dictionary initialization step (Artetxe et al., 2018).

Multilingual Word Embeddings BWE methods can be extended to the case of multiple languages by simply mapping all the languages to the vector space of a selected language. However, directly learning multilingual word embeddings (MWE) in a shared space has been shown to improve performance (Ammar et al., 2016; Duong et al., 2017; Chen and Cardie, 2018; Alaux et al., 2018). Yet, all these approaches are mainly evaluated on word translation and their effectiveness on cross-lingual sentiment analysis have not been empirically compared.

Sentimental Embeddings Continuous word representations encode the syntactic context of a word but often ignore the information of sentiment polarity. This drawback makes them hard to distinguish words with similar syntactic context but opposite sentiment polarity (e.g. *good* and *bad*), resulting in unsatisfactory performance on sentiment analysis. Tang et al. (2014) learned word representations that encode both syntactic context and sentiment polarity by adding an objective to classify the polarity of an n-gram. This method can be generalized to the cross-lingual setting by training monolingual sentimental embeddings on both languages then aligning them in a common space. However, it requires sentiment resources in the target language thus is impractical for low-resource languages.

There are also approaches to learn sentimental embeddings in the bilingual space without any sentiment resources in the target language. Barnes et al. (2018) jointly minimized an alignment objective based on a seed dictionary, and a classification objective based on the sentiment corpus. Its performance is compared to our method in Section 4. Xu and Wan (2017) learned multilingual sentimental embeddings by extending the BiSkip model (Luong et al., 2015). However, their method does not apply to pretrained embeddings and requires large-scale parallel corpora thus is not

included in our experiments.

3 Proposed Method

3.1 The Overall Framework

This subsection first introduces the proposed mappings for aligning the monolingual embeddings in the bilingual space, then describes the general self-learning algorithm used to learn these bilingual mappings. The details of our algorithm are explained in Section 3.2 - Section 3.6.

3.1.1 The Alignment Model

We assume we have normalized monolingual embeddings $\mathbf{S} \in \mathbb{R}^{v \times d}$ and $\mathbf{T} \in \mathbb{R}^{v \times d}$, where the i-th row of $\mathbf{S}$ is the vector representation of word i in the source language. The normalization procedure is as follows: (i) l_2-normalize each vector (ii) center the vectors (iii) l_2-normalize each vector again (Artetxe et al., 2018).

Given these monolingual embeddings, existing BWE methods typically learn a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ from the source vector space to the target vector space. However, these methods are unsuitable in our setting for two reasons: (i) most methods constrain $\mathbf{W}$ to be orthogonal or near-orthogonal, thus preserving distances between word vectors; (ii) word vectors in the target language space remain unchanged. These two properties prevent us from separating words with opposite sentiment polarity in the bilingual space. In this work, we propose to align the monolingual embeddings with two non-linear mappings:

$$f_s(\mathbf{x}) = \frac{\mathbf{W}_s\mathbf{x}}{\|\mathbf{W}_s\mathbf{x}\|} \qquad f_t(\mathbf{x}) = \frac{\mathbf{W}_t\mathbf{x}}{\|\mathbf{W}_t\mathbf{x}\|}$$

where $\|\cdot\|$ denotes the l_2-norm, $\mathbf{W}_s$ ($\mathbf{W}_t$) is the projection matrix for the source (target) embeddings, and $\mathbf{x}$ is a d-dimension word vector. Each mapping can be seen as a linear projection followed by a re-normalization step.

We propose the following convex domain $\mathcal{D} = \{\mathbf{W} \in \mathbb{R}^{d \times d} \mid \|\mathbf{W}\|_2 \leq r\}$ as an alternative for the orthogonal constraint, where $\|\cdot\|_2$ denotes the spectral norm and r is a hyperparameter that determines to what extent we want to preserve word distances. This is inspired by the unit spectral norm constraint proposed by (Joulin et al., 2018).

3.1.2 The Self-learning Procedure

Given a bilingual seed dictionary, we can learn the projection matrices $\mathbf{W}_s$ and $\mathbf{W}_t$ by forcing the word pairs in the dictionary to have similar representations in the bilingual space. In the unsupervised case, such a dictionary can be induced from the monolingual embeddings $\mathbf{S}$ and $\mathbf{T}$ (Artetxe et al., 2018). However, the quality of this dictionary is usually not good, which in turn degrades the quality of the projection matrices learned from this dictionary. Previous work (Artetxe et al., 2017, 2018) showed that an iterative self-learning procedure can induce a good bilingual dictionary and hence good projection matrices. Given an initial dictionary D_{bi}, this procedure iterates over two steps: (i) it aligns the monolingual embeddings in a common space based on D_{bi}, yielding $\mathbf{S}'$ and $\mathbf{T}'$; (ii) it computes a new dictionary D_{bi} using nearest neighbour retrieval over the approximately aligned embeddings $\mathbf{S}'$ and $\mathbf{T}'$.

In our method, there are three objects $\mathbf{W}_s$, $\mathbf{W}_t$ and D_{bi} to update through the self-learning procedure. Thus we iterates over the following three steps:

1. Solve $\mathbf{W}_s$ by minimizing a sentiment-specific objective $\mathcal{L}_s$ over $\mathcal{D}$, as described in Section 3.3;

2. Solve $\mathbf{W}_t$ by minimizing an alignment-specific objective $\mathcal{L}_t$ over $\mathcal{D}$, as described in Section 3.4;

3. Derive a new bilingual dictionary D_{bi} based on $\mathbf{S}' = \mathbf{S}\mathbf{W}_s^\top$ and $\mathbf{T}' = \mathbf{T}\mathbf{W}_t^\top$, as described in Section 3.5.

Re-normalization is applied as a final step after we have obtained $\mathbf{W}_s$ and $\mathbf{W}_t$.

3.2 Preliminaries

3.2.1 Unsupervised Bilingual Dictionary Initialization

The normalized embeddings $\mathbf{S}$ and $\mathbf{T}$ are not aligned along the first axis, i.e., the i-th row of $\mathbf{S}$ does not correspond to the i-th row of $\mathbf{T}$. Therefore, an initial bilingual dictionary is required in order to access the correspondence between the two languages. Following (Artetxe et al., 2018), we first compute the similarity matrices $\mathbf{M}_s = \sqrt{\mathbf{S}\mathbf{S}^\top}$ and $\mathbf{M}_t = \sqrt{\mathbf{T}\mathbf{T}^\top}$, sort them along the second axis and normalize the rows, yielding $\mathbf{M}'_s$ and $\mathbf{M}'_t$. For each row in $\mathbf{M}'_s$, we apply nearest neighbour retrieval over the rows of $\mathbf{M}'_t$ to find its corresponding translation, yielding a dictionary $D_{s \to t} =$

$\{(1, T_{s\to t}(1)), (2, T_{s\to t}(2)), \ldots, (v, T_{s\to t}(v))\}$, where $T_{s\to t}(i)$ is the translation of the source word i. The same procedure is repeated in the other direction, yielding $D_{t\to s}$. The two dictionaries are then concatenated to produce the initial bilingual dictionary $D_{bi} = D_{s\to t} \cup D_{t\to s}$.

3.2.2 Learning Sentiment-Specific Vectors

In order to incorporate the sentiment information into the bilingual word embeddings, we need a set of d-dimension vectors with known sentiment polarity. We propose a neural network based approach to learn these sentiment-specific vectors. Let the training corpus in the source language be $\mathcal{C} = \{(z_1, y_1), (z_2, y_2), \ldots, (z_{|\mathcal{C}|}, y_{|\mathcal{C}|})\}$, where z_i is a text and y_i is its corresponding label. A d-dimension vector with sentiment polarity y_i can be obtained by calculating the weighted average of the word vectors in z_i:

$$\mathbf{h}_i = \frac{\sum_{j\in z_i} \exp(\alpha_j)\mathbf{S}_{j\cdot}}{\sum_{j\in z_i} \exp(\alpha_j)} \quad (1)$$

where $\mathbf{S}_{j\cdot}$ is the vector representation of the word j in the source language (corresponding to the j-th row of $\mathbf{S}$) and α_j is a scalar that scores the importance of word j on the sentiment polarity. α_j is computed by $\alpha_j = \max(\mathbf{A}\mathbf{S}_{j\cdot} + \mathbf{b})$, where $\mathbf{A} \in \mathbb{R}^{h\times d}$ and $\mathbf{b} \in \mathbb{R}^h$ are the parameters to learn. This function can be seen as a convolution layer with h filters followed by a max pooling layer. The number of filters h is set to 4. Each $\mathbf{h}_i$ is then forwarded to a linear classifier to predict the sentiment label y_i.

Once we have trained the model by minimizing the cross-entropy loss, we re-compute $\mathbf{h}_i$ for each training example z_i. We denote the set of vectors (i.e., $\mathbf{h}_i$) with positive labels as $\mathcal{P} = \{\mathbf{h}_1^p, \mathbf{h}_2^p, \ldots, \mathbf{h}_{|\mathcal{P}|}^p\}$ and the set of vectors with negative labels as $\mathcal{N} = \{\mathbf{h}_1^n, \mathbf{h}_2^n, \ldots, \mathbf{h}_{|\mathcal{N}|}^n\}$. In the 4-class setup, we have four sets: $\mathcal{P}$, $\mathcal{N}$, $\mathcal{SP}$ (the set of strongly positive vectors), $\mathcal{SN}$ (the set of strongly negative vectors).

3.3 Solving $\mathbf{W}_s$

Given a set of positive d-dimension vectors $\mathcal{P} = \{\mathbf{h}_1^p, \mathbf{h}_2^p, \ldots, \mathbf{h}_{|\mathcal{P}|}^p\}$ and a set of negative d-dimension vectors $\mathcal{N} = \{\mathbf{h}_1^n, \mathbf{h}_2^n, \ldots, \mathbf{h}_{|\mathcal{N}|}^n\}$ (or four sets in the 4-class setup), our goal is to distinguish the positive vectors from the negative vectors in the bilingual space, i.e., to separate $\mathbf{W}_s\mathbf{h}_i^p$ from $\mathbf{W}_s\mathbf{h}_j^n$ for any pair of i, j.

We introduce a new d-dimension vector $\mathbf{a}^p \in \mathcal{O} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| \le 1\}$ to represent the "positive direction", which is to be learned. In order to separate positive vectors from negative vectors in the bilingual space, we try to make $\mathbf{W}_s\mathbf{h}_i^p$ ($i = 1, \ldots, |\mathcal{P}|$) to be close to $\mathbf{a}^p$ and $\mathbf{W}_s\mathbf{h}_i^n$ ($i = 1, \ldots, |\mathcal{N}|$) to be distant from $\mathbf{a}^p$.

For a given $\mathbf{a}^p$, we first compute ${\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^p$ for $i = 1, 2, \ldots, |\mathcal{P}|$ and denote the set of i with $\lambda|\mathcal{P}|$ smallest values as $\mathcal{Q}_+^p$, where $\lambda \in [0, 1]$ is a hyperparameter[1]. These $\mathbf{W}_s\mathbf{h}_i^p$ are least similar with $\mathbf{a}^p$ (dot product is used as the similarity metric), hence we maximize the average of ${\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^p$ over $\mathcal{Q}_+^p$. Likewise, we denote the set of $i \in \{1, 2, \ldots, |\mathcal{N}|\}$ with $\lambda|\mathcal{N}|$ largest values of ${\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^n$ as $\mathcal{Q}_-^p$. These $\mathbf{W}_s\mathbf{h}_i^n$ are most similar to $\mathbf{a}^p$, hence we minimize the average of ${\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^n$ over $\mathcal{Q}_-^p$. The overall objective is as follows:

$$\begin{aligned}\min_{\substack{\mathbf{W}_s\in\mathcal{D}\\ \mathbf{a}^p\in\mathcal{O}}} \mathcal{L}_s(\mathbf{W}_s, \mathbf{a}^p) &= \mathcal{L}'(\mathbf{W}_s, \mathbf{a}^p, \mathcal{P}, \mathcal{N}) \\ &\triangleq -\frac{1}{\lambda|\mathcal{P}|}\sum_{i\in\mathcal{Q}_+^p} {\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^p \\ &\quad + \frac{1}{\lambda|\mathcal{N}|}\sum_{i\in\mathcal{Q}_-^p} {\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^n \quad (2)\end{aligned}$$

where $\mathcal{D}$ is the convex set defined in Section 3.1.1. The rationale for this objective is that, instead of forcing every $\mathbf{W}_s\mathbf{h}_i^p$ to be close to $\mathbf{a}^p$, we only focus on a fraction of positive vectors that are most distant from $\mathbf{a}^p$, and vice versa for those negative vectors. We observe that this objective can be rewritten as:

$$\begin{aligned}&\min_{\substack{\mathbf{W}_s\in\mathcal{D}\\ \mathbf{a}^p\in\mathcal{O}}} \mathcal{L}_s(\mathbf{W}_s, \mathbf{a}^p) \\ &= \frac{1}{\lambda|\mathcal{P}|}\max_{\mathcal{Q}\in\mathcal{S}_{\lambda|\mathcal{P}|}(|\mathcal{P}|)} -\sum_{i\in\mathcal{Q}} {\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^p \\ &\quad + \frac{1}{\lambda|\mathcal{N}|}\max_{\mathcal{Q}\in\mathcal{S}_{\lambda|\mathcal{N}|}(|\mathcal{N}|)}\sum_{i\in\mathcal{Q}} {\mathbf{a}^p}^\top \mathbf{W}_s\mathbf{h}_i^n \quad (3)\end{aligned}$$

where $\mathcal{S}_{\lambda|\mathcal{P}|}(|\mathcal{P}|)$ represents all subsets of $\{1, 2, \ldots, |\mathcal{P}|\}$ of size $\lambda|\mathcal{P}|$, and $\mathcal{S}_{\lambda|\mathcal{N}|}(|\mathcal{N}|)$ is similarly defined. [2] This formulation shows that

[1] For simplicity, we assume $\lambda|\mathcal{P}|$ is already rounded to an integer.

[2] There is no need to introduce a new vector $\mathbf{a}^n$ to represent the "negative direction" and introduce a new objective, since the new objective is exactly the same after replacing $\mathbf{a}^p$ with $-\mathbf{a}^n$.

both terms of this objective can be seen as a maximum of linear functions of either $\mathbf{W}_s$ or $\mathbf{a}^p$. Therefore, our objective is convex with respect to either $\mathbf{W}_s$ or $\mathbf{a}^p$, thus can be efficiently minimized by using the projected gradient descent algorithm. We first minimize this objective with respect to $\mathbf{a}^p$ over $\mathcal{O}$, then minimize it with respect to $\mathbf{W}_s$ over $\mathcal{D}$.

While this objective is useful in the binary setup, it does not separate a strongly positive vector in $\mathcal{SP}$ from a weakly positive vector in $\mathcal{P}$ (similarly for $\mathcal{SN}$ and $\mathcal{N}$). In order to achieve better performance in the 4-class setup, we adopt the one-versus-rest strategy to write $\mathcal{L}_s$ as the sum of four terms:

$$\begin{aligned}\min_{\substack{\mathbf{W}_s\in\mathcal{D}\\ \mathbf{a}^p\in\mathcal{O}\\ \mathbf{a}^{sp}\in\mathcal{O}\\ \mathbf{a}^n\in\mathcal{O}\\ \mathbf{a}^{sn}\in\mathcal{O}}} & \mathcal{L}_s(\mathbf{W}_s,\mathbf{a}^p,\mathbf{a}^{sp},\mathbf{a}^n,\mathbf{a}^{sn})\\ &= \mathcal{L}'(\mathbf{W}_s,\mathbf{a}^p,\mathcal{P},\mathcal{N}\cup\mathcal{SP}\cup\mathcal{SN})\\ &+ \mathcal{L}'(\mathbf{W}_s,\mathbf{a}^{sp},\mathcal{SP},\mathcal{P}\cup\mathcal{N}\cup\mathcal{SN})\\ &+ \mathcal{L}'(\mathbf{W}_s,\mathbf{a}^n,\mathcal{N},\mathcal{P}\cup\mathcal{SP}\cup\mathcal{SN})\\ &+ \mathcal{L}'(\mathbf{W}_s,\mathbf{a}^{sn},\mathcal{SN},\mathcal{P}\cup\mathcal{SP}\cup\mathcal{N}) \quad (4)\end{aligned}$$

where $\mathcal{L}'$ is defined in Eq.(2) and $\mathbf{a}^c$ is a d-dimension vector representing the "direction" of class c.

3.4 Solving $\mathbf{W}_t$

Based on the current bilingual dictionary D_{bi}, we construct two sets of vectors $\{\mathbf{x}^s_1,\mathbf{x}^s_2,\ldots,\mathbf{x}^s_{2v}\}$ and $\{\mathbf{x}^t_1,\mathbf{x}^t_2,\ldots,\mathbf{x}^t_{2v}\}$, where $\mathbf{x}^s_i$ and $\mathbf{x}^t_i$ are the vector representations of the i-th word pair in D_{bi}. With $\mathbf{W}_s$ fixed, we can solve $\mathbf{W}_t$ by minimizing:

$$\min_{\mathbf{W}_t\in\mathcal{D}} \mathcal{L}_t(\mathbf{W}_t) = \sum_{i=1}^{2v} \|\mathbf{W}_s\mathbf{x}^s_i - \mathbf{W}_t\mathbf{x}^t_i\|^2 \quad (5)$$

where $\mathcal{D}$ is the convex set defined in Section 3.1.1. This objective is convex with respect to $\mathbf{W}_t$, thus can be minimized efficiently by using the projected gradient descent algorithm.

3.5 Bilingual Dictionary Induction

Once we have computed $\mathbf{W}_s$ and $\mathbf{W}_t$, we can obtain the aligned embeddings $\mathbf{S}' = \mathbf{S}\mathbf{W}_s^\top$ and $\mathbf{T}' = \mathbf{T}\mathbf{W}_t^\top$. Then we induce a new dictionary D_{bi} using nearest neighbour retrieval over the rows of $\mathbf{S}'$ and $\mathbf{T}'$. We perform the induction in two directions to produce $D_{s\to t}$ and $D_{t\to s}$, then concatenate them to produce D_{bi}.

In this work, we propose a modified version of CSLS(Conneau et al., 2017) to be used as the similarity metric to preform nearest neighbour retrieval:

$$\begin{aligned}\text{CSLS}'(\mathbf{x},\mathbf{y}) &= \mathbf{x}^\top\mathbf{y}\\ &- \frac{1}{k}\sum_{\mathbf{y}'\in\mathcal{N}_Y(\mathbf{x})}\mathbf{x}^\top\mathbf{y}' - \frac{1}{k}\sum_{\mathbf{x}'\in\mathcal{N}_X(\mathbf{y})}\mathbf{x}'^\top\mathbf{y} \quad (6)\end{aligned}$$

where $\mathcal{N}_Y(\mathbf{x})$ is the set of k nearest neighbours of $\mathbf{x}$ in the set of vectors Y (in our case Y is the set of rows of $\mathbf{T}'$). We set k to 10 following the original paper.

3.6 Iterative Constraint Relaxation

As mentioned in Section 3.1.1, we introduce a hyperparameter r to define the convex domain $\mathcal{D}$. There is a trade-off to make for r: a large r better incorporates sentimental similarity but significantly harms the quality of the alignment, while a small r constrains $\mathbf{W}_s$ to be near-orthogonal thus prevents it to capture the sentimental similarity.

In order to address this problem, we propose to first set r to 1, letting the the monolingual embeddings to be properly aligned. Then r is iteratively increased by Δr, causing the positive vectors in the bilingual space to be gradually moved further away from the negative vectors. The training process stops when r reaches a maximum value r_{max}, where r_{max} is a hyperparameter [3].

The pseudo code of UBISE in the binary setup is shown in Algorithm 1. For the 4-class UBISE, lines 3,6,7 are replaced by their counterparts in the 4-class setup.

4 Experiments

4.1 Datasets

We use the multilingual sentiment dataset provided by (Barnes et al., 2018). It contains annotated hotel reviews in English (EN), Spanish (ES), Catalan (CA) and Basque (EU). In our experiment, we use EN as the source language and ES, CA, EU as the target languages. For each target language, the dataset is divided into a target development set and a target test set. We also combine the strong and weak labels to produce a binary setup.

[3] Although having the number of iterations be implicitly defined by r_{max} and Δr makes choosing a small r_{max} impractical, it allows us to tune r_{max} in a single training process.

Algorithm 1 binary UBISE

Input: $\lambda, r_{max}, \Delta r, \mathbf{S}, \mathbf{T}, \mathcal{C}$
Output: $\mathbf{S}', \mathbf{T}', \mathbf{W}_s, \mathbf{W}_t$

$r \leftarrow 1$
Initialize $\mathbf{W}_s$ and $\mathbf{W}_t$ to identity matrices
Learn $\mathcal{P}$, $\mathcal{N}$ from $\mathbf{S}$ and $\mathcal{C}$, according to Section 3.2.2
Compute the initial bilingual dictionary D_{bi} from $\mathbf{S}$ and $\mathbf{T}$, according to Section 3.2.1
while $r \leq r_{max}$ **do**
  $\mathbf{a}^p \leftarrow \text{argmin}_{\mathbf{a}^p \in \mathcal{O}} \mathcal{L}_s(\mathbf{a}^p, \mathbf{W}_s)$
  $\mathbf{W}_s \leftarrow \text{argmin}_{\mathbf{W}_s \in \mathcal{D}} \mathcal{L}_s(\mathbf{a}^p, \mathbf{W}_s)$
  $\mathbf{W}_t \leftarrow \text{argmin}_{\mathbf{W}_t \in \mathcal{D}} \mathcal{L}_t(\mathbf{W}_t)$
  $\mathbf{S}' \leftarrow \mathbf{S}\mathbf{W}_s^\top$
  $\mathbf{T}' \leftarrow \mathbf{T}\mathbf{W}_t^\top$
  Derive a new bilingual dictionary D_{bi} from $\mathbf{S}'$ and $\mathbf{T}'$, according to Section 3.5
  $r \leftarrow r + \Delta r$
end while
Normalize the rows of $\mathbf{S}'$, $\mathbf{T}'$ to unit length
return $\mathbf{S}', \mathbf{T}', \mathbf{W}_s, \mathbf{W}_t$

The normalized 300-dimension fastText vectors (Bojanowski et al., 2017) are used by all methods.

The MUSE dataset(Conneau et al., 2017) is used by approaches that require bilingual supervision [4]. Each dictionary contains 5000 unique source words.

4.2 Implementation details

We empirically set $\Delta r = 0.01$ and $v = 10000$. The vocabulary of each language is limited to the v most frequent words so that the embedding matrix has shape $v \times d$. Hyper parameters λ and r_{max} are tuned on the target development set via a grid search. We apply stochastic dictionary induction by randomly setting the elements of the similarity matrix used for nearest neighbour retrieval to zero with probability $1 - p$, as described in (Artetxe et al., 2018). p is initialized to 0.1 and increased by 0.005 at each iteration. We empirically stop updating the dictionary when r exceeds 3.

4.3 Baselines

We compare our method with the following baselines, including state-of-the-art BWE methods that are originally evaluated on the word translation task, as well as bilingual sentimental embeddings methods that are optimized for cross-lingual sentiment analysis. The bilingual word embeddings learned by each method are later evaluated on cross-lingual sentiment analysis using the same classifier for fairness.

4.3.1 Unsupervised BWE Methods

ADVERSARIAL Conneau et al. (2017) proposed an unsupervised BWE method based on adversarial training. After a near-orthogonal projection matrix is learned through adversarial training, a refinement procedure is applied to improve the quality of the alignment.

VECMAP Artetxe et al. (2018) proposed an unsupervised BWE learning framework. It consists of an unsupervised dictionary initialization step and the self-learning procedure mentioned in Section 3.1.2.

4.3.2 Supervised BWE Methods

PROCRUSTES Artetxe et al. (2016) proposed a simple and effective supervised BWE method that requires a seed dictionary. It computes the optimal projection matrix by taking singular value decomposition (SVD).

RCSLS Joulin et al. (2018) proposed an supervised BWE method that also requires a seed dictionary. They proposed a training objective that is consistent with the retieval criterion that can be minimized by using gradient descent. It achieves state-of-the-art results on the word translation task.

4.3.3 Bilingual Sentimental Embedding Methods

BLSE Barnes et al. (2018) exploited both bilingual supervision and the sentiment corpus to learn bilingual sentimental embeddings. They jointly minimize an alignment-specific objective and a classification objective to learn the projection matrices. The trade-off between the two objectives is controlled by a hyperparameter $\alpha \in [0, 1]$. We tune α on the target development set as described in the original paper. Once the projection matrices have been learned, the classifier in this model is abandoned. The quality of the resulting BWE is evaluated using the classifier mentioned in Section 4.4.

4.4 Evaluation

We use DAN (Iyyer et al., 2015) as the classifier to preform cross-lingual sentiment analysis. The loss

[4] This dataset does not contain a dictionary for EN-EU, thus we translate the EN-ES dictionary into EN-EU

of each instance is weighted by its inverse class frequency to address the class imbalance problem. For each method, the dropout rate is fixed at 0.3 and the l_2-regularization strength is tuned on the target development set[5]. We train five classifiers for each method and report the average macro-F1 on the target test set.

4.5 Results and Analysis

Table 1 presents the results of different BWE methods. UBISE outperforms all unsupervised methods on all six tasks and outperforms all baselines on four out of six tasks.

All methods, especially unsupervised methods, suffer from distant language pairs, which is consistent with the observation of (Søgaard et al., 2018). VECMAP and ADVERSARIAL perform significantly worse on EN-EU compared to supervised methods. Yet, UBISE outperforms the strongest baseline by 2.1% on EN-EU, indicating that incorporating sentiment is vital for cross-lingual sentiment analysis on distant languages.

Despite the similar performance across different BWE methods in the binary setup, UBISE outperform all baselines in the 4-class setup by a large margin (average of +2.2%). This may indicate that the original monolingual embeddings are able to distinguish positive words from negative words(e.g., *good* and *bad*), but bad at distinguishing strongly positive words from weakly positive words (e.g., *good* and *perfect*).

The performance of BLSE is merely comparative with other baselines.[6] We suspect that this is due to the classifier we use to perform cross-lingual sentiment analysis. The original paper used SVM or logistic regression to perform classification, in which case BLSE achieved better performance due to the utilization of sentiment information. But if we use a deeper neural network to perform cross-lingual sentiment analysis, preserving the original semantic similarity is more important. A qualitative comparison between BLSE and UBISE is presented in Section 4.8.

4.6 Effect of the Sentiment Information

We perform an ablation test to demonstrate the effect of the sentiment information provided by $\mathcal{L}_s$. We create a new model UBISE_MIN that does not utilize the sentiment information by eliminating lines 6,7,12 in Algorithm 1. UBISE_MIN runs 500 iterations for every language pairs.

The comparative results in Table 2 show that utilizing the sentiment information leads to an average improvement of +3.1% in the binary setup or +4.1% in the 4-class setup.

4.7 Effect of Re-Normalization

Re-normalization is useful in the sense that it leads to better alignment by constraining all the bilingual vectors to be on the unit sphere. While this property does not matter for word translation as long as cosine-similarity is used as the retrieval criterion, it matters for cross-lingual sentiment analysis. Another effect of re-normalization is that it introduces non-linearity between the linear projection and the classifier, which is vital for separating words with opposite sentiment polarity. Without non-linearity the linear projection and the first layer of the classifier would collapse into a single linear projection, thus eliminating the effect of $\mathbf{W}_s$. Figure 2 illustrates how this non-linearity helps separating positive words from negative words in the bilingual space. This effect is demonstrated in Section 4.8.

4.8 Visualization of the Bilingual Space

To illustrate how UBISE transfers sentiment information from the source language to the target language, we visualize six categories of words in the bilingual space of UBISE and BLSE using t-SNE (Maaten and Hinton, 2008). As shown in Figure 1, both methods manage to separate positive words from negative words without any annotated data in Spanish. However, Barnes et al. (2018) abandon the original semantic similarity, which degrades its performance as shown in Section 4.5. In contrast, our method preserves semantic similarity by limiting the largest singular values of $\mathbf{W}_s$ and $\mathbf{W}_t$ to be smaller than r_{max}. The trade-off between semantic similarity and sentimental similarity is made by choosing an appropriate r_{max}.

5 Conclusion

This paper presents a method to learn bilingual sentiment-specific word embeddings without any cross-lingual supervision. We propose a novel sentiment-specific objective that separates words

[5]The optimal regularization strength depends on the BWE method. Stronger regularization is favourable to BLSE and UBISE.

[6]We already obtain significantly better results after replacing the original classifier with DAN, compared with the original reported results.

Bilingual Supervision	Method	Binary ES	Binary CA	Binary EU	4-class ES	4-class CA	4-class EU
$5k$ dict.	PROCRUSTES	80.4	**83.1**	74.1	49.1	50.9	43.0
	RCSLS	**80.7**	81.4	73.4	50.3	47.8	41.3
	BLSE	80.2	82.2	73.5	50.0	47.0	35.1
None	VECMAP	80.0	80.2	69.2	51.2	52.2	38.8
	ADVERSARIAL	79.8	79.9	60.3	45.8	47.9	31.0
None	UBISE	80.5	80.4	**76.7**	**54.4**	**54.1**	**44.6**

Table 1: Macro F1 of different BWE approaches. The best score for each language pair is shown in **bold**. The best score among unsupervised BWE methods for each language pair is underlined.

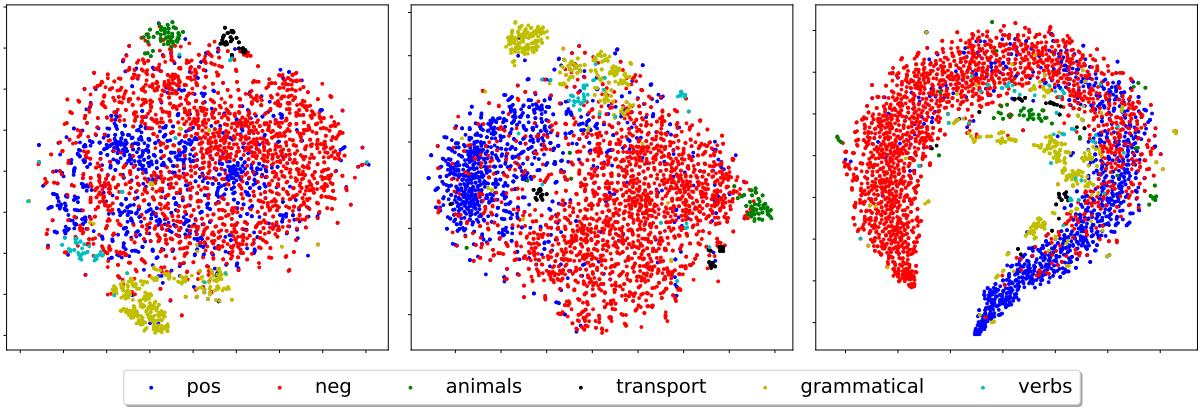
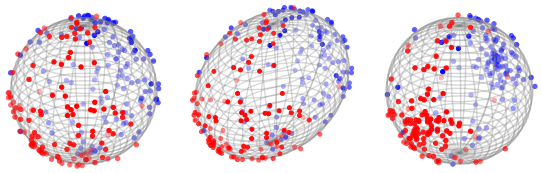


Figure 1: t-SNE Visualization of the Spanish word vectors. Grammatical words include pronouns, prepositions, articles, conjunctions, etc. *left*: original normalized vectors; *middle*: the bilingual space of binary UBISE with $\lambda = 0.5$ and $r_{max} = 5.5$; *right*: the bilingual space of BLSE.

Setup	Method	ES	CA	EU
Binary	UBISE_MIN	77.4	80.2	70.8
	UBISE	80.5	80.4	76.7
4-class	UBISE_MIN	51.7	49.3	39.9
	UBISE	54.4	54.1	44.6

Table 2: Comparison between UBISE and UBISE_MIN

.



(a) (b) (c)

Figure 2: Illustration of the effect of re-normalization. (a) the original normalized embeddings (b) embeddings after linear projection (c) embeddings after re-normalization

with opposite sentiment polarity in the bilingual space, and an alignment objective that enables the transfer of sentiment information from the source language to the target language. An iterative constraint relaxation procedure is applied to gradually incorporate the sentiment information into the bilingual word embeddings. We empirically evaluate our method on three language pairs for cross-lingual sentiment analysis and demonstrate its effectiveness. Experimental results show that incorporating sentiment information significantly improves the performance on fine-grained cross-lingual sentiment analysis.

Acknowledgment

This work was supported by National Natural Science Foundation of China (61772036) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings. *arXiv preprint arXiv:1811.01124*.

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

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

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

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. 2018. Bilingual sentiment embeddings: Joint projection of sentiment across languages. *arXiv preprint arXiv:1805.09016*.

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

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. *arXiv preprint arXiv:1808.08933*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Erkin Demirtas and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 9. ACM.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 894–904.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1234–1244.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1681–1691.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1555–1565.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-volume 1*, pages 235–243. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2012. Multi-view adaboost for multilingual subjectivity analysis. *Proceedings of COLING 2012*, pages 2851–2866.

Kui Xu and Xiaojun Wan. 2017. Towards a universal sentiment classifier in multiple languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 511–520.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1959–1970.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 430–440.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 247–256.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1403–1412.