Towards a Universal Sentiment Classifier in Multiple languages

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

Abstract

Existing sentiment classifiers usually work for only one specific language, and different classification models are used in different languages. In this paper we aim to build a universal sentiment classifier with a single classification model in multiple different languages. In order to achieve this goal, we propose to learn multilingual sentiment-aware word embeddings simultaneously based only on the labeled reviews in English and unlabeled parallel data available in a few language pairs. It is not required that the parallel data exist between English and any other language, because the sentiment information can be transferred into any language via a pivot languages. We present the evaluation results of our universal sentiment classifier in five languages, and the results are very promising even when the parallel data between English and the target languages are not used. Furthermore, the universal single classifier is compared with a few cross-language sentiment classifiers relying on direct parallel data between the source and target languages, and the results show that the performance of our universal sentiment classifier is very promising compared to that of different cross-language classifiers in multiple target languages.

1 Introduction

Nowadays, a large amount of user-generated content (UGC) appears online everyday, such as tweets, comments and product reviews. Sentiment classification on these data has become a popular research topic over the past few years (Pang et al., 2002; Blitzer et al., 2007; Agarwal et al., 2011; Liu, 2012). Distributed representations of words or word embeddings have been widely explored, and have proved its great usability for the sentiment classification task (Tang et al., 2014; Zhou et al., 2015; Xu et al., 2015; Bollegala et al., 2016; Ferreira et al., 2016).

Most existing sentiment classifiers rely on labeled training data and the data are usually language-dependent. In other words, a sentiment classifier is learned from a labeled dataset in a specific language and this sentiment classifier can be used for sentiment classification in this language. However, labeled training data for sentiment classification are not available or not easy to obtain in many languages in the world (e.g., Malaysian, Mongolian, Uighur). Without reliable labeled data, it is hard to build a sentiment classifier in these resource-poor languages.

Fortunately, there are a few studies investigating the task of cross-language sentiment classification (Banea et al., 2008; Wan, 2009; Meng et al., 2012; Xiao and Guo, 2013; Gao et al., 2015; Chen et al., 2015; Zhou et al., 2015; Li et al., 2017; Zhou et al., 2016a,b), which aims to make use of the labeled data in a source language (English in most cases) to build a sentiment classifier in a target language. However, cross-language sentiment classification methods rely on parallel data between the source and target languages[1] In a resource-poor language, the parallel data between this language and the source language may not be available or is not easy to obtain. In this circumstance, previous cross-language sentiment classification meth-

[1]Note that a few methods rely on a machine translation system to produce parallel data between the two languages, while the machine translation system is built on a large amount of parallel data between the two languages. In this sense, the methods rely on both the parallel data for machine translation and the pseudo parallel data produced by machine translation systems.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 520–529
Copenhagen, Denmark, September 7–11, 2017.

ods will fail to work.

Another shortcoming of previous cross-language sentiment classification researches is that we have to build an individual cross-language sentiment classifier for each target language, even when we want to perform sentiment classification in a couple of languages at the same time.

In this study, instead of building a sentiment classifier for each target language, we aim to build a universal sentiment classifier in multiple languages and this universal sentiment classifier only learns one single sentiment classification model and it can be applied for sentiment classification in many languages.

In order to achieve this goal, we propose an approach to learn multilingual sentiment-aware word embeddings simultaneously based only on the labeled reviews in English and unlabeled parallel data available in a few language pairs. As mentioned earlier, in some resource-poor languages, there do not exist direct parallel data between these languages and the source English language. In order to address this problem, we propose a pivot-based model to transfer the sentiment information from the source language to any resource-poor language via pivot languages. Finally, a universal sentiment classifier can be built because the multilingual word embeddings are in the same semantic space.

We build three different models (Bilingual Model, Pivot-Driven Bilingual Model and Universal Multilingual Model) and compare them empirically in order to answer two questions in this paper: 1) Can pivot-based models learn bilingual sentiment-aware word embeddings effectively? 2) Can an effective universal sentiment classifier be built for multiple languages?

Without loss of generality, we present and compare the evaluation results of the models in five languages. Evaluation results show that pivot-driven bilingual models perform as well as the bilingual model using direct parallel data, which lays the solid foundation of our universal model. Moreover, it is very promising that our universal sentiment classifier can work well in five languages, and it can achieve very promising classification results as compared to several typical cross-language sentiment classification models.

The main contributions of our study in this paper are summarized as follows:

- We are the first to build a universal sentiment classifier in multiple languages by learning multilingual sentiment-aware word embeddings, which can not be addressed by previous researches on cross-language sentiment classification.
- We propose pivot-based models to bridge two languages in which there do not exist parallel data, and thus the sentiment information can be transferred to any target language.
- Evaluation results on five languages demonstrate the efficacy of our proposed pivot-based models and the universal sentiment classifier.

2 Our Approach

In order to build a universal sentiment classifier, we propose an approach to learn multilingual sentiment-aware word embeddings simultaneously, and then train a universal sentiment classification model in the embedding space by averaging the word embeddings in a document as the document representation. Note that in this study, we focus on only using the labeled data in English and do not make use of any labeled data in other languages, which makes the task more challenging[2]. Formally, we aims to build a single sentiment classifier which can perform sentiment classification in many languages $\{S, T_1, T_2, ..., T_N\}$, where S refers to English language, and T_1 to T_N refer to other N languages.

In our approach, the multilingual sentiment-aware word embeddings play the key role in building the universal sentiment classifier, and now the question is how to learn the multilingual sentiment-aware word embeddings? Inspired by previous studies on cross-lingual sentiment classification and bilingual word embedding learning, we can leverage the labeled data in S (i.e., English) and unlabeled parallel data between S and language T to learn bilingual sentiment-aware word embeddings in both English and T languages with a bilingual model. However, such unlabeled parallel data are not always easy to obtain for all other languages. For a specific language T, if the unlabeled parallel data between T and S do not exist, the bilingual model cannot be applied. In order to address this problem, we propose a pivot-driven bilingual model to leverage pivot languages

[2]Note that the labeled data in other languages can be easily used by our approach in the same way as the English labeled data, and we believe more labeled data will eventually improve the performance of the sentiment classifier.

to bridge T and S. We choose a pivot language P where the parallel data between P and S, and the parallel data between P and T are easy to obtain, and then leverage them to learn the multilingual sentiment-aware word embeddings in the three languages P, T and S. Furthermore, we can leverage more parallel data between multiple languages, some of which are parallel data between S and other languages, and some of which are parallel data within other languages, to build an universal multilingual model. The sentiment information will be directly or indirectly transfered to each language as well, and thus we obtain multilingual sentiment-aware word embeddings in many languages.

The bilingual model, pivot-driven bilingual model and universal multilingual model will be described in next sections, respectively.

2.1 Bilingual Model

The bilingual model tries to induce bilingual word embeddings from a parallel corpus, and in the meantime make similar words from the two languages share adjacent vector representations in the same vector space.

Formally, we assume a source language S with $|S|$ words and a target language T with $|T|$ words. We use s and t to represent a word from S and T, respectively. Given the bilingual parallel corpus $\mathcal{C}$ between language S and T, it can be divided into a corpus $\mathcal{C}_\mathcal{S}$ in language S and a corpus $\mathcal{C}_\mathcal{T}$ in language T. And we use a notation $S-T$ to indicate a parallel corpus between languages S and T.

Previous studies have proposed some bilingual models for learning bilingual word embeddings, so we extend the well-behaved BiSkip model (Luong et al., 2015) to Bilingual Model (**BM**). This model requires word alignment information, and in this study word alignment is automatically obtained from parallel sentences by using a word alignment tool.

In our bilingual model, every word s in language S is required to predict the adjacent words of itself and the aligned word t in the target language T. For corpus $\mathcal{C}_\mathcal{S}$, the monolingual constraint on itself ($\mathcal{C}_\mathcal{S} \rightarrow \mathcal{C}_\mathcal{S}$) is:

$$Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{S}) = \sum_{s\in\mathcal{C}_\mathcal{S}} \sum_{w\in adj(s)} \log p(w|s), \quad (1)$$

and the cross-lingual constraint on $\mathcal{C}_\mathcal{T}$ ($\mathcal{C}_\mathcal{S} \rightarrow \mathcal{C}_\mathcal{T}$) is:

$$Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{S}) = \sum_{s\in\mathcal{C}_\mathcal{S}} \sum_{w\in adj(t), s\leftrightarrow t} \log p(w|s) \quad (2)$$

where $s\leftrightarrow t$ means word $s(\in \mathcal{C}_\mathcal{S})$ is aligned to word $t(\in \mathcal{C}_\mathcal{T})$ and $adj(s)$ or $adj(t)$ mean the adjacent words of word s or t.

Similarly, for corpus $\mathcal{C}_\mathcal{T}$ we can obtain:

$$Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{T}) = \sum_{t\in\mathcal{C}_\mathcal{T}} \sum_{w\in adj(t)} \log p(w|t), \quad (3)$$

and

$$Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{T}) = \sum_{t\in\mathcal{C}_\mathcal{T}} \sum_{w\in adj(s), t\leftrightarrow s} \log p(w|t) \quad (4)$$

Combining equations 1, 2, 3 and 4, we get the objective for obtaining bilingual word embeddings from parallel corpus:

$$Obj(\mathcal{C}) = \alpha_1 Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{S}) + \alpha_2 Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{S}) + \alpha_3 Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{T}) + \alpha_4 Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{T})$$

where $\alpha_1, \alpha_2, \alpha_3$ and α_4 are scalar parameters.

We still have to incorporate the sentiment information into the bilingual word embeddings. Similar to previous studies (Zhou et al., 2015), we make use of the sentiment polarity of texts as supervision in the learning process. Given a labeled sentimental corpus $\mathcal{C}_\mathcal{L}$[3], we use S^* to represent a sentence in $\mathcal{C}_\mathcal{L}$ and w as a word in S^*. And x^T is a sum of word embeddings in S^*. We simply adopt the logistic regression classifier to enforce the sentiment constraint, and thus make the bilingual word embeddings absorb the corresponding sentiment information. The objective function is:

$$L(\mathcal{C}_\mathcal{L}) = \sum_{S^*\in\mathcal{C}_\mathcal{L}} y \log\sigma(Wx^T + b) + (1-y)\log\sigma(1-(Wx^T+b)) \quad (5)$$

where y is the label of the sentence S^*, W is a weight vector and b is a bias.

The overall objective function for inducing bilingual sentiment-aware word embeddings is to maximize:

$$Obj(\mathcal{C}) + L(\mathcal{C}_\mathcal{L})$$

[3]Note that the labeled corpus is usually provided in the source language S, which means L is S. but the labeled corpus usually does not overlap with the parallel corpus.

2.2 Pivot-Driven Bilingual Model

For some resource-poor target language T, it is quite expensive to get direct parallel corpus between T and the source language S. Without such parallel corpus, it is not possible to apply the above bilingual model to learn bilingual sentiment-aware word embeddings. In order to address this problem we propose our Pivot-Driven Bilingual Model (**PDBM**) by using a pivot language to bridge T and S. The model is inspired by (Wu and Wang, 2007), in which pivot languages are used for phrase-based SMT. A pivot language P is chosen if the parallel corpus between P and S, and the parallel corpus between P and T are easy to obtain. Given two parallel corpora: S-P and P-T, our PDBM model tries to get the trilingual word embeddings by putting constrains on the two corpora. Under the well-designed constraint, the pivot language P can pass the sentiment information from the source language S to the target language T. Similarly, we further assume the pivot language P with $|P|$ words, and use $\mathcal{C}_\mathcal{P}$ to denote the corpus in language P.

We design constraints on the two parallel corpora S-P and P-T, instead of direct constraints on S and T. Derived from the BM model, we can get three monolingual constraints $\mathcal{C}_\mathcal{S}\rightarrow\mathcal{C}_\mathcal{S}$, $\mathcal{C}_\mathcal{T}\rightarrow\mathcal{C}_\mathcal{T}$, $\mathcal{C}_\mathcal{P}\rightarrow\mathcal{C}_\mathcal{P}$ and four bilingual constraints $\mathcal{C}_\mathcal{S}\rightarrow\mathcal{C}_\mathcal{P}$, $\mathcal{C}_\mathcal{T}\rightarrow\mathcal{C}_\mathcal{P}$, $\mathcal{C}_\mathcal{P}\rightarrow\mathcal{C}_\mathcal{S}$ and $\mathcal{C}_\mathcal{P}\rightarrow\mathcal{C}_\mathcal{T}$. The final objective function for learning the trilingual word embeddings can be summarized as:

$$Obj_p(\mathcal{C}) = \beta_1 Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{S}) + \beta_2 Obj(\mathcal{C}_\mathcal{S}|\mathcal{C}_\mathcal{P})$$
$$+\beta_3 Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{T}) + \beta_4 Obj(\mathcal{C}_\mathcal{T}|\mathcal{C}_\mathcal{P})$$
$$+\beta_5 Obj(\mathcal{C}_\mathcal{P}|\mathcal{C}_\mathcal{S}) + \beta_6 Obj(\mathcal{C}_\mathcal{P}|\mathcal{C}_\mathcal{T})$$
$$+\beta_7 Obj(\mathcal{C}_\mathcal{P}|\mathcal{C}_\mathcal{P})$$

where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are scalar parameters. Similarly, the objective for enforcing the sentiment constraint is the same as equation 5, so we combine them together to get the overall objective function:

$$Obj_p(\mathcal{C}) + L(\mathcal{C}_\mathcal{L})$$

Through the pivot language, the sentiment information can be passed from a source language to a target language by maximizing the above objective function.

2.3 Universal Multilingual Model

The bilingual model and the pivot-driven bilingual model lay the foundations of build a universal multilingual model for sentiment classification in many languages. Given a source language S and a few other languages $\{T_1, T_2, ..., T_N\}$. If there exist parallel data between a language T_i and S, then the bilingual sentiment-aware word embeddings can be learned by the bilingual model. If the parallel data between languages T_i and S are not available, a pivot language can be selected and the pivot-driven model can be applied to learn the trilingual sentiment-aware word embeddings. Even when a single pivot language cannot be found for languages T_i and S, we still can find two or more pivot languages $\{P_1, P_2, ..., P_M\}$ to form a pivot chain and the sentiment information in the source language can be passed through the pivot chain ($S - P_1 - ... - P_M - T_i$) to the target language.

Therefore, in this model, we will make use of all parallel corpora between any pair of languages (including parallel corpora between the source language and any other language, and parallel corpora between other languages) and learn the sentiment-aware word embeddings in all the languages simultaneously. The monolingual objective in each language and the cross-lingual objective for any available parallel corpus are defined in the same way as in the above models, and we sum all the objectives and denote it as $Obj_{universal}(C)$, and this objective is then combined with the sentiment constraint as follows:

$$Obj_{universal}(\mathcal{C}) + L(\mathcal{C}_\mathcal{L})$$

By maximizing the above objective function, the sentiment-aware word embeddings in all the languages will be learned.

3 Evaluations

3.1 Dataset

Without loss of generality, we evaluate our models in five languages (including three western languages and two Asian languages): English (en), German (de), French (fr), Japanese (jp) and Chinese (en/zh). Among these languages, the English language is the source language with labeled training data, and we do no use any labeled data in the other languages.

Particularly, we use the multilingual multi-domain Amazon review dataset [4] provided by (Prettenhofer and Stein, 2010) and the NLPC-C2013 dataset [5]. The review dataset provided by (Prettenhofer and Stein, 2010) contains labeled data in four languages: English, German, French and Japanese, and the NLPCC2013 dataset further provides labeled data in Chinese. The reviews in each language are divided into three domains: *dvd*, *music* and *books*. Each domain of product reviews contains a balanced training set and test set, each of which consists of 1000 positive and 1000 negative reviews for each language except for Chinese. While for Chinese language, the test set consists of 2000 positive and 2000 negative reviews. We only use English training data as the labeled data in the experiments.

We further obtain unlabeled parallel data from Europarl v7 [6] (Koehn, 2004) (*Eu v7*) and The United Nations Parallel Corpus v1.0 [7] (Ziemski et al., 2016) (*UN v1.0*). The Europarl corpus contains bilingual parallel corpus between English and other 20 Europe languages. The United Nations Parallel Corpus is composed of official records and other parliamentary documents of the United Nations that are in the public domain. These documents are mostly available in the six official languages of the United Nations. Besides, we use the *cldc-2009-004* [8] Chinese-English (*CN-EN*) news parallel corpus and *Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles Version 2.01* [9] (*JP-EN*), which is created manually by translating Japanese Wikipedia articles (related to Kyoto) into English. In addition, *CJWikiCorpus* (*CN-JP*) is a Chinese-Japanese Parallel Corpus Constructed from Wikipedia [10]

For the the BM model, we use *en-de* ($\in$ *Eu v7*) and *en-fr* ($\in$ *Eu v7*), *en-zh* ($\in$ *CN-EN*), and *en-jp* ($\in$ *JP-EN*).

For the PDBM model, we use *en-fr* ($\in$ *UN v1.0*) with *fr-de* ($\in$ *Eu v7*) to get the case *en-fr-de* (*fr* acts as a pivot), *en-zh* ($\in$ *CN-EN*) with *zh-jp* ($\in$ *CN-JP*) to build the case *en-zh-jp* (*zh* acts as a pivot), *en-zh* ($\in$ *CN-EN*) with *zh-fr* ($\in$ *UN v1.0*) to build *en-zh-fr* (*zh* acts as a pivot), and *en-fr* ($\in$ *Eu v7*) with *zh-fr* ($\in$ *UN v1.0*) to build *en-fr-zh* (*fr* acts as a pivot). Note that any pivot language can be selected if the parallel corpora between the pivot language and other languages can be obtained, but in our experiments, we only use one pivot language in each test case to validate the feasibility of our proposed model. In practice, a popular language (such as English, Chinese) can be used as the pivot because it can act as a link between two unpopular languages.

While for the UMM model, we use all the corpora used in PDBM to build a universal model. All the details can be found in Table 1.

3.2 Comparison Methods

In addition to the comparison between our models, we further compare them with popular cross-lingual (CL) sentiment classification methods.

For comparison in German, French and Japanese, we adopt a few typical CL classification methods, and the results are directly borrowed from the corresponding published papers:

MT-BOW: It is a simple model to train a linear classifier based on the bag-of-words features, and it uses a machine translator to translate the test data into the source language (Prettenhofer and Stein, 2010) .

CL-SCL: It is the cross-lingual structural correspondence learning algorithm proposed by (Prettenhofer and Stein, 2010) and the features in the two languages are mapped to a unified space.

BSE: It is introduced in (Tang and Wan, 2014) by forcing the representations of words from both the source and target languages to share the same feature space. In this way, bilingual word embeddings are learned for cross-lingual sentiment classification.

CR-RL: It is the bilingual word representation learning method of (Xiao and Guo, 2013). It learns different representations for words in different languages. Part of the word vector is shared among different languages and the rest is language dependent. The document representation is calculated by taking average over all words in the document.

Bi-PV: It extends the paragraph vector model into bilingual setting by sharing the document representation of a pair of parallel documents (Pham et al., 2015).

For comparison in Chinese, we adopt several typical CL classification methods:

MT-LR and **MT-SVM**: We use logistic regres-

[4] https://www.uni-weimar.de/medien/webis/corpora/corpus-webis-cls-10/

[5] http://tcci.ccf.org.cn/conference/2013/pages/page04_evares.html

[6] http://www.statmt.org/europarl/v7/

[7] https://conferences.unite.un.org/UNCorpus/

[8] http://www.chineseldc.org/resource_info.php?rid=141

[9] http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

[10] http://lotus.kuee.kyoto-u.ac.jp/ chu/resource/wiki_zh_ja.tgz

Model	Parallel corpora with size	Test case
BM	*en-de* (∈ *Eu v7*, 1.92M)	*en-de*
	en-fr (∈ *Eu v7*, 2.0M)	*en-fr*
	en-zh (∈ *CN-EN*, 1.0M)	*en-zh*
	en-jp (∈ *JP-EN*, 0.5M)	*en-jp*
PDBM	*en-fr* (∈ *UN v1.0*, 2.0M) + *fr-de* (∈ *Eu v7*, 1.5M)	*en-fr-de*
	en-zh (∈ *CN-EN*, 1.0M) + *zh-jp* (∈ *CN-JP*, 0.12M)	*en-zh-jp*
	en-zh (∈ *CN-EN*, 1.0M) + *zh-fr* (∈ *UN v1.0*, 2.0M)	*en-zh-fr*
	en-fr (∈ *Eu v7*, 2.0M) + *zh-fr* (∈ *UN v1.0*, 2.0M)	*en-fr-zh*
UMM	*all the corpora used in PDBM*	*en,de,fr,zh,jp*

Table 1: Parallel corpora used in our models.

sion and SVM to learn different classifiers based on the translated Chinese training data. Bag of words features are used for classification.

Bi-PV: The same as that described above.

BSWE: It uses the bilingual sentiment word embedding algorithm based on denoising autoencoders (Zhou et al., 2015) to learns word representations. Each document is then represented by the sentiment words and the corresponding negation words.

3.3 Settings and Preprocessing

We utilize *cdec* (Dyer et al., 2010) as an alignment tool to get word-level alignment, and we also use it to lowercase the characters in western languages. We use the *stanford-segmenter* [11] to segment Chinese words, and use *Mecab* [12] to segment Japanese words. The *SnowNLP* [13] is used to convert traditional words to simplified ones. Besides, we remove all the irregular characters (e.g., ©, £, ♡) in the texts.

For all the three models, we use stochastic gradient descent (SGD) for learning, with a default learning rate of 0.025, negative sampling with 30 samples, skip-gram with context window of size 5, and a subsampling rate of value 1e-4. The embedding size is set to 400. The training epochs are all set to 10. All the parameters of α and β used in the three models are simply set to 1. The word embeddings in a document are averaged to get the document representation, and then the logistic regression classier is adopted for sentiment classification.

[11] http://nlp.stanford.edu/software/segmenter.shtml
[12] http://taku910.github.io/mecab/
[13] https://github.com/isnowfy/snownlp

3.4 Results

The sentiment classification results of our three models and the CL classification methods in the three domains and in the German, French and Japanese languages are presented in Table 2. The results in the Chinese language are presented in Table 3. Note that the results of the CL methods are not reported on English test sets, and we only compare our three models on English test sets in Table 3.

First and most importantly, we compare our three models. The BM model relies on the direct parallel data between the source and target languages, and it generally works slightly better than the other models, including the PMDB model and the UMM model. The reason is that direct parallel data can be used for transferring the sentiment information from the source language to the target language directly. However, the performance achieved by the PDBM model is very close to the BM model in most test cases. In some cases (DE-DVD, JP-book and EN-music), the PDBM model can even outperform the BM model. Note that the PDBM model does not leverage the direct parallel data between the source and target languages, but uses a pivot language as a bridge. The results demonstrate that the pivot-driven model is very effective for learning bilingual / trilingual sentiment-aware word embeddings. The results also verify the feasibility of using pivot languages to address the problem of sentiment classification in resource-poor languages, which lays a good foundation for building a universal sentiment classifier in multiple languages. When comparing the UMM model with BM and PDBM, the results of UMM are very close to that of BM and PDBM in most cases, Note that the UMM model does not use the direct parallel corpora of *en-de*

TL	Domain	BM	PDBM	UMM	MT-BOW	CL-SCL	BSE	CR-RL	Bi-PV
DE	book	82.46	81.97	81.65	79.68	79.50	80.27	79.89	79.51
	DVD	81.47	82.67	81.27	77.92	76.92	77.16	77.14	78.60
	music	82.95	81.93	81.32	77.22	77.79	77.98	77.27	82.45
FR	book	82.47	81.01	80.27	80.76	78.49	-	78.25	84.25
	DVD	81.86	81.68	80.27	78.83	78.80	-	74.83	79.60
	music	81.51	80.03	79.41	75.78	77.92	-	78.71	80.09
JP	book	70.93	71.59	71.23	70.22	73.09	70.75	71.11	71.75
	DVD	74.62	72.82	72.55	71.30	71.07	74.96	73.12	75.40
	music	76.48	76.26	75.38	72.02	75.11	77.06	74.38	75.45

Table 2: Comparison results (accuracy) on DE (German), FR (French) and JP (Japanese).

TL	Domain	BM	PDBM	UMM	MT-LR	MT-SVM	Bi-PV	BSWE
CN	book	79.7	77.8	78.4	76.5	77.9	78.5	81.1
	DVD	81.7	80.9	79.8	79.6	81.4	82.0	81.6
	music	79.2	77.3	75.8	74.1	70.7	75.3	79.4
EN	book	81.6	80.5	80.2	-	-	-	-
	DVD	81.7	80.8	79.5	-	-	-	-
	music	76.8	78.8	77.9	-	-	-	-

Table 3: Comparison results (accuracy) on CN (Chinese) and EN (English).

and *en-jp*, but relies on pivot-based methods for bridging language gaps. We also find that the different parallel corpora used by the UMM model are of different quality and genres, and if they are used at the same time, they may have some negative influence on each other and thus the learned word embeddings are not always better than the BM and PDBM models using only one or two parallel corpora. What's more, the available parallel data in different language pairs are of various sizes (0.12M $\sim$ 2.0M). Considering all these issues, the results of UMM are promising because the learned single sentiment classifier can work generally well in multiple languages. We believe that if more high-quality and balanced parallel data are used, the performance of the universal sentiment classifier will be improved.

Second, we compare our models with typical CL classification methods. In Table 2, we can see our models can outperform MT-BOW, CL-SCL, and CR-RL in most test cases, and outperform BSE in the German language. Our models can achieve very close results with the other sophisticated CL methods, including Bi-PV. In Table 3, we can see our models can generally outperform MT-LR and MT-SVM, and achieve very competitive results with other strong CL methods, including Bi-PV and BSWE. Most CL classification methods rely on commercial machine translation systems (e.g. Google Translate) for translating the reviews (including the training reviews, the test reviews and additional unlabeled reviews) to get parallel data. Compared with the large amount of parallel data used by commercial machine translation systems, the parallel data used by our models are of a very small size. Though our models are simply based on word embeddings, and the parallel data used by our models are in a small scale, the performance achieved by our models are very competitive.

In Figure 1, we show the visualization of word embeddings learned by the UMM model for some example words. We can see that similar sentiment words in different languages appear nearby with each other. The figure demonstrate that the UMM model are successful in learning sentiment-aware word embeddings in multiple languages.

4 Related Work

The most closely related work is cross-lingual sentiment classification, which aims to leverage the labeled sentiment data from a language with rich sentiment resources (e.g., English) to perform sentiment classification in a target language lacking sentiment resources (e.g., Japanese). Some studies tried to transfer labeled data from the source

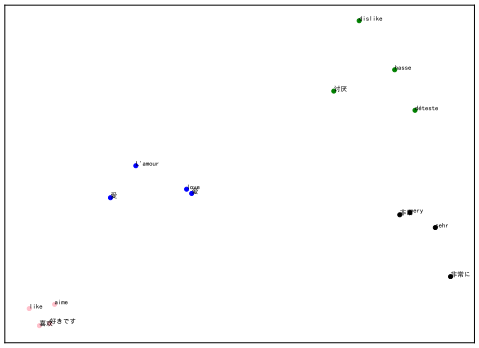

Figure 1: Visualization of word embeddings in UMM (Chinese, Japanese, English, French, German). The similar words are marked in the same color.

language to the target language (Banea et al., 2008; Wan, 2009; Gao et al., 2015; Chen et al., 2015), and some other studies tried to build a unified feature/semantic space in both two languages(Prettenhofer and Stein, 2010; Xiao and Guo, 2013; Zhou et al., 2015, 2016b,a; Li et al., 2017). In the latter case, the sentiment classifier learned in the source language can be used for sentiment classification in both languages. Particularly, Wan (2009) used machine translation to translate the source language to the target language to bridge the gap and applied the co-training approach. Prettenhofer and Stein (2010) provided a CL-SCL model based on structural correspondence learning (SCL) for sentiment classification. Lu et al. (2011) explored to increase the labeled data in both the source and target languages by applying an extra unlabeled parallel data. Xiao and Guo (2013) expected to get cross-lingual discriminative word embeddings to perform the multiple document classification tasks. Their intuitive thought is based on a delicate log-losses function, which aims to increase the probability of the documents with their labels. Like Lu et al. (2011), Meng et al. (2012) also proposed their cross-lingual mixture model to leverage an unlabeled parallel dataset. They intended to learn the previously unseen sentimental words from the big parallel corpus. Some studies have attempted to address multi-lingual sentiment classification (Deriu et al., 2017), but different from our study, they directly leverage training data in multiple languages, by assuming the training data can be obtained directly or in a distant supervision way in each language, and they did not consider the resource or data transfer problem at all.

Word embeddings have shown its great practicable usability in plenty of natural language processing tasks, such as information retrieval (Diaz et al., 2016; Zuccon et al., 2015), machine translation (Shi et al., 2016; Zhang et al., 2014), sentiment analysis (Ren et al., 2016; Xu et al., 2015; Tang et al., 2014) and so on. Bilingual word embeddings have been induced for cross-lingual NLP tasks (Vulić and Moens, 2015; Guo et al., 2014; Zou et al., 2013; Tang et al., 2014; Luong et al., 2015; Zhou et al., 2015). In particular, Luong et al. (2015) proposed the BiSkip model to induce bilingual word embeddings, which is extended from the monolingual skip-gram model in *word2vec* to a bilingual model. They added constraint mutually on both the source language and the target language, while the monolingual model only has constraint on a single language. Zhou et al. (2015) proposed an approach to learning bilingual sentiment word embeddings by using sentiment information of text as supervision, based on labeled corpora and their translations. Ferreira et al. (2016) used a single optimization problem by combining a co-regularizer for the bilingual embeddings with a task-specific loss. However, these methods for inducing bilingual word embeddings usually rely on directly parallel corpus.

5 Conclusion and Future Work

In this paper, we proposed an approach to build a universal sentiment classifier in multiple languages. Particularly we proposed a pivot-based model to transfer the sentiment information from the source language to any resource-poor language via pivot languages. Evaluation results show that the pivot-based model can learn bilingual sentiment-aware word embeddings as well as the bilingual model using direct parallel data. Moreover, the universal sentiment classifier built in the five languages can achieve promising results.

In future work, we will investigate using more advanced document embedding techniques (e.g., CNN, RNN) to directly model document-level sentiment information. We will also extend our model to other languages.

Acknowledgments

This work was supported by NSFC (61331011), 863 Program of China (2015AA015403) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for helpful comments. Xiaojun Wan is the corresponding author.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135. Association for Computational Linguistics.

John Blitzer, Mark Dredze, Fernando Pereira, et al. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, volume 7, pages 440–447.

Danushka Bollegala, Tingting Mu, and J Y Goulermas. 2016. Cross-domain sentiment classification using sentiment sensitive embeddings. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):398–410.

Qiang Chen, Wenjie Li, Yu Lei, Xule Liu, and Yanxiang He. 2015. Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *Association for Computational Linguistics*, pages 419–429.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multilanguage sentiment classification. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1045–1052. International World Wide Web Conferences Steering Committee.

Fernando Diaz, Bhaskar Mitra, and Nick Craswell. 2016. Query expansion with locally-trained word embeddings. *Association for Computational Linguistics*, pages 367–377.

Chris Dyer, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. 2010. cdec: a decoder, alignment, and learning framework for finite-state and context-free translation models. In *ACL 2010, Proceedings of the Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, pages 7–12.

Daniel C. Ferreira, Andr F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Meeting of the Association for Computational Linguistics*, pages 2019–2028.

Dehong Gao, Furu Wei, Wenjie Li, Xiaohua Liu, and Ming Zhou. 2015. Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. *Computational Linguistics*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning sense-specific word embeddings by exploiting bilingual resources. In *International Conference on Computational LinguisticsCOLING*, pages 497–507.

Philipp Koehn. 2004. A parallel corpus for statistical machine translation. *Proceedings of the Third Workshop on Statistical Machine Translation*, (1):3–4.

Nana Li, Shuangfei Zhai, Zhongfei Zhang, and Boying Liu. 2017. Structural correspondence learning for cross-lingual sentiment classification with one-to-many mappings. pages 3490–3496.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Bin Lu, Chenhao Tan, Claire Cardie, and Benjamin K Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 320–330. Association for Computational Linguistics.

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

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Ge Xu, and Houfeng Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Meeting of the Association for Computational Linguistics: Long Papers*, pages 572–581.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Hieu Pham, Thang Luong, and Christopher Manning. 2015. Learning distributed representations for multilingual text sequences. In *The Workshop on Vector Space Modeling for Natural Language Processing*, pages 88–94.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *ACL 2010, Proceedings of the Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 1118–1127.

Yafeng Ren, Ruimin Wang, and Donghong Ji. 2016. A topic-enhanced word embedding for twitter sentiment classification. *Information Sciences*, 369:188–198.

Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang. 2016. Knowledge-based semantic embedding for machine translation. pages 2245–2254.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. pages 1555–1565.

Xuewei Tang and Xiaojun Wan. 2014. Learning bilingual embedding model for cross-language sentiment classification. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02*, pages 134–141. IEEE Computer Society.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 363–372. ACM.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *ACL 2009, Proceedings of the Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing of the Afnlp, 2-7 August 2009, Singapore*, pages 235–243.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *Conference on Empirical Methods in Natural Language Processing*, pages 1465–1475.

Ruifeng Xu, Tao Chen, Yunqing Xia, Qin Lu, Bin Liu, and Xuan Wang. 2015. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, 7(2):226–240.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, Chengqing Zong, et al. 2014. Bilingually-constrained phrase embeddings for machine translation. pages 111–121.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Association for Computational Linguistics*, pages 430–440.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016a. Attention-based lstm network for cross-lingual sentiment classification.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2016b. Cross-lingual sentiment classification with bilingual document representation learning. In *Meeting of the Association for Computational Linguistics*, pages 1403–1412.

Micha Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Lrec*.

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

Guido Zuccon, Bevan Koopman, Peter Bruza, and Leif Azzopardi. 2015. Integrating and evaluating neural word embeddings in information retrieval. page 12.