Bilingual Word Representations with Cross-cultural Socio-linguistic Features

Abstract

Capturing cross-cultural differences is an important challenge in bilingual text understanding and machine translation. This paper presents a novel framework for obtaining bilingual word representations from social media by leveraging socio-linguistic features. Such representations can act as a building block for cross-lingual studies in computational social science. We evaluated our framework on two such tasks: detection of cross-cultural differences in named entities and bilingual lexicon extraction for Internet slangs. Experimental results show that the proposed word representations outperform a number of baseline methods by substantial margins.

1 Introduction

Computing similarity between terms is one of the most fundamental computational task in natural language understanding. Much work has been done in this area, most notably using the distributional properties drawn from large monolingual textual corpora to train vector representations for words or other linguistic units [Mikolov *et al.*, 2013b; Pennington *et al.*, 2014]. Recently there is growing interest in cross-lingual and cross-cultural similarity computation [Luong *et al.*, 2015; Pennebaker *et al.*, 2016]. For example, consider this question: "is there any perceptional difference between Nagoya (a city in Japan) for native English speakers and 名古屋 (Nagoya in Chinese) for Chinese?" Such questions are important not only in cross-cultural social studies but also in machine translation if we want to produce culturally sensitive results. The key challenge is how to map distributional representations in one language to another without losing cultural characteristics. Each of existing models requires a different form of cross-lingual supervision: a parallel corpus with alignments, a bilingual lexicon or comparable documents [Upadhyay *et al.*, 2016]. However, it is costly to build aligned parallel corpora for mining cross-cultural differences and most work has not purposely preserved socio-linguistic features which reflect cultural and social context of terms.

In this paper, we propose an approach to project two monolingual word vector spaces into a common higher-dimensional space, known as social vector space (*SocVec*). Each dimension of this space represents a socio-linguistic feature derived from semantic similarities within the two original vector spaces respectively. As a result, term W in language L_1 and term U in language L_2 can each be projected into a vector in *SocVec*, where cross-lingual similarity between W and U can be computed directly. To reflect cultural perception of each term in such projection, we make use of a set of opinion-related words. This approach is backed by the following observations: i) the perception toward a concept or entity can be captured by opinion-related context; ii) such contexts can be mined from online social media such as Twitter and Weibo; iii) perception changes over time, so one can only talk about cross-cultural differences with respect to a particular time frame.

We evaluated the proposed framework on two novel and interesting cross-cultural and cross-lingual tasks. The first task is mining cross-cultural differences in the perception of named entities (persons, places and organizations). Perception about named entities can be very different from culture to culture. Back in 2012, in the case of "Nagoya", while most English speakers considered the city to be a nice travel destination with a few live concerts taking place, the Chinese people overwhelmingly greeted the city with anger and condemnation because of the city mayor's comments that denied the truthfulness of Nanjing Massacre in 1937. Enabling machines to understand such cultural differences toward named entities, can be useful in various cross-lingual language processing tasks and human-computer interactions.

The second task is bilingual lexicon extraction for Internet slangs. Online social media has been the rich soil to produce new and emerging slangs in all languages and cultures. For example, "浮云" (literally means "floating clouds", now means "nothingness"). State-of-the-art machine translation systems are inadequate in translating such emerging slangs. We wish to automatically explain such dynamic slangs from one language by another language.

In summary, this paper makes the following contributions:

- We propose a direct and uncomplicated bilingual word representation model as a building block for cross-cultural social studies and socio-linguistic research (Section 2).
- We propose two novel bilingual socio-linguistic tasks and evaluate our model on both of them. Results show

that our model outperforms strong baseline methods by significant margins (Section 3).

- We open-source a prototype tool of building cross-cultural bilingual vector space and release two valuable datasets on the above tasks and a bilingual socio-linguistic lexicon, which will benefit future research in this area.

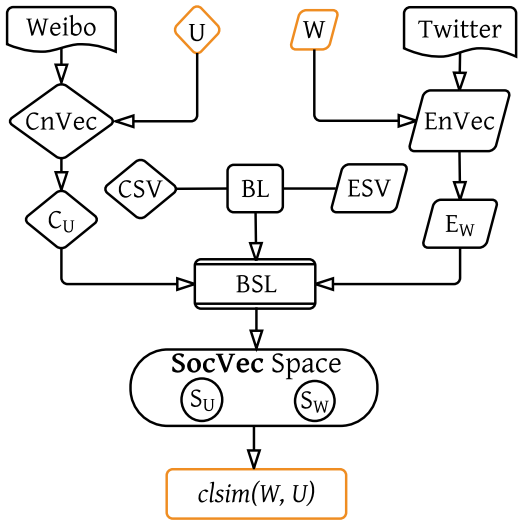

Figure 1: Workflow of SocVec for computing the cross-lingual similarity between an English word *W* and a Chinese word *U*. E_x, C_x and S_x denote the word vectors of word x in *EnVec* space, *CnVec* space and *SocVec* space respectively.

2 Approach

Our problem is that given an English term W and a Chinese term U, compute a cross-lingual similarity score, $clsim(W, U)$, that represents the similarity in perception of W and U in the respective cultures. While this paper mainly considers English and Chinese as our target languages, the techniques developed here are language independent and thus can be used for any two natural languages.

Next sections first discuss the intuition behind our model informally, then give the overall workflow of our approach, and finally present the details of the *SocVec* framework.

2.1 The Intuition Behind SocVec

A very naive solution to the problem is to translate U to its English counterpart U' through a bilingual lexicon and then simply consider the cosine similarity between W and U' by their word embeddings in English. However, this solution is infeasible under two situations: i) if U is an OOV (Out of Vocabulary) term, then no U' exists in the bilingual lexicon; ii) if W and U refer to the same named entity, $U' = W$, $clsim(W, U)$ remains to be 1 and cannot capture any cross-cultural difference.

Our intuition is thus to project English and Chinese embeddings to a common third space. This projection needs to carry *socio-linguistic* context such as opinions, sentiments and cognition associated with the terms in respective languages. This information will be encoded as features in the common space.

2.2 Overall Workflow

The *SocVec* model attacks the problem with the help of four external resources: i) English Twitter corpus; ii) Chinese Sina Weibo corpus; iii) a bilingual lexicon (denoted as *BL*) between English and Chinese common words; and iv) English and Chinese socio-linguistic vocabularies (denoted as *ESV* and *CSV*) consisting of opinion words.

Figure 1 shows our framework. First, we train English and Chinese word embeddings (known as *EnVec* and *CnVec*) from Twitter and Weibo corpus respectively. Then we build a bilingual socio-linguistic lexicon (denoted as *BSL*) from the bilingual lexicon and two monolingual socio-linguistic vocabularies. *BSL* is used to map previously incomparable *EnVec* and *CnVec* into the common higher-dimensional *SocVec* space and thus induce two new vectors S_W and S_U, which are comparable to each other in *SocVec* now.

2.3 SocVec Modeling

In this section, we present the details of building *BSL* and constructing *SocVec* space.

Building Bilingual Socio-linguistic Lexicon

For convenience, we use the term "social word" here to represent the words in a socio-linguistic vocabulary. The process of building *BSL* is illustrated in Figure 2. To build *BSL*, we first use the bilingual lexicon to translate each word in *ESV* into several Chinese words. Then, we filter out Chinese words that are not in *CSV*. After that, we have a group of Chinese social words as the translations of each English social word. The final step of building *BSL* is to construct a "pseudo-word" for each English social word by a pseudo-word generator.

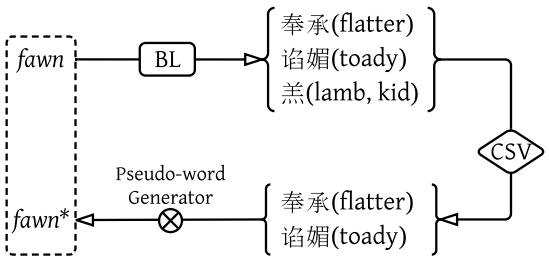

Figure 2: Generating an entry in BSL for "*fawn*" and its pseudo-word "*fawn**".

For example, in Figure 2, English social word "*fawn*" has three Chinese translations in the bilingual lexicon, but only two of them are Chinese social words. The pseudo-word generator takes the *CnVec* of the two words as input and generates the pseudo-word vector of "*fawn*", denoted as "*fawn**".

Denote $\mathbf{C_i}$ as the *CnVec* of the i^{th} out of N Chinese translation words of a given English social word. Four common pseudo-word generator functions are as follows:

Max.: Maximum of the values in each dimension.

$$Pseudo(\mathbf{C_1}, ..., \mathbf{C_N}) = \begin{bmatrix} max(C_1^{(1)}, ..., C_N^{(1)}) \\ \vdots \\ max(C_1^{(N)}, ..., C_N^{(N)}) \end{bmatrix}^{\mathrm{T}}$$

Avg.: Average of the values in each dimension.

$$Pseudo(\mathbf{C_1}, ..., \mathbf{C_N}) = \frac{1}{N} \sum_{i}^{N} \mathbf{C_i}$$

W.Avg.: Weighted average value of each dimension with respect to translation confidence (see Section 3.1).

$$Pseudo(\mathbf{C_1}, ..., \mathbf{C_N}) = \frac{1}{N} \sum_{i}^{N} w_i \mathbf{C_i}$$

Top: Choose the most confident translation, $\mathbf{C_{top}}$.

$$Pseudo(\mathbf{C_1}, ..., \mathbf{C_N}) = \mathbf{C_{top}}$$

Projecting EnVec and CnVec into SocVec Space

Let B_i be i^{th} English word in *BSL*, B_i^* be its corresponding Chinese pseudo-word, $\mathbf{E_x}$, $\mathbf{C_x}$ and $\mathbf{S_x}$ be the word vectors of word x in *EnVec*, *CnVec* and *SocVec* respectively and L be the number of entries in the *BSL*. The last step of computing $clsim(W, U)$ is to project $\mathbf{E_W}$ and $\mathbf{C_U}$ to $\mathbf{S_W}$ and $\mathbf{S_U}$ so that they are comparable to each other. We define the function f to compute cross-lingual similarity between W and U as follows:

$$\begin{aligned} clsim(W, U) &:= f(\mathbf{E_W}, \mathbf{C_U}) \\ &= sim\left(\begin{bmatrix} cos(\mathbf{E_W}, \mathbf{E_{B_1}}) \\ \vdots \\ cos(\mathbf{E_W}, \mathbf{E_{B_L}}) \end{bmatrix}^{\mathrm{T}}, \begin{bmatrix} cos(\mathbf{C_U}, \mathbf{C_{B_1^*}}) \\ \vdots \\ cos(\mathbf{C_U}, \mathbf{C_{B_L^*}}) \end{bmatrix}^{\mathrm{T}} \right) \\ &= sim(\mathbf{S_W}, \mathbf{S_U}) \end{aligned}$$

This calculation process is shown in Algorithm 1. Here cos stands for cosine similarity, while sim function is a generic similarity function, for which a number of metrics can be considered, as we discuss later in Section 3.

Algorithm 1: Compute cross-lingual similarity between an English word *W* and a Chinese word *U*

Input: *EnVec*, *CnVec*, BSL with L word pairs
Output: the cross-lingual similarity $clsim(W, U)$

E_W = word vector of W in *EnVec*
C_U = word vector of U in *CnVec*
S_W = zero vector with L dimension
S_U = zero vector with L dimension
for $1 \leq i \leq L$ **do**
    B_i = i^{th} English word in *BSL*
    B_i^* = Chinese pseudo-word of B_i
    $S_W[i]$ = $cos(E_W, E_{B_i})$
    $S_U[i]$ = $cos(C_U, C_{B_i^*})$
return $clsim(W, U)$ = $sim(S_W, S_U)$

Parameters Description

Here, we briefly summarize the main parameters of *SocVec* model:

- Two trained monolingual word embeddings, *EnVec* and *CnVec*;
- A bilingual lexicon;
- Chinese and English socio-linguistic vocabularies;
- Pseudo-word generator function;
- Option of similarity function sim;

We conduct several experiments for testing the these parameters in Section 3.2.

3 Evaluation

We evaluate our framework on two tasks: mining cross-cultural differences in named entities, bilingual lexicon extracting for Internet slangs. Next sections first discuss some preliminary setup then present our experiments for the two tasks.

3.1 Experiment Setup

Prior to evaluating the two tasks, we need to first preprocess the Twitter and Weibo corpus, perform entity linking, then train word embedding for all terms including entity names, and finally construct the bilingual lexicon for common words and socio-linguistic vocabularies, which contain opinion and sentiment related words.

Microblog Corpora

The English Twitter corpus is from Archive Team's Twitter stream grab[1]. We select only English tweets and leave out all the metadata. Our Sina Weibo corpus comes from Open Weiboscope Data Access[2] [Fu *et al.*, 2013]. Both corpora cover the whole year of 2012. We then downsample each corpus to 100 million messages, where each message must contain at least 10 characters. We also normalize the tweets [Han *et al.*, 2012] to reduce the noise, lemmatize the text [Manning *et al.*, 2014] and filter out stop words. For Weibo corpus, we used LTP [Che *et al.*, 2010] to do word segmentation.

Entity Linking and Monolingual Word Vectors

After preprocessing the corpora, we first do entity linking. For Twitter corpus, we used Wikifier [Cheng and Roth, 2013; Ratinov *et al.*, 2011], a widely used entity linker for Wikipedia. Because no suitable Chinese entity linking tool is available, we implement our own tool aiming for high precision. We utilize context information for selecting entity candidates. That is, only when a surface form is an exact match, or the candidate has been linked in the surrounding text before, or satisfies the occurrence frequency criterion, will it be linked by our tool. Also, we only focus on entities that have both English and Chinese Wikipedia pages. We argue that such precision oriented approach is sufficiently good for our tasks, because even if an entity is not recognized, it can still be captured as a normal word and contribute to the semantics of other terms. Next we use Word2Vec [Mikolov *et al.*, 2013b] to train English and Chinese word embedding on Twitter and Weibo corpus respectively.

Bilingual Lexicon

We collect a bilingual lexicon using Bing Translate API[3], which contains 22,082 English and 37,645 Chinese common words, translation pairs in which are many-to-many. In addition, each translation pair is associated with a confidence score ranging from 0 to 1.

[1] `https://archive.org/details/twitterstream`
[2] `http://weiboscope.jmsc.hku.hk/datazip/`
[3] `http://www.bing.com/translator`

Socio-linguistic Vocabulary

Our socio-linguistic vocabularies come from Empath [Fast *et al.*, 2016] and OpinionFinder [Choi *et al.*, 2005] for English, and TextMind [Gao *et al.*, 2013] for Chinese. Empath is similar to LIWC [Pennebaker *et al.*, 2001], but with more words and more categories. We manually select 91 categories of words that are more relevant to human perception and psychological processes. In summary, our selected Empath vocabulary contains 3343 English words while the OpinionFinder vocabulary contains 3861 words, and the union of two vocabularies consists of 5574 unique words. Some example words from our vocabularies are *fawn, inept, tremendous, gratitude, terror, terrific, kiss, loving, traumatic*, etc.

3.2 Mining Cross-cultural Differences of Named Entity

This task is to discover and quantify cross-cultural differences of concerns and topics towards name entities. We first explain how we obtain the ground truth, then present several baseline methods to this problem and finally show our experiment results in detail.

Ground Truth

Lacking settled measures of cultural differences in named entities, we propose to apply the distributional hypothesis of Harris [1954], which states that the meaning of words is evidenced by the contexts they occur with. Similarly, we assume that the cultural properties of entity can be evidenced by the terms they co-occur with. Thus, for each named entity, we present four human annotators with two list of top 20 words ranking by their occurrence with the named entity, from Twitter and Weibo respectively. We select 700 named entities for annotators to label. These entities are the most frequently mentioned both in Twitter and Weibo. Annotators rate relatedness between the two word lists from 1 to 5, where 1 indicates the lists are very different and 5 means they are the most similar. The ground truth similarity score for each entity is averaged over all annotators' scores. For classification problem, an entity is considered culturally similar if the score is larger than 3.0, and culturally different otherwise. The inter-annotator agreement is 0.531 by Cohen's kappa coefficient, suggesting moderate correlation.

Baseline and Our Methods

We propose five baselines that can be categorized into *distributional* and *transform*-based work. First category comprises of three baselines, comparing the lists of surrounding English and Chinese terms, denoted as L_E and L_C, by computing the cross-lingual relatedness between the two lists, though different baselines differ in terms of word selection and similarity computation. Second category computes the vector representation in English and Chinese corpus respectively then trains a transformation.

The first three baselines fall into the distributional category:

- Bilingual Lexicon Jaccard Similarity (BL-JS) BL-JS uses the bilingual lexicon to translate L_E to a Chinese word list L_E^* as a medium and then calculates the Jaccard Similarity between L_E^* and L_C as J_{EC}. Similarly, we can compute J_{CE}. Finally, use $\frac{J_{EC}+J_{CE}}{2}$ as the cross-cultural similarity of this given name entity.
- WordNet Wu-Palmer Similarity (WN-WUP) Instead of using the bilingual lexicon and Jaccard Similarity, WN-WUP uses Open Multilingual Wordnet [Wang and Bond, 2013; Bond and Foster, 2013] to calculate the average similarity of two lists of words from different languages.
- Word Embedding based Jaccard Similarity (EM-JS) EM-JS is very similar to BL-JS, except for that its L_E and L_C are generated by ranking the similarities between the name of entities and all English words and Chinese words respectively.

The second type baseline methods are as follows:

- Linear Transformation (LTrans) We follow the steps in Mikolov et al. [2013a] to train a transformation matrix between *EnVec* and *CnVec*, using 3000 translation pairs with confidence of 1 in the bilingual lexicon. Given a named entity, this solution would simply calculate cosine similarity between the *EnVec* of its English name and the *transformed CnVec* of its Chinese name.
- Bilingual Lexicon Space (BLex) This baseline is similar to *SocVec* but it does not utilize socio-linguistic vocabulary and simply projecting *EnVec* and *CnVec* by the similarities to all the word pairs in lexicon.

Given a named entity with its English name and Chinese name, our method is to simply regard the similarity between their *SocVec*s as its cross-cultural difference.

Experimental Results

Table 1 shows some of the most culturally different named entities obtained from our method. The listed hot and trending topics on Twitter and Weibo are manually summarized from our ground truth words for presentation purpose. It is obvious that the listed entities all have large divergence on topics over Twitter and Weibo messages, thus reflecting the mined cross-cultural differences.

Table 1: Selected culturally different named entities, with Twitter and Weibo's trending topics manually summarized

Entity	Twitter topics	Weibo topics
Maldives	coup, president Nasheed quit, political crisis	holiday, travel, honeymoon, paradise, beach
Nagoya	tour, concert, travel, attractive, Osaka	Mayor Takashi Kawamura, Nanjing Massacre, denial of history
Quebec	Conservative Party, Liberal Party, politicians, prime minister, power failure	travel, autumn, maples, study abroad, immigration, independence
Philippines	gunman attack, police, quake, tsunami	South China Sea, sovereignty dispute, confrontation, protest
Yao Ming	NBA, Chinese, good player, Asian	patriotism, collective values, Jeremy Lin, Liu Xiang, Chinese Law maker, gold medal superstar
University of Southern California	college football, baseball, Stanford, Alabama, win, lose	top study abroad destination, Chinese student murdered, scholars, economics, Sino American politics

In Table 2, we evaluate the baseline methods and our approach with three metrics: Spearman correlation and Pearson correlation on the ranking problem, and Mean Average Precision (MAP) on the classification problem. Monolingual word vectors are trained with 5-word context window and 150 dimensions. We choose cosine similarity as the *sim* function to compute the similarity within the SocVec space. The *BSL* of *SocVec:opn* uses only OpinionFinder as English socio-linguistic vocabulary, while *SocVec:all* also uses Emapth lexicon. All models uses "*Top*" pseudo-word generator (see Section 2.3). Results show that our *SocVec* models perform best and the more lexicon are devised the correlation is better.

Following the above setups, we also evaluate the effect of four different similarity options in *SocVec*, namely, Pearson Correlation Coefficient (*PCorr.*), L1-normalized Manhattan distance (*L1+M*), Cosine Similarity (*Cos*) and L2-normalized Euclidean distance (*L2+E*). We show the results in Table 3. We conclude that among these 4 options, *Cos* and *L2+E* perform the best. Table 4 shows effect of using four different pseudo-word generator functions (see Section 2.3). We can conclude that "*Top*" generator function performs best for it reduces the noise brought by ambiguity.

Table 2: Comparison of Different Methods

Method	Spearman	Pearson	MAP
BL-JS	0.276	0.265	0.644
WN-WUP	0.335	0.349	0.677
EM-JS	0.221	0.210	0.571
LTrans	0.366	0.385	0.644
BLex	0.596	0.595	0.765
SocVec:opn	0.668	0.662	**0.834**
SocVec:all	**0.676**	**0.671**	**0.834**

Table 3: Evaluation of Different Similarity Functions

Similarity	Spearman	Pearson	MAP
PCorr.	0.631	0.625	0.806
L1 + M	0.666	0.656	0.824
Cos	**0.676**	0.669	**0.834**
L2 + E	**0.676**	**0.671**	**0.834**

3.3 Bilingual Lexicon Extraction for Internet Slangs

In this section, we evaluate our model on the second task. This task aims to translate, define or understand the meaning of emerging Internet slangs from Chinese to English and from English to Chinese. We first construct the ground truth and then implement several baseline methods for comparison. Finally, we analyze the experimental results quantitatively and qualitatively.

Data Preparation

We use an online Chinese Internet slang glossary[4] consisting of 200 popular Chinese Internet slangs with English explanation. For English slangs, we resort to another slang dictionary and crawl their word list[5] as well as explanations. We randomly downsampled the list to 200 English slangs.

Ground Truth

To evaluate the performance of our model on automatically translating and explaining slangs in another language, we propose to build the ground truth based on above-mentioned explanations of slangs. Since the subtle and latent semantics of Internet slangs are too difficult to translate without losing any information, exact translation of them are always missing from all the dictionary. Thus, we argue that using the relevant terms in the explanation as the target word list and then computing the similarity between the results and the list is a better approach to evaluate bilingual slang lexicon induction system. In the following example, we manually selected and annotated words from original glossary that are related to the meaning of the slang, constructing the ground truth:

二百五 A ***foolish*** person who is lacking in sense but still ***stubborn***, ***rude***, and ***impetuous***.

Baseline and Our Methods

We proposed several baselines regarding Internet slang translation. One type of baseline for translation comparison are from Internet translators. Google, Bing and Baidu are all well-known online translators, thus with our test set's slang as input, we retrieve the output of translation. An additional baseline specific to Chinese slangs is from CC-CEDICT[6] (CC), an online public-domain Chinese-English dictionary, which is well updated with some popular slang inside.

Except for simply using Linear Transform (LTrans) to find the most cross-lingually similar words of the given slang, we propose a strong baseline leveraging our *bilingual lexicon* (BLex). Given an Internet slang of one language, for each common word of target language in our bilingual lexicon, we obtain its translations back into the the source language and then calculate the word similarities between the input slang and aforementioned translation word(s) within monolingual word vector. A word may have multiple possible translation words in the other language. In this case, we choose to take average over all of them in terms of similarity score. We then rank the words by their scores and take top 5 words to form a word set, while other online translation baselines directly produce a word set for later comparison with the ground truth word set.

Our method simply uses top 5 most similar words (in target language) with the given slang in SocVec space.

Table 4: Evaluation of Different Pseudo-word Generators

Generator	Spearman	Pearson	MAP
Max.	0.413	0.401	0.726
Avg.	0.667	0.625	0.831
W.Avg.	0.671	0.660	0.832
Top	**0.676**	**0.671**	**0.834**

Experimental Results

In Table 5, we present several examples of translation results for Chinese and English slangs with their explanations from the glossaries. Our results are highly correlated with these explanations and capture their core semantics, whereas most online translators are inadequate for extracting subtle meanings of such slangs. They often give just literal meanings as translation or nothing.

Additionally, we take a step forward to directly translate between English slangs and Chinese slangs by simply filtering out common words in the original result. Examples are shown in Table 6.

To quantitatively compare our methods with the baselines, we need to measure the similarity between the translation word set and the ground truth word set. Jaccard similarity coefficient is too strict to capture valuable relatedness between two word sets, since it takes only exact matches into account. We argue that average cosine similarity (ACS) between two sets of word vectors is a better metrics to evaluate similarity between two word sets. The following equation illustrates such computation, where A and B are the two word sets, $\mathbf{A_i}$ and $\mathbf{B_j}$ denotes the word vector of the i^{th} word in A and j^{th} word in B respectively.

$$ACS(A,B) = \frac{1}{|A||B|}\sum_{i=1}^{|A|}\sum_{j=1}^{|B|}\frac{\mathbf{A_i}\cdot\mathbf{B_j}}{\|\mathbf{A_i}\|\|\mathbf{B_j}\|}$$

Experiment results of Chinese and English online slangs translation are in Table 7a and Table 7b. Typically, the performance of online translators in translating slangs depends on the a number of human-set rules and supervised learning on well-annotated parallel corpora. However, such parallel corpora are rare and costly, especially for social media where Internet slangs emerge the most. It could be a possible reason why they do not perform well. Linear transformation model is trained on translation pairs with high confidence in the bilingual lexicon, which contains no Internet slangs and almost no opinion-related words. Thus, the matrix transforms slangs badly. *BL* is competitive for its similarity is based on monolingual word similarity, while it is limited by the bilingual lexicon and consequently

[4] https://www.chinasmack.com/glossary

[5] http://onlineslangdictionary.com/word-list/

[6] https://cc-cedict.org/wiki/

Table 5: Slang Translation Examples

Slang	Explanation	Google	Bing	Baidu	Ours
浮云	something as ephemeral and unimportant as "passing clouds"	clouds	nothing	floating clouds	nothingness, illusion
水军	"water army", referring to people paid to post comments on the Internet to help shape public opinion by slandering competitors and promoting themselves	Water army	Navy	Navy	propaganda, complicit, fraudulent
城管	"City administrators", who enforce city regulations, with poor reputation as being corrupt and violent, best known for physically bullying illegal street peddlers	urban management	urban management	urban management	terrorist, rioting, threaten
floozy	a woman with a reputation for promiscuity	floozy	劣根性 (depravity)	荡妇(slut)	骚货(slut),妖精(promiscuous)
fruitcake	a crazy person, someone who is completely insane	水果蛋糕 (fruit cake)	水果蛋糕 (fruit cake)	水果蛋糕 (fruit cake)	怪诞(bizarre),令人厌烦(annoying)
nonce	A person convicted (or simply guilty) of sexual crimes, especially pedophilia. Or a common British insult regardless of the tendencies of the person	随机数 (random numbers)	杜撰 (fabricate)	杜撰 (fabricate)	伤风败俗(immoral),十恶不赦(extremely evil),畜类(beast),令人发指(heinous)

loses the information from the related words which are not in the lexicon.

Our *SocVec* utilizes comparable English and Chinese social media corpora and encodes the context and usage of a given slang by computing its similarities with opinion and sentiment words in the socio-linguistic vocabulary of the source language. Therefore, our model keeps the cross-cultural socio-linguistic features, which is a most important reason why we outperform baselines.

Table 6: Slang-to-Slang Translation Examples

Chinese Slang	English Slang	Explanation
萌	adorbz, adorb, adorbs, tweeny, attractiveee	cute, adorable
二百五	shithead, stupidit, douchbag	A foolish and senseless person
鸭梨	antsy, stressy, fidgety, grouchy, badmood	stress, pressure, burden

Table 7: ACS Result of Slang Translation

Google	Bing	Baidu	CC	LTrans	BLex	SocVec
18.24	16.38	17.11	17.38	9.14	20.92	**23.01**

(a) Chinese Slang Translation

Google	Bing	Baidu	LTrans	BLex	SocVec
6.40	15.96	15.44	7.32	11.43	**17.31**

(b) English Slang Translation

4 Related Work

Most existing approaches for learning cross-lingual word representations rely on expensive parallel corpora with word or sentence alignments [Klementiev *et al.*, 2012; Kočiskỳ *et al.*, 2014] or a supervised model to learn a transformation matrix between two monolingual vector spaces [Mikolov *et al.*, 2013a]. These work aims for improving monolingual tasks and cross-lingual document classification, which does not require cross-cultural signals. However, they fails to capture socio-linguistic information. We propose an uncomplicated framework to quantify the cross-cultural differences by leveraging publicly accessible resources such as Bing Translator and OpinionFinder lexicon.

Cross-cultural studies have been conducted in sociology, anthropology and psychology for many years. Recently, some researchers propose studying cross-cultural analysis through text mining and natural language processing. Nakasaki et al. [2009] and Elahi et al. [2012] show that User Generated Content (UGC) like microblogs, is a valuable resources to cross-cultural analysis. The most relevant work to our first task is Pennebaker et al. [2016], which studies the cross-cultural differences in word usage between Australian and American English through socio-linguistic features. Nevertheless, their supervised model are dependent on large volume of training data and limited to identifying differences of monolingual word usage. To the best of our knowledge, we are among the first to focus on cross-cultural differences in named entities and to propose an effective unsupervised approach.

Previous work about Internet slangs mainly focuses on automatic discovering of slangs [Elsahar and Elbeltagy, 2014] and normalization of noisy texts [Han *et al.*, 2012]. However, research on automatic translation and explanation for slangs in another language is missing from literature. Our work on the second task fills the void by directly computing cross-lingual similarities to find the most related words in another language.

5 Conclusion

In this paper, we conclude that the cultural properties and usages of a term (including named entities and slangs) can be effectively represented by its similarities to socio-linguistic words. Bilingual socio-linguistic lexicon enables two incomparable monolingual semantic spaces to be comparable with each other. Our proposed framework can assist cross-cultural social studies and cross-lingual linguistic research, such as detection of cross-cultural differences in named entities and extraction of bilingual lexicon for Internet slangs.

References

[Bond and Foster, 2013] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362, 2013.

[Che *et al.*, 2010] Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics, 2010.

[Cheng and Roth, 2013] X. Cheng and D. Roth. Relational inference for wikification. In *EMNLP*, 2013.

[Choi *et al.*, 2005] Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2005.

[Elahi and Monachesi, 2012] Mohammad Fazleh Elahi and Paola Monachesi. An examination of cross-cultural similarities and differences from social media data with respect to language use. In *LREC*, pages 4080–4086, 2012.

[Elsahar and Elbeltagy, 2014] Hady Elsahar and Samhaa R Elbeltagy. A fully automated approach for arabic slang lexicon extraction from microblogs. pages 79–91, 2014.

[Fast *et al.*, 2016] Ethan Fast, Binbin Chen, and Michael S Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016.

[Fu *et al.*, 2013] King-wa Fu, Chung-hong Chan, and Michael Chau. Assessing censorship on microblogs in china: Discriminatory keyword analysis and the real-name registration policy. *IEEE Internet Computing*, 17(3):42–50, 2013.

[Gao *et al.*, 2013] Rui Gao, Bibo Hao, He Li, Yusong Gao, and Tingshao Zhu. Developing simplified chinese psychological linguistic analysis dictionary for microblog. In *International Conference on Brain and Health Informatics*, pages 359–368. Springer, 2013.

[Han *et al.*, 2012] Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 421–432. Association for Computational Linguistics, 2012.

[Harris, 1954] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[Klementiev *et al.*, 2012] A Klementiev, I Titov, and B Bhattarai. Inducing crosslingual distributed representations of words. 2012.

[Kočiskỳ *et al.*, 2014] Tomáš Kočiskỳ, Karl Moritz Hermann, and Phil Blunsom. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*, 2014.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D Manning. 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, 2015.

[Manning *et al.*, 2014] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv.org*, September 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Nakasaki *et al.*, 2009] Hiroyuki Nakasaki, Mariko Kawaba, Sayuri Yamazaki, Takehito Utsuro, and Tomohiro Fukuhara. Visualizing cross-lingual/cross-cultural differences in concerns in multilingual blogs. In *ICWSM*, 2009.

[Pennebaker *et al.*, 2001] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[Pennebaker *et al.*, 2016] James Pennebaker, Aparna Garimella, and Rada Mihalcea. Identifying cross-cultural differences in word usage. 2016.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[Ratinov *et al.*, 2011] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1375–1384. Association for Computational Linguistics, 2011.

[Upadhyay *et al.*, 2016] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425*, 2016.

[Wang and Bond, 2013] Shan Wang and Francis Bond. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP*, pages 10–18, 2013.