Mining Cross-Cultural Differences and Similarities in Social Media

Bill Yuchen Lin[1]* **Frank F. Xu**[1]* **Kenny Q. Zhu**[1] **Seung-won Hwang**[2]
[1]Shanghai Jiao Tong University, Shanghai, China
{yuchenlin, frankxu}@sjtu.edu.cn, kzhu@cs.sjtu.edu.cn
[2]Yonsei University, Seoul, Republic of Korea
seungwonh@yonsei.ac.kr

Abstract

Cross-cultural differences and similarities are common in cross-lingual natural language understanding, especially for research in social media. For instance, people of distinct cultures often hold different opinions on a single named entity. Also, understanding slang terms across languages requires knowledge of cross-cultural similarities. In this paper, we study the problem of computing such cross-cultural differences and similarities. We present a lightweight yet effective approach, and evaluate it on two novel tasks: 1) mining cross-cultural differences of named entities and 2) finding similar terms for slang across languages. Experimental results show that our framework substantially outperforms a number of baseline methods on both tasks. The framework could be useful for machine translation applications and research in computational social science.

1 Introduction

Computing similarities between terms is one of the most fundamental computational tasks 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 of words or other linguistic units (Pennington et al., 2014; Le and Mikolov, 2014). However, computing cross-cultural similarities of terms between different cultures is still an open research question, which is important in cross-lingual natural language understanding. In this paper, we address cross-cultural research questions such as these:

Figure 1: Two social media messages about Nagoya from different cultures in 2012

1. *Were there any cross-cultural differences between Nagoya (a city in Japan) for native English speakers and 名古屋 (Nagoya in Chinese) for Chinese people in 2012?*
2. *What English terms can be used to explain "浮云" (a Chinese slang term)?*

These kinds of questions about cross-cultural differences and similarities are important in cross-cultural social studies, multi-lingual sentiment analysis, culturally sensitive machine translation, and many other NLP tasks, especially in social media. We propose two novel tasks in mining them from social media.

The first task (Section 4) is to mine cross-cultural differences in the perception of named entities (e.g., persons, places and organizations). Back in 2012, in the case of "Nagoya", many native English speakers posted their pleasant travel experiences in Nagoya on Twitter. However, Chinese people overwhelmingly greeted the city with anger and condemnation on *Weibo* (a Chinese version of Twitter), because the city mayor denied the truthfulness of the Nanjing Massacre. Figure 1 illustrates two example microblog messages about Nagoya in Twitter and Weibo respectively.

The second task (Section 5) is to find similar terms for slang across cultures and languages. Social media is always a rich soil where slang terms emerge in many cultures. For example,

*Both authors contributed equally.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 709–719
Melbourne, Australia, July 15 - 20, 2018.

"浮云" literally means "floating clouds", but now almost equals to "nothingness" on the Chinese web. Our experiments show that well-known online machine translators such as *Google Translate* are only able to translate such slang terms to their literal meanings, even under clear contexts where slang meanings are much more appropriate.

Enabling intelligent agents to understand such cross-cultural knowledge can benefit their performances in various cross-lingual language processing tasks. Both tasks share the same core problem, which is **how to compute cross-cultural differences (or similarities) between two terms from different cultures.** A term here can be either an ordinary word, an entity name, or a slang term. We focus on names and slang in this paper for they convey more social and cultural connotations.

There are many works on cross-lingual word representation (Ruder et al., 2017) to compute general cross-lingual similarities (Camacho-Collados et al., 2017). Most existing models require bilingual supervision such as aligned parallel corpora, bilingual lexicons, or comparable documents (Sarath et al., 2014; Kočiský et al., 2014; Upadhyay et al., 2016). However, they do not purposely preserve social or cultural characteristics of named entities or slang terms, and the required parallel corpora are rare and expensive.

In this paper, we propose a lightweight yet effective approach to project two incompatible monolingual word vector spaces into a single bilingual word vector space, known as social vector space (*SocVec*). A key element of SocVec is the idea of "bilingual social lexicon", which contains bilingual mappings of selected words reflecting psychological processes, which we believe are central to capturing the socio-linguistic characteristics. Our contribution in this paper is two-fold:

(a) We present an effective approach (SocVec) to mine cross-cultural similarities and differences of terms, which could benefit research in machine translation, cross-cultural social media analysis, and other cross-lingual research in natural language processing and computational social science.

(b) We propose two novel and important tasks in cross-cultural social studies and social media analysis. Experimental results on our annotated datasets show that the proposed method outperforms many strong baseline methods.

2 The *SocVec* Framework

In this section, we first discuss the intuition behind our model, the concept of "social words" and our notations. Then, we present the overall workflow of our approach. We finally describe the *SocVec* framework in detail.

2.1 Problem Statement

We choose (English, Chinese) to be the target language pair throughout this paper for the salient cross-cultural differences between the east and the west[1]. Given an English term W and a Chinese term U, the core research question is how to compute a similarity score, $ccsim(W, U)$, to represent the *cross-cultural similarities* between them.

We cannot directly calculate the similarity between the monolingual word vectors of W and U, because they are trained separately and the semantics of dimension are not aligned. Thus, the challenge is to devise a way to compute similarities across two different vector spaces while retaining their respective cultural characteristics.

A very intuitive solution is to firstly translate the Chinese term U to its English counterpart U' through a Chinese-English bilingual lexicon, and then regard $ccsim(W, U)$ as the (cosine) similarity between W and U' with their monolingual word embeddings. However, this solution is not promising in some common cases for three reasons:

(a) if U is an OOV (Out of Vocabulary) term, e.g., a novel slang term, then there is probably no translation U' in bilingual lexicons.

(b) if W and U are names referring to the same named entity, then we have $U' = W$. Therefore, $ccsim(W, U)$ is just the similarity between W and itself, and we cannot capture any cross-cultural differences with this method.

(c) this approach does not explicitly preserve the cultural and social contexts of the terms.

To overcome the above problems, our intuition is to project both English and Chinese word vectors into a single third space, known as *SocVec*, and the projection is supposed to purposely carry cultural features of terms.

2.2 Social Words and Our Notations

Some research in psychology and sociology (Kitayama et al., 2000; Gareis and Wilkins, 2011)

[1]Nevertheless, the techniques are language independent and thus can be utilized for any language pairs so long as the necessary resources outlined in Section 2.3 are available.

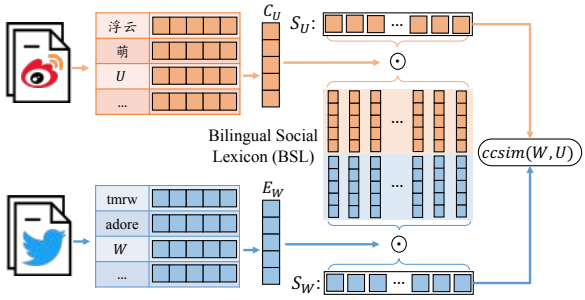

Figure 2: Workflow for computing the cross-cultural similarity between an English word *W* and a Chinese word *U*, denoted by $ccsim(W, U)$

show that culture can be highly related to emotions and opinions people express in their discussions. As suggested by Tausczik and Pennebaker (2009), we thus define the concept of "**social word**" as the words directly reflecting opinion, sentiment, cognition and other human psychological processes[2], which are important to capturing cultural and social characteristics. Both Elahi and Monachesi (2012) and Garimella et al. (2016a) find such *social words* are most effective culture/socio-linguistic features in identifying cross-cultural differences.

We use these notations throughout the paper: *CnVec* and *EnVec* denote the Chinese and English word vector space, respectively; *CSV* and *ESV* denote the Chinese and English social word vocab; *BL* means Bilingual Lexicon, and *BSL* is short for Bilingual Social Lexicon; finally, we use $\mathbf{E_x}$, $\mathbf{C_x}$ and $\mathbf{S_x}$ to denote the word vectors of the word x in *EnVec*, *CnVec* and *SocVec* spaces respectively.

2.3 Overall Workflow

Figure 2 shows the workflow of our framework to construct the *SocVec* and compute $ccsim(W, U)$. Our proposed *SocVec* model attacks the problem with the help of three low-cost external resources: (i) an English corpus and a Chinese corpus from social media; (ii) an English-to-Chinese bilingual lexicon (*BL*); (iii) an English social word vocabulary (*ESV*) and a Chinese one (*CSV*).

We train English and Chinese word embeddings (*EnVec* and *CnVec*) on the English and Chinese social media corpus respectively. Then, we build a *BSL* from the *CSV*, *ESV* and *BL* (see Section 2.4). The *BSL* further maps the previously incompatible *EnVec* and *CnVec* into a single common vector space *SocVec*, where two new vectors, S_W for W and S_U for U, are finally comparable.

2.4 Building the BSL

The process of building the *BSL* is illustrated in Figure 3. We first extract our bilingual lexicon (*BL*), where confidence score w_i represents the probability distribution on the multiple translations for each word. Afterwards, we use BL to translate each social word in the *ESV* to a set of Chinese words and then filter out all the words that are not in the *CSV*. Now, we have a set of Chinese social words for each English social word, which is denoted by a "translation set". The final step is to generate a Chinese "pseudo-word" for each English social word using their corresponding translation sets. A "pseudo-word" can be either a real word that is the most representative word in the translation set, or an imaginary word whose vector is a certain combination of the vectors of the words in the translation set.

For example, in Figure 3, the English social word "*fawn*" has three Chinese translations in the bilingual lexicon, but only two of them (underlined) are in the CSV. Thus, we only keep these two in the translation set in the filtered bilingual lexicon. The pseudo-word generator takes the word vectors of the two words (in the black box), namely 奉承 (flatter) and 谄媚 (toady), as input, and generates the pseudo-word vector denoted by "*fawn**". Note that the direction of building *BSL* can also be from Chinese to English, in the same manner. However, we find that the current direction gives better results due to the better translation quality of our *BL* in this direction.

Given an English social word, we denote $\mathbf{t_i}$ as the i^{th} Chinese word of its translation set consisting of N social words. We design four intuitive types of pseudo-word generator as follows, which are tested in the experiments:

(1) Max. Maximum of the values in each dimension, assuming dimensionality is K:

$$\text{Pseudo}(\mathbf{C_{t_1}}, ..., \mathbf{C_{t_N}}) = \begin{bmatrix} max(C^{(1)}_{t_1}, ..., C^{(1)}_{t_N}) \\ \vdots \\ max(C^{(K)}_{t_1}, ..., C^{(K)}_{t_N}) \end{bmatrix}^{\mathrm{T}}$$

(2) Avg. Average of the values in every dimension:

$$\text{Pseudo}(\mathbf{C_{t_1}}, ..., \mathbf{C_{t_N}}) = \frac{1}{N} \sum_{i}^{N} \mathbf{C_{t_i}}$$

[2]Example social words in English include *fawn, inept, tremendous, gratitude, terror, terrific, loving, traumatic*, etc. We discuss the sources of such social words in Section 3.

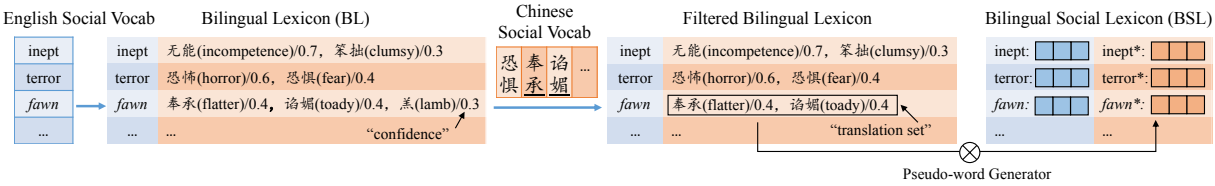

Figure 3: Generating an entry in the BSL for "*fawn*" and its pseudo-word "*fawn**"

(3) WAvg. Weighted average value of every dimension with respect to the translation confidence:

$$\text{Pseudo}(\mathbf{C_{t_1}}, ..., \mathbf{C_{t_N}}) = \frac{1}{N}\sum_{i}^{N} w_i \mathbf{C_{t_i}}$$

(4) Top. The most confident translation:

$$\text{Pseudo}(\mathbf{C_{t_1}}, ..., \mathbf{C_{t_N}}) = \mathbf{C_{t_k}}, k = \underset{i}{\text{argmax}}\, w_i$$

Finally, the *BSL* contains a set of English-Chinese word vector pairs, where each entry represents an English social word and its Chinese pseudo-word based on its "translation set".

2.5 Constructing the SocVec Space

Let B_i denote the English word of the i^{th} entry of the *BSL*, and its corresponding Chinese pseudo-word is denoted by B_i^*. We can project the English word vector $\mathbf{E_W}$ into the *SocVec* space by computing the cosine similarities between $\mathbf{E_W}$ and each English word vector in *BSL* as values on SocVec dimensions, effectively constructing a new vector $\mathbf{S_W}$ of size L. Similarly, we map a Chinese word vector $\mathbf{C_U}$ to be a new vector $\mathbf{S_U}$. $\mathbf{S_W}$ and $\mathbf{S_U}$ belong to the same vector space *SocVec* and are comparable. The following equation illustrates the projection, and how to compute $ccsim$[3].

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

For example, if W is "Nagoya" and U is "名古屋", we compute the cosine similarities between "Nagoya" and each English social word in the *BSL* with their monolingual word embeddings in English. Such similarities compose $\mathbf{S}_{\text{nagoya}}$. Similarly, we compute the cosine similarities between "名 古 屋" and each Chinese pseudo-word, and compose the social word vector $\mathbf{S}_{名古屋}$.

In other words, for each culture/language, the new word vectors like S_W are constructed based on the monolingual similarities of each word to the vectors of a set of task-related words ("social words" in our case). This is also a significant part of the novelty of our transformation method.

3 Experimental Setup

Prior to evaluating *SocVec* with our two proposed tasks in Section 4 and Section 5, we present our preparation steps as follows.

Social Media Corpora Our English Twitter corpus is obtained from Archive Team's Twitter stream grab[4]. The Chinese Weibo corpus comes from Open Weiboscope Data Access[5] (Fu et al., 2013). Both corpora cover the whole year of 2012. We then randomly down-sample each corpus to 100 million messages where each message contains at least 10 characters, normalize the text (Han et al., 2012), lemmatize the text (Manning et al., 2014) and use LTP (Che et al., 2010) to perform word segmentation for the Chinese corpus.

Entity Linking and Word Embedding Entity linking is a preprocessing step which links various entity mentions (surface forms) to the identity of corresponding entities. For the Twitter corpus, we use Wikifier (Ratinov et al., 2011; Cheng and Roth, 2013), a widely used entity linker in English. Because no sophisticated tool for Chinese short text is available, we implement our own tool that is greedy for high precision. We train English and Chinese monolingual word embedding respectively using *word2vec*'s skip-gram method with a window size of 5 (Mikolov et al., 2013b).

Bilingual Lexicon Our bilingual lexicon is collected from *Microsoft Translator*[6], which translates English words to multiple Chinese words

[3]The function sim is a generic similarity function, for which several metrics are considered in experiments.

[4]https://archive.org/details/twitterstream

[5]http://weiboscope.jmsc.hku.hk/datazip/

[6]http://www.bing.com/translator/api/Dictionary/Lookup?from=en&to=zh-CHS&text=<input_word>

with confidence scores. Note that all named entities and slang terms used in the following experiments are excluded from this bilingual lexicon.

Social Word Vocabulary Our social word 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 (Tausczik and Pennebaker, 2009), but has more words and more categories and is publicly available. We manually select 91 categories of words that are relevant to human perception and psychological processes following Garimella et al. (2016a). OpinionFinder consists of words relevant to opinions and sentiments, and TextMind is a Chinese counterpart for Empath. In summary, we obtain 3,343 words from Empath, 3,861 words from OpinionFinder, and 5,574 unique social words in total.

4 Task 1: Mining cross-cultural differences of named entities

Task definition: This task is to discover and quantify cross-cultural differences of concerns towards named entities. Specifically, the input in this task is a list of 700 named entities of interest and two monolingual social media corpora; the output is the scores for the 700 entities indicating the cross-cultural differences of the concerns towards them between two corpora. The ground truth is from the labels collected from human annotators.

4.1 Ground Truth Scores

Harris (1954) states that the meaning of words is evidenced by the contexts they occur with. Likewise, we assume that the cultural properties of an entity can be captured by the terms they always co-occur within a large social media corpus. Thus, for each of randomly selected 700 named entities, we present human annotators with two lists of 20 most co-occurred terms within Twitter and Weibo corpus respectively.

Our annotators are instructed to rate the topic-relatedness between the two word lists using one of following labels: "very different", "different", "hard to say", "similar" and "very similar". We do this for efficiency and avoiding subjectivity. As the word lists presented come from social media messages, the social and cultural elements are already embedded in their chances of occurrence. All four annotators are native Chinese speakers but have excellent command of English and lived in the US extensively, and they are trained with many selected examples to form shared understanding of the labels. The inter-annotator agreement is 0.67 by Cohen's kappa coefficient, suggesting substantial correlation (Landis and Koch, 1977).

4.2 Baseline and Our Methods

We propose eight baseline methods for this novel task: **distribution-based** methods (BL-JS, E-BL-JS, and WN-WUP) compute cross-lingual relatedness between two lists of the words surrounding the input English and Chinese terms respectively ($\mathcal{L}_E$ and $\mathcal{L}_C$); **transformation-based** methods (LTrans and BLex) compute the vector representation in English and Chinese corpus respectively, and then train a transformation; MCCA, MCluster and Duong are three typical **bilingual word representation models** for computing general cross-lingual word similarities.

The $\mathcal{L}_E$ and $\mathcal{L}_C$ in the BL-JS and WN-WUP methods are the same as the lists that annotators judge. **BL-JS** (*Bilingual Lexicon Jaccard Similarity*) uses the bilingual lexicon to translate $\mathcal{L}_E$ to a Chinese word list $\mathcal{L}_E^*$ as a medium, and then calculates the Jaccard Similarity between $\mathcal{L}_E^*$ and $\mathcal{L}_C$ as J_{EC}. Similarly, we compute J_{CE}. Finally, we regard $(J_{EC} + J_{CE})/2$ as the score of this named entity. **E-BL-JS** (*Embedding-based Jaccard Similarity*) differs from BL-JS in that it instead compares the two lists of words gathered from the rankings of word embedding similarities between the name of entities and all English words and Chinese words respectively. **WN-WUP** (*WordNet Wu-Palmer Similarity*) uses Open Multilingual Wordnet (Wang and Bond, 2013) to compute the average similarities over all English-Chinese word pairs constructed from the $\mathcal{L}_E$ and $\mathcal{L}_C$.

We follow the steps of Mikolov et al. (2013a) to train a linear transformation (**LTrans**) matrix between *EnVec* and *CnVec*, using 3,000 translation pairs with maximum confidences in the bilingual lexicon. Given a named entity, this solution simply calculates the cosine similarity between the vector of its English name and the *transformed* vector of its Chinese name. **BLex** (*Bilingual Lexicon Space*) is similar to our *SocVec* but it does not use any social word vocabularies but uses bilingual lexicon entries as pivots instead.

MCCA (Ammar et al., 2016) takes two trained monolingual word embeddings with a bilingual lexicon as input, and develop a bilingual word em-

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
USC	college football, baseball, Stanford, Alabama, win, lose	top destination for overseas education, Chinese student murdered, scholars, economics, Sino American politics

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

bedding space. It is extended from the work of Faruqui and Dyer (2014), which performs slightly worse in the experiments. **MCluster** (Ammar et al., 2016) requires re-training the bilingual word embeddings from the two mono-lingual corpora with a bilingual lexicon. Similarly, **Duong** (Duong et al., 2016) retrains the embeddings from mono-lingual corpora with an EM-like training algorithm. We also use our BSL as the bilingual lexicon in these methods to investigate its effectiveness and generalizability. The dimensionality is tuned from $\{50, 100, 150, 200\}$ in all these bilingual word embedding methods.

With our constructed ***SocVec*** space, given a named entity with its English and Chinese names, we can simply compute the similarity between their *SocVec*s as its cross-cultural difference score. Our method is based on monolingual word embeddings and a BSL, and thus does not need the time-consuming re-training on the corpora.

4.3 Experimental Results

For qualitative evaluation, Table 1 shows some of the most culturally different entities mined by the SocVec method. The hot and trendy topics on Twitter and Weibo are manually summarized to help explain the cross-cultural differences. The perception of these entities diverges widely between English and Chinese social media, thus suggesting significant cross-cultural differences. Note that some cultural differences are time-specific. We believe such temporal variations of cultural differences can be valuable and beneficial for social studies as well. Investigating temporal factors of cross-cultural differences in social media can be an interesting future research topic in this task.

In Table 2, we evaluate the benchmark methods and our approach with three metrics: Spearman and Pearson, where correlation is computed between truth averaged scores (quantifying the labels from 1.0 to 5.0) and computed cultural difference scores from different methods; Mean Average Precision (MAP), which converts averaged scores as binary labels, by setting 3.0 as the threshold. The ***SocVec:opn*** considers only OpinionFinder as the ESV, while ***SocVec:all*** uses the union of Empath and OpinionFinder vocabularies[7].

Method	Spearman	Pearson	MAP
BL-JS	0.276	0.265	0.644
WN-WUP	0.335	0.349	0.677
E-BL-JS	0.221	0.210	0.571
LTrans	0.366	0.385	0.644
BLex	0.596	0.595	0.765
MCCA-BL(100d)	0.325	0.343	0.651
MCCA-BSL(150d)	0.357	0.376	0.671
MCluster-BL(100d)	0.365	0.388	0.693
MCluster-BSL(100d)	0.391	0.425	0.713
Duong-BL(100d)	0.618	0.627	0.785
Duong-BSL(100d)	0.625	0.631	0.791
SocVec:opn	0.668	0.662	**0.834**
SocVec:all	**0.676**	**0.671**	**0.834**
SocVec:noun	0.564	0.562	0.756
SocVec:verb	0.615	0.618	0.779
SocVec:adj.	0.636	0.639	0.800

Table 2: Comparison of Different Methods

Lexicon Ablation Test. To show the effectiveness of social words versus other type of words as the bridge between the two cultures, we also compare the results using sets of nouns (***SocVec:noun***), verbs (***SocVec:verb***) and adjectives (***SocVec:adj.***). All vocabularies under comparison are of similar sizes (around 5,000), indicating that the improvement of our method is significant. Results show that our *SocVec* models, and in particular, the *SocVec* model using the social words as cross-lingual media, performs the best.

[7]The following tuned parameters are used in *SocVec* methods: 5-word context window, 150 dimensions monolingual word vectors, cosine similarity as the *sim* function, and "*Top*" as the pseudo-word generator.

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**

Table 3: Different Similarity Functions

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**

Table 4: Different Pseudo-word Generators

Similarity Options. We also evaluate the effectiveness 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*). From Table 3, we conclude that among these four options, *Cos* and *L2+E* perform the best.

Pseudo-word Generators. Table 4 shows effect of using four pseudo-word generator functions, from which we can infer that "*Top*" generator function performs best for it reduces some noisy translation pairs.

5 Task 2: Finding most similar words for slang across languages

Task Description: This task is to find the most similar English words of a given Chinese slang term in terms of its slang meanings and sentiment, and vice versa. The input is a list of English/Chinese slang terms of interest and two monolingual social media corpora; the output is a list of Chinese/English word sets corresponding to each input slang term. Simply put, for each given slang term, we want to find a set of the words in a different language that are most similar to itself and thus can help people understand it across languages. We propose Average Cosine Similarity (Section 5.3) to evaluate a method's performance with the ground truth (presented below).

5.1 Ground Truth

Slang Terms. We collect the Chinese slang terms from an online Chinese slang glossary[8] consisting of 200 popular slang terms with English explanations. For English, we resort to a slang word list from OnlineSlangDictionary[9] with explanations and downsample the list to 200 terms.

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

Gg	Bi	Bd	CC	LT
18.24	16.38	17.11	17.38	9.14
TransBL	MCCA	MCluster	Duong	SV
18.13	17.29	17.47	20.92	**23.01**

(a) Chinese Slang to English

Gg	Bi	Bd	LT	TransBL
6.40	15.96	15.44	7.32	11.43
MCCA	MCluster	Duong	SV	
15.29	14.97	15.13	**17.31**	

(b) English Slang to Chinese

Table 5: ACS Sum Results of Slang Translation

Truth Sets. For each Chinese slang term, its truth set is a set of words extracted from its English explanation. For example, we construct the truth set of the Chinese slang term "二百五" by manually extracting significant words about its slang meanings (bold) in the glossary:

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

Similarly, for each English slang term, its Chinese word sets are the translation of the words hand picked from its English explanation.

5.2 Baseline and Our Methods

We propose two types of baseline methods for this task. The first is based on well-known *online translators*, namely Google (Gg), Bing (Bi) and Baidu (Bd). Note that experiments using them are done in August, 2017. Another baseline method for Chinese is CC-CEDICT[10] (CC), an online public Chinese-English dictionary, which is constantly updated for popular slang terms.

Considering situations where many slang terms have literal meanings, it may be unfair to retrieve target terms from such machine translators by solely inputing slang terms without specific contexts. Thus, we utilize example sentences of their slang meanings from some websites (mainly from Urban Dictionary[11]). The following example shows how we obtain the target translation terms for the slang word "fruitcake" (an insane person):

Input sentence: *Oh man, you don't want to date that girl. She's always drunk and yelling. She is a total* ***fruitcake***.[12]

[9] http://onlineslangdictionary.com/word-list/
[10] https://cc-cedict.org/wiki/
[11] http://www.urbandictionary.com/
[12] http://www.englishbaby.com/lessons/4349/slang/fruitcake

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", people paid to slander competitors on the Internet and to help shape public opinion	Water army	Navy	Navy	propaganda, complicit, fraudulent
floozy	a woman with a reputation for promiscuity	N/A	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),厌烦(annoying)

Table 6: Bidirectional Slang Translation Examples Produced by SocVec

Google Translation: 哦, 男人, 你不想约会那个女孩。她总是喝醉了, 大喊大叫。她是一个**水果蛋糕**。

Another lines of baseline methods is scoring-based. The basic idea is to score all words in our bilingual lexicon and consider the top K words as the target terms. Given a source term to be translated, the Linear Transform (LT), MCCA, MCluster and Duong methods score the candidate target terms by computing cosine similarities in their constructed bilingual vector space (with the tuned best settings in previous evaluation). A more sophisticated baseline (TransBL) leverages the bilingual lexicon: for each candidate target term w in the target language, we first obtain its translations T_w back into the source language and then calculate the average word similarities between the source term and the translations T_w as w's score.

Our *SocVec-based method* (**SV**) is also scoring-based. It simply calculates the cosine similarities between the source term and each candidate target term within *SocVec* space as their scores.

5.3 Experimental Results

To quantitatively evaluate our methods, we need to measure similarities between a produced word set and the ground truth set. Exact-matching Jaccard similarity is too strict to capture valuable relatedness between two word sets. We argue that average cosine similarity (ACS) between two sets of word vectors is a better metric for evaluating the similarity between two word sets.

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

The above equation illustrates such computation, where A and B are the two word sets: A is the truth set and B is a similar list produced by each method. In the previous case of "二百五" (Section 5.1), A is {foolish, stubborn, rude, impetuous} while B can be {imbecile, brainless, scumbag, imposter}. $\mathbf{A_i}$ and $\mathbf{B_j}$ denote the word vector of the i^{th} word in A and j^{th} word in B respectively. The embeddings used in ACS computations are pre-trained *GloVe* word vectors[13] and thus the computation is fair among different methods.

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

Table 7: Slang-to-Slang Translation Examples

Experimental results of Chinese and English slang translation in terms of the sum of *ACS* over 200 terms are shown in Table 5. The performance of online translators for slang typically depends on human-set rules and supervised learning on well-annotated parallel corpora, which are rare and costly, especially for social media where slang emerges the most. This is probably the reason why they do not perform well. The Linear Transformation (LT) model is trained on highly confident translation pairs in the bilingual lexicon, which lacks OOV slang terms and social contexts around them. The TransBL method is competitive because its similarity computations are within monolingual semantic spaces and it makes great use of the bilingual lexicon, but it loses the information from the related words that are not in the bilingual lexicon. Our method (SV) outperforms baselines by directly using the distances in the *SocVec* space, which proves that the *SocVec* well captures the cross-cultural similarities between terms.

To qualitatively evaluate our model, in Table 6, we present several examples of our translations for Chinese and English slang terms as well as their

[13] https://nlp.stanford.edu/projects/glove/

explanations from the glossary. Our results are highly correlated with these explanations and capture their significant semantics, whereas most online translators just offer literal translations, even within obviously slang contexts. We take a step further to directly translate Chinese slang terms to English slang terms by filtering out ordinary (non-slang) words in the original target term lists, with examples shown in Table 7.

6 Related Work

Although social media messages have been essential resources for research in computational social science, most works based on them only focus on a single culture and language (Petrovic et al., 2010; Paul and Dredze, 2011; Rosenthal and McKeown, 2015; Wang and Yang, 2015; Zhang et al., 2015; Lin et al., 2017). Cross-cultural studies have been conducted on the basis of a questionnaire-based approach for many years. There are only a few of such studies using NLP techniques.

Nakasaki et al. (2009) present a framework to visualize the cross-cultural differences in concerns in multilingual blogs collected with a topic keyword. Elahi and Monachesi (2012) show that cross-cultural analysis through language in social media data is effective, especially using emotion terms as culture features, but the work is restricted in monolingual analysis and a single domain (love and relationship). Garimella et al. (2016a) investigate the cross-cultural differences in word usages between Australian and American English through their proposed "socio-linguistic features" (similar to our social words) in a supervised way. With the data of social network structures and user interactions, Garimella et al. (2016b) study how to quantify the controversy of topics within a culture and language. Gutiérrez et al. (2016) propose an approach to detect differences of word usage in the cross-lingual topics of multilingual topic modeling results. To the best of our knowledge, our work for Task 1 is among the first to mine and quantify the cross-cultural differences in concerns about named entities across different languages.

Existing research on slang mainly focuses on automatic discovering of slang terms (Elsahar and Elbeltagy, 2014) and normalization of noisy texts (Han et al., 2012) as well as slang formation. Ni and Wang (2017) are among the first to propose an automatic supervised framework to monolingually explain slang terms using external resources. However, research on automatic translation or cross-lingually explanation for slang terms is missing from the literature. Our work in Task 2 fills the gap by computing cross-cultural similarities with our bilingual word representations (*SocVec*) in an unsupervised way. We believe this application is useful in machine translation for social media (Ling et al., 2013).

Many existing cross-lingual word embedding models rely on expensive parallel corpora with word or sentence alignments (Klementiev et al., 2012; Kočiský et al., 2014). These works often aim to improve the performance on monolingual tasks and cross-lingual model transfer for document classification, which does not require cross-cultural signals. We position our work in a broader context of "monolingual mapping" based cross-lingual word embedding models in the survey of Ruder et al. (2017). The SocVec uses only lexicon resource and maps monolingual vector spaces into a common high-dimensional third space by incorporating social words as pivot, where orthogonality is approximated by setting clear meaning to each dimension of the *SocVec* space.

7 Conclusion

We present the SocVec method to compute cross-cultural differences and similarities, and evaluate it on two novel tasks about mining cross-cultural differences in named entities and computing cross-cultural similarities in slang terms. Through extensive experiments, we demonstrate that the proposed lightweight yet effective method outperforms a number of baselines, and can be useful in translation applications and cross-cultural studies in computational social science. Future directions include: 1) mining cross-cultural differences in general concepts other than names and slang, 2) merging the mined knowledge into existing knowledge bases, and 3) applying the SocVec in downstream tasks like machine translation.[14]

Acknowledgment

Kenny Zhu is the contact author and was supported by NSFC grants 91646205 and 61373031. Seung-won Hwang was supported by Microsoft Research Asia. Thanks to the anonymous reviewers and Hanyuan Shi for their valuable feedback.

[14]We will make our code and data available at https://github.com/adapt-sjtu/socvec.

References

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

José Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proc. of SemEval@ACL*.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proc. of COLING 2010: Demonstrations*.

Xiao Cheng and Dan Roth. 2013. Relational inference for wikification. In *Proc. of EMNLP*.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proc. of HLT-EMNLP*.

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

Mohammad Fazleh Elahi and Paola Monachesi. 2012. An examination of cross-cultural similarities and differences from social media data with respect to language use. In *Proc. of LREC*.

Hady Elsahar and Samhaa R Elbeltagy. 2014. A fully automated approach for arabic slang lexicon extraction from microblogs. In *Proc. of CICLing*.

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

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proc. of CHI*.

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

Rui Gao, Bibo Hao, He Li, Yusong Gao, and Tingshao Zhu. 2013. Developing simplified chinese psychological linguistic analysis dictionary for microblog. In *Proceedings of International Conference on Brain and Health Informatics*. Springer.

Elisabeth Gareis and Richard Wilkins. 2011. Love expression in the united states and germany. *International Journal of Intercultural Relations*, 35(3):307–319.

Aparna Garimella, Rada Mihalcea, and James W. Pennebaker. 2016a. Identifying cross-cultural differences in word usage. In *Proc. of COLING*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2016b. Quantifying controversy in social media. In *Proc. of WSDM*.

E. Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *TACL*, 4:47–60.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proc. of EMNLP-CoNLL*.

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

Shinobu Kitayama, Hazel Rose Markus, and Masaru Kurokawa. 2000. Culture, emotion, and well-being: Good feelings in japan and the united states. *Cognition & Emotion*, 14(1):93–124.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proc. of COLING*.

Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. In *Proc. of ACL*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of ICML*.

Bill Y. Lin, Frank F. Xu, Zhiyi Luo, and Kenny Q. Zhu. 2017. Multi-channel bilstm-crf model for emerging named entity recognition in social media. In *Proc. of W-NUT@EMNLP*.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proc. of ACL*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proc. of ACL (System Demonstrations)*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*.

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

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard english words and phrases. In *Proc. of IJCNLP*.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Proc. of ICWSM*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Proc. of HLT-NAACL*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proc. of ACL*.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proc. of SIGDIAL*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*.

Chandar A P Sarath, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proc. in NIPS*.

Yla R. Tausczik and James W. Pennebaker. 2009. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. In *Proc. of ACL*.

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

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proc. of EMNLP*.

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Sujian Li, Chin-Yew Lin, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bülent Yener. 2015. Context-aware entity morph decoding. In *Proc. of ACL*.