56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)

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
[2] Yonsei University, Seoul, Republic of Korea

SHANGHAI JIAO TONG UNIVERSITY

Introduction

Example questions in cross-lingual understanding:

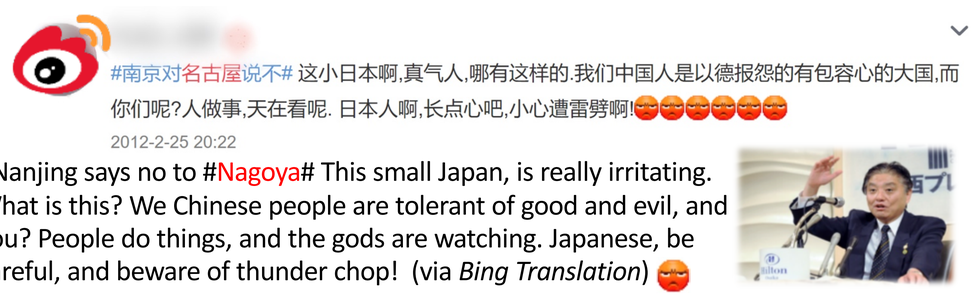
Named Entities: *Were there any* ***cross-cultural differences*** *between* ***Nagoya*** *(a city in Japan) for native English speakers and 名古屋 (Nagoya in Chinese) for Chinese people in 2012? (see the right figure)*

Slang terms: *What English terms can be used to explain "浮云" (a Chinese slang term, which literally means "floating cloud" but now equals to "nothingness")?* → ***cross-cultural similarities***

Motivation:

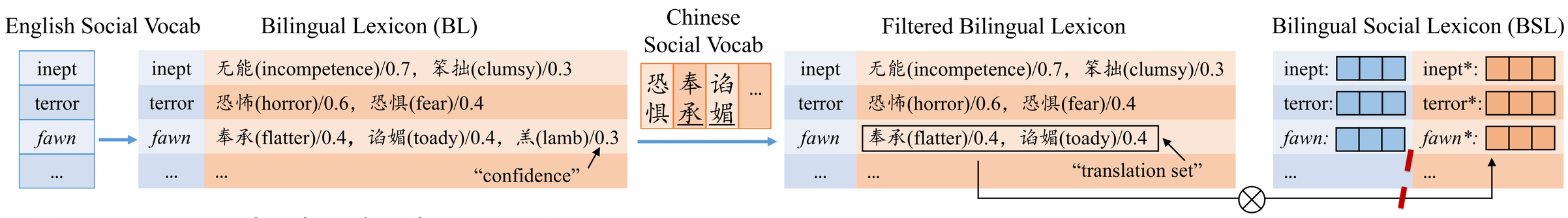
Enabling intelligent agents to understand such **cross-cultural knowledge** can benefit their performances in various **cross-lingual language processing tasks**.

The core problem: **how to compute cross-cultural differences (or similarities) between two terms from different cultures.**



Back in 2012, 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***, because **the city mayor** denied the truthfulness of the *Nanjing Massacre*.

The *SocVec* Framework

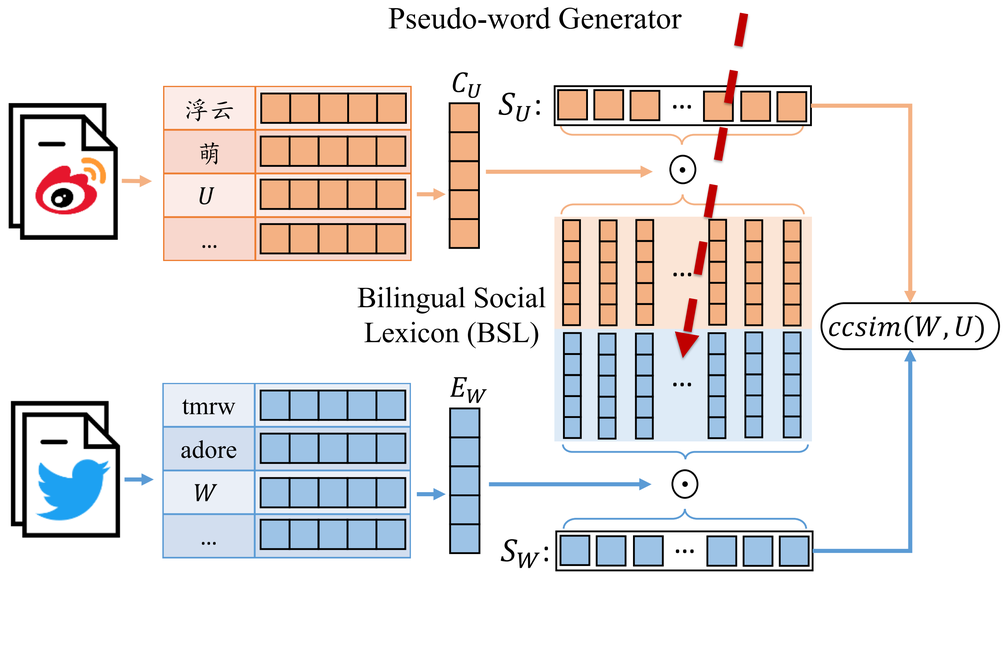


To compute cross-cultural word similarities, our SocVec:

1) firstly builds a Bilingual Social Lexicon (BSL) from a bilingual lexicon and the two monolingual "social vocabularies" in English and Chinese;
2) develops a bilingual vector space (named **SocVec** space) based on the cosine similarities between monolingual embeddings and BSL words.

"Social Words" are the words directly reflecting opinion, sentiment, cognition and other human psychological processes, which are important to capturing cultural and social characteristics. We use the lexicons from Empathy, OpinionFinder, and TextMind.

$$
\begin{aligned}
ccsim(W,U) &:= f(\mathbf{E}_W, \mathbf{C}_U) \\
&= sim\left(\begin{bmatrix} cos(\mathbf{E}_W, \mathbf{E}_{B_1}) \\ \vdots \\ cos(\mathbf{E}_W, \mathbf{E}_{B_L}) \end{bmatrix}^T, \begin{bmatrix} cos(\mathbf{C}_U, \mathbf{C}_{B_1^*}) \\ \vdots \\ cos(\mathbf{C}_U, \mathbf{C}_{B_L^*}) \end{bmatrix}^T\right) \\
&= sim(\mathbf{S}_W, \mathbf{S}_U)
\end{aligned}
$$



Evaluation

Task 1: Mining cross-cultural differences of named entities

This task is to discover and quantify cross-cultural differences of concerns towards named entities.

Input: a list of 700 named entities of interest and two monolingual social media corpora;
Output: the scores for the 700 entities indicating the cross-cultural difference. The ground truth is from the labels collected from human annotators.

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
MCluster-BL(100d)	0.365	0.388	0.693
Duong-BL(100d)	0.618	0.627	0.785
SocVec:all	**0.676**	**0.671**	**0.834**

Task 2: Finding most similar words for slang across languages

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.

Input: a list of English/Chinese slang terms and monolingual social media corpora;
Output: a list of Chinese/English word sets corresponding to each input slang term

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)

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

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.