Mining Cross-Cultural Differences of Named Entities: A Preliminary Study

Bill Y. Lin*
Shanghai Jiao Tong University
Shanghai, China
yuchenlin@sjtu.edu.cn

Frank F. Xu*
Shanghai Jiao Tong University
Shanghai, China
frankxu@sjtu.edu.cn

Kenny Q. Zhu
Shanghai Jiao Tong University
Shanghai, China
kzhu@cs.sjtu.edu.cn

ABSTRACT

It is very common that people of different cultures hold different opinions on the same named entities. Knowledge about such cross-cultural differences of named entities in existing knowledge bases can benefit a lot of downstream applications, especially for Computational Social Science. However, research on mining such knowledge is almost missing from the literature. In this paper, we propose this novel research topic with a preliminary study on collecting datasets and proposing several approaches.

1 INTRODUCTION

Opinions about a certain named entity, such as a famous person, a world-wide organization, or a place, may differ from culture to culture. For example, *Kashmir* is a large mountainous region on the China-India border. Due to decades of border disputes between China and India about that region, to the Chinese people, this region is almost synonymous to military conflicts and political struggles. On the contrary, that same region is considered as a picturesque travel destination by the westerners due to its perfect location in the Himalayas, since the border dispute between China and India is hardly their concern. This type of cross-cultural differences of named entities are evident from the most popular images about Kashmir on the English and Chinese search engines[1](see Figure 1).

Our goal in this paper is to identify an entity with significantly different cultural understanding, which can contribute to applications such as instant messenger or machine translator, to avoid culturally sensitive mentions or translations. Apart from these applications, a list of such entities with cultural differences in its own right is a valuable resource for cross-cultural studies. However, understanding subtle cultural differences requires not only perfect understanding of the two languages, but also devouring large volumes of biligual texts to sufficiently observe how they are mentioned in each culture and how they differ.

We transform this problem into a computational task, by proposing a quantitative evaluation metric measuring the cultural similarity between two cultures of a given named entity. To calculate such scores, we propose two approaches to compute the cultural similarity scores based on the word embeddings trained in each mono-lingual corpus receptively. The first solution connects the two results using linear transformation so that we can directly compute the cosine similarity between the English name and the Chinese name of a certain named entity. Another way is to construct a higher-dimensional vector space. Every dimension of this space is representing a pair of words, which are an English word and its corresponding Chinese translation. We call this space "translation space". The values in the English name vector of a certain entity are the cosine similarities between this entity's English name and each of other English words. We similarly compute the Chinese name vector. Because an English word may has many different Chinese translations, we duplicate the cosine similarities into several dimensions in translation space. After constructing this new comprehensive vector space, we can simply calculate cosine similarity between the English and Chinese vector of a certain entity.

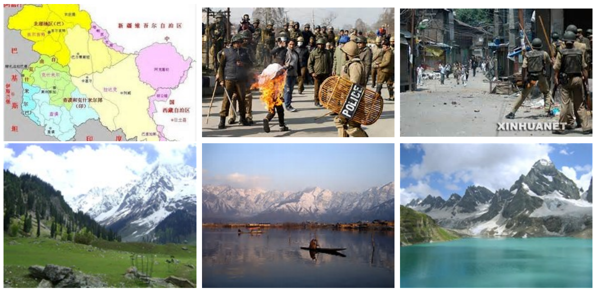


Figure 1: Popular images about Kashmir on Chinese web (top) and English web (bottom)

2 APPROACH

Our overall approach is illustrated in Figure 2. The starting point is a set of cross-lingual named entities harvested from English and Chinese Wikipedia/Wikidata, as well as a set of translation pairs of ordinary words from Bing translator API (Section 2.1). These two sets serve as our vocabulary. Then we conduct named entity recognition [3–6] and entity linking[1, 9, 12], which connects named entities harvested with text mentions in the English and Chinese corpora (Section 2.2). This step enables us to understand named entities in the distributional semantic space, by creating English and Chinese word vector spaces respectively using word2vec [8], for both named entities and ordinary words. Finally, with two separate methods (Section 2.3), we compute the cultural similarity scores for each cross-lingual entity pair by either linearly transforming words from the Chinese vector space into English or by merging the two spaces into a new, higher-dimension translation space. [2]

*The two authors contributed equally.

[1] Image search result from Bing on June 2017.

[2] Thanks to the salient cross-cultural differences between the east and the west, this paper mainly considers English and Chinese as the target languages. Nevertheless, the techniques developed here are language independent and thus can be used for any two natural languages so long as we have the necessary resources.

Conference'17, July 2017, Washington, DC, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

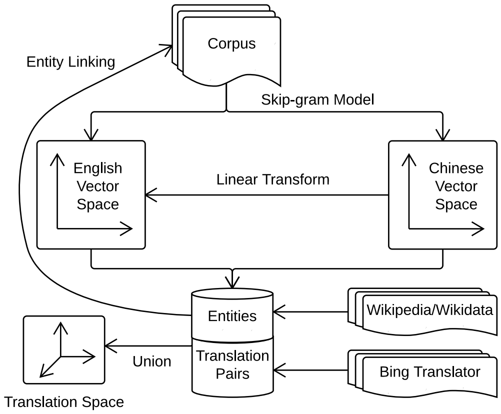


Figure 2: Overall Workflow

2.1 Vocabulary Building

Our vocabulary has two parts: i) a set of named entities of interest drawn from a standard ontology and ii) a set of ordinary English-Chinese word pairs. The obvious choice of this ontology is Wikipedia which keeps a unique identifier for every documented named entity. Many of these entities in Wikipedia have both English and Chinese instances and thus make up the first part of the vocabulary. The ordinary words from the two languages can be connected through online dictionary or translation APIs. We discuss each step in further details below.

2.1.1 *Named Entities.*

We focus on three categories of named entities, namely people, locations and organizations. We ensure that an entity is a person if it belongs to the Wikipedia category "Births by year"[3]. We consider an entity to be a location, if its Wikipedia page contains longitude-latitude coordinates. An entity is considered as an organization, if it appears under the subcategories of "Organization" in Wikidata while it carries a Wikipedia page. We use the inter-language links offered by Wikipedia to make sure all named entity exist both in English and Chinese Wikipedia.

2.1.2 *Translation Pairs.*

To construct the set of translation pairs of ordinary words, we first collect common English words from a large lemmatized English corpus (illustrated in Section 3.1) and translate these words into Chinese translations using online dictionary and translation APIs, specifically Bing[4] in this work. As each English word can be translated into multiple Chinese words, and a Chinese into multiple English words, this phase generates a many-to-many mapping.

2.2 Entity Linking

After preprocessing the corpora, we first do entity linking. For the English corpus, we utilize Wikifier [2, 10], a widely used entity linker to type the mentions of entities to Wikipedia entries. Because no suitable Chinese entity linking tool is available, we implement our own tool that is optimized for high precision. This tool prefers to link an entity with a surface form that appears more frequently in our corpus. The purpose of entity linking is to link mentions of entities of our interest in a large text corpus to our vocabulary. This enables to project a named entity in the distributional semantic space, together with ordinary words.

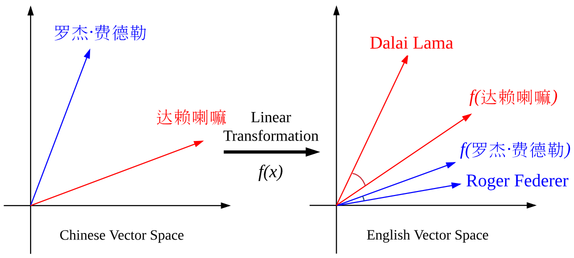


Figure 3: Linear transformation from Chinese to English

2.3 Cultural Similarity Computation

Next we introduce two algorithms for computing cultural similarity between the English vector and the Chinese vector of the same entity. The cultural difference can then be readily induced from the similarity.

2.3.1 *Linear-transformation Algorithm.*

English and Chinese vector spaces trained from the Skip-gram model are not directly comparable due to unknown meaning of each dimension. However, Mikolov et al. (2013) [7] have shown that the relationship between these vector spaces can be captured by rotation and scaling, represented by a linear transformation matrix W. In this paper, we borrow this idea and train this matrix using a number of human annotated "seed entities" with *little* cultural difference and using the following optimization problem:

$$\underset{W}{\text{argmin}} \sum_{i=1}^{n} ||Wx_i - t_i||^2, \qquad (1)$$

where x_i is a word in Chinese while t_i is its corresponding translation in English and n is the size of training samples.

With the linear transformation matrix from Chinese to English spaces, we can map each Chinese word vector to the English space so that two types of vectors are in the same coordinate. Figure 3 shows an illustrative example of how we linearly transformed the embedding space of one language to match with that of another language. This example shows that, after the transformation, both Chinese and English word vectors are in the same coordinate, while the angle between Chinese and English version of "Dalai Lama" is larger than that between Chinese and English version of "Roger Federer". This suggests that Dalai Lama, a controversial political figure, has a larger cultural difference in Chinese and English, than Roger Federer, a famous tennis player.

Despite the power of linear transformation, its performance strongly depends on the quality of the seed entities. However, obtaining high quality seed entities requires time and bilingual annotators. We thus propose an alternative unsupervised approach, called *translation space algorithm.*

2.3.2 *Translation Space Algorithm.*

This section combines English and Chinese semantic spaces into a rich and higher dimensional space , leveraging many-to-many translation pairs created in Section 2.1.

[3] https://en.wikipedia.org/wiki/Category:Births_by_year
[4] http://www.bing.com/translator

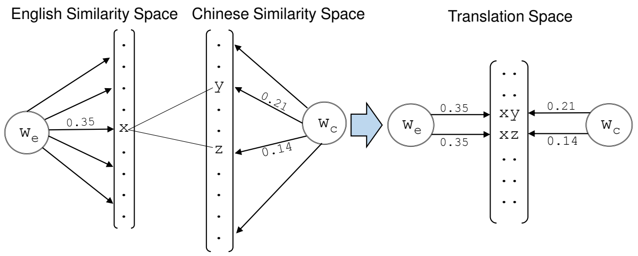


Figure 4: Translation Space Algorithm

Specifically, we first represent each entity in a word vector space by its cosine similarity with all tokens (including entities) in the same space, including itself. Suppose we want to compute the cultural similarity score for a pair of entities w_e and w_c in the English and Chinese vector spaces respectively. We first represent w_e by a *similarity vector* of size l_e where l_e is the total number of words/entities in the English space, and each dimension of this vector is the cosine similarity between w_e and all words in English. The cosine between w_e and itself is 1. We represent w_c similarly. Because the English and Chinese vocabularies are of different sizes, these two similarity vectors are of different sizes, too. Furthermore, since the translation is many to many, the two vectors are not directly comparable.

Our solution is to "expand" these two vector spaces in a higher dimensional space, where each dimension represents a translation from one English word to the corresponding Chinese word. As such, the new space, known as the "translation space", is k-dimensional, where k is equal to the total number of translation pairs or edges between the two vocabularies. In Figure 4, as an example, consider a word x in the similarity vector of w_e. If x is translated to y and z in Chinese, without prior information, we assume x is translated to y and z with equal probability.[5] As a result, the dimension for x is then expanded into two new dimensions, namely xy and xz, where each dimension stores the same value as the value for x.

At this point, the similarity vectors of w_e and w_c are mapped to the new translation space and are now comparable. Now we can calculate the cosine similarity between w_e and w_c pairs in the translation space as the cultural similarity score between the two entities.

3 EVALUATION

This section evaluates the performance of our approach to mine cultural differences of named entities from large text. First of all, we introduce the details in building the English and Chinese corpus. Then, we present how we construct the ground truth from human annotation. Finally, we report the evaluation metrics we use to evaluate our approach as well as the experiment results.

3.1 Data Preparation

To build English corpus, we crawled news articles from Daily Mail and New York Times published between Jan 1st, 2012 to Aug 5th, 2016, for these two sources are among the most representative news media of western cultures. Similarly, we crawled China News and iFeng News in the same time period to build our Chinese corpus. In total, there are 1,857,581 English news and 673,655 Chinese news. An average English news has 558.2 words while the average length of a Chinese news is 507.3 words.[6]

3.2 Ground Truth

As shown in Section 1, cultural differences of a given entity is visible from the most popular images in the image search results. It is because that people from different cultures have different views on the same entity so the kind of images that they search or create on the Internet are very different, too. With the help of the online image search engine such as Bing, we can get the most interesting images of a given named entity in western culture with Bing's global site and in Chinese culture with its Chinese site.

Thus, we obtain manual labels of 885 named entities by showing human annotators the top 20 pictures of a certain named entity from global Bing image search and the Chinese Bing image search respectively. This set of 885 entities is the intersection of the 2000 most frequent entities in the English corpus and Chinese corpus respectively. We invited 14 annotators from different cultures (both Chinese and international students) to judge whether the two sets of image search results of a given entity are visually different, without considering the actual meaning of the entity.

We choose 497 entities that most annotator agree on as our evaluation ground truth. The inter-annotator agreement by Cohen's kappa coefficient among these annotator is 0.6. Among these annotated pairs, we set aside 100 entities for which all annotators consider culturally similar. These are used as the training set for the linear transformation model. Consequently, our final test dataset consists of 397 entities, out of which 173 are labeled as culturally different and 224 labeled as culturally similar. Considering the percentage of annotators who label each pair as similar, we obtain the scores of each entity. Thus, we propose a ranking-based evaluation to investigate the performance of our method.

3.3 Baselines

We compare our approach with two baseline methods:

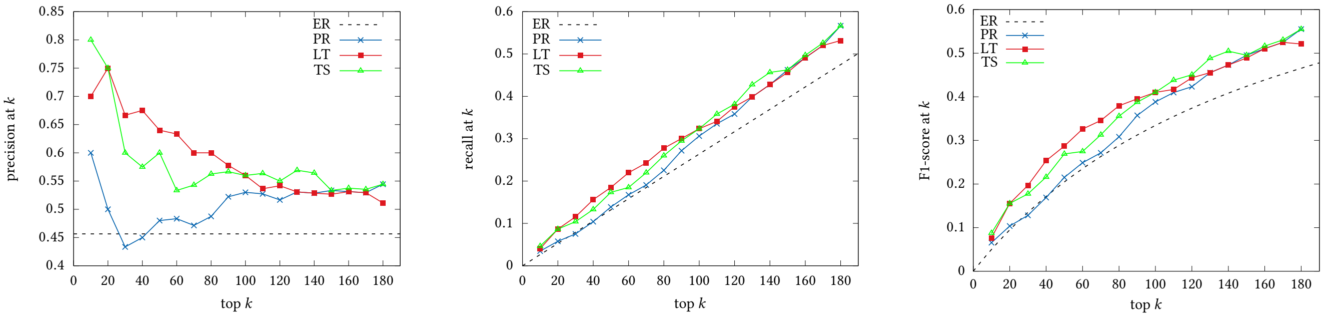
3.3.1 Biased Random Classifier. To judge whether a named entity is culturally different or not is actually a classification problem. Thus, a biased random classifier with a prior probability computed by the ratio of the number of culturally different entities to the total number of entities in the training data can be a simple baseline.

3.3.2 Ranking by Popularity. A stronger baseline is to rank the entities in test dataset by the sum of the relative frequency of this entity in the English corpus and Chinese corpus. [7]

[5] Admittedly, this is an over-simplified assumption. However, in our preliminary experiments, we found that considering translation confidence scores from Bing Translator as weights did not help improve the performance.

[6] Our vocabulary contains 45,740 English terms and 47,854 Chinese terms, including 4,212 terms representing the named entities common to two term sets. For the purpose of implementing the Translation Space algorithm, we build 122,284 translation pairs between the two term sets.

[7] The reason why it is stronger than the former one is that when a named entity is more likely to occur in different cultures, it has more chance to be viewed in different ways. If an object is not very common in different cultures, it has almost no opportunity to be exposed to multiple cultural views. Based on this assumption, we consider this baseline is a stronger competitor to our two algorithms.



(a) Precision at top k

(b) Recall at top k

(c) F1-score at top k

Figure 5: The precision, recall and F1-score at top k of the 4 methods. (ER=Expected Random, PR=Popularity Ranking, LT=Linear Transform, TS=Translation Space)

3.4 Experimental Results

3.4.1 Entity Linking Accuracy. In order to see the performance of our entity linking method, we randomly sampled 50 pieces of English news and Chinese news respectively. On the whole, there are 1,530 links in English samples, 50 among which are incorrect. Chinese samples contain 436 links and there are 23 errors. We thus achieve accuracies of **96.7%** and **94.7%**, respectively.

3.4.2 Word-embedding Results. To evaluate the correctness of our word embedding results of entities, we show qualitative results of high cosine similarity neighbors. To illustrate, Table 1 shows the top similar entities of "Adolf Hitler" in the two cultures, including similar semantic information with Benito Mussolini, Nazi Germany and the word "dictator".

English Space	Sim.	Chinese Space	Sim.
Hitler	0.929	*Nazi Germany*	0.869
Benito Mussolini	0.827	Nazi	0.811
Fuhrer	0.817	*Nazi Party*	0.769
Stalin	0.798	Napoleon	0.753
Nazi Germany	0.790	Stalin	0.729
Nazi	0.774	*Benito Mussolini*	0.716
Heinrich Himmler	0.751	dictator	0.704

Table 1: Top 7 most similar terms to the named entity "Adolf Hitler" by cosine similarity. (The Chinese terms in the table have already been translated into English. The words in italic are named entities in our vocabulary.)

3.4.3 Precision, Recall and F1-score at Top k. We can regard our cross-cultural entity similarity mining experiment as a ranked retrieval problem. Figure 5a reports quantitative comparison of our two algorithms with the two baselines. Note the accuracy of Expected Random Classifier baseline is fixed as 173/379 and its recall-at-k as $k/379$, shown as a dotted line. In the figure, our two algorithms consistently outperform the two baselines, until k reaches 150 where all algorithms converge. Translational space performs comparably to Linear transform requiring seed annotation, and even outperforms when $k < 20$ or $k > 100$.

Our algorithms, focusing on precision, are comparable in terms of recall with baselines as shown in Figure 5b, such that in terms of F1-measure, we outperform the baselines in Figure 5c.

Table 2 reports the mean average precision (MAP) [11]. Biased random as a baseline achieves 0.456, which is improved by our two proposed algorithms by 35.3% and 34.2% respectively.

Method	MAP
Biased Random	0.456
Popularity Ranking	0.543
Linear Transform	0.612
Translation Space	**0.617**

Table 2: Performance comparison

Linear Transform	Translation Space
Bihar	Baltimore
Sichuan	Human Rights Watch
Gujarat	APEC
China Central Television	Beijing
West Bengal	Greenpeace
Madhya Pradesh	China Central Television
Korean Central News Agency	Korean Central News Agency
Bharatiya Janata Party	African Union

Table 3: Most culturally different named entities.

3.4.4 The Most Culturally Different Entities. Table 3 shows the most culturally different entities we mined from our two algorithms. As discussed in Section 1, entities in the list include entities located in China or neighboring countries (e.g., Bihar, Sichuan, CCTV and Korean Central News Agency), for which the volume of interests is significantly different in the two cultures. In the case of more common entities such as Beijing, it carries more political connotations for the westerners but is instead more of a cultural and geographic landmark for the Chinese people, which shows different directions of interest.

4 CONCLUSION

In this paper, we propose a new research topic in Information Extraction and Text Mining, and develop a framework to compute cross-cultural differences of named entities. Leveraging the quantity of corpus from news articles and the quality of named entity information from Wikipedia and Wikidata, we managed to come up with an approach of calculating the distance to represent cross-cultural differences.

5 FUTURE WORK

Firstly, popular images on search engines for different languages are a good source for detecting cultural difference. Also, the reliability of the human annotators can be improved by more annotators from more diverse cultures. Apart from that, the cultural differences were mined from news articles which often reflect the official opinions rather than the opinion of the masses. It would be interesting to do similar research using other text resources, such as the social media data. Obviously this poses new challenges as social media data is a lot noisier and more ambiguous.

REFERENCES

[1] Xiao Cheng and Dan Roth. 2013. Relational Inference for Wikification. In *EMNLP*. http://cogcomp.org/papers/ChengRo13.pdf

[2] Xiao Cheng and Dan Roth. 2013. Relational Inference for Wikification. In *EMNLP*.

[3] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 363–370.

[4] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

[5] Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2017. Empower Sequence Labeling with Task-Aware Neural Language Model. *arXiv preprint arXiv:1709.04109* (2017).

[6] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bidirectional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).

[7] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *arXiv.org* (Sept. 2013). arXiv:1309.4168v1

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*.

[9] L. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *ACL*. http://cogcomp.org/papers/RRDA11.pdf

[10] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *ACL*. Association for Computational Linguistics, 1375–1384.

[11] Hinrich Schütze. 2008. Introduction to Information Retrieval. In *Proceedings of the international communication of association for computing machinery conference*.

[12] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2015), 443–460.