

Identifying Cultural Differences through Multi-Lingual Wikipedia

Yufei Tian^{2,*}, Tuhin Chakrabarty^{1,3,†}, Fred Morstatter¹ and Nanyun Peng¹

¹Information Sciences Institute, University of Southern California

²Department of Automation, Tsinghua University

³Department of Computer Science, Columbia University

tyf16@mails.tsinghua.edu.cn, {tuhinc, fredmors, npeng}@isi.edu

Abstract

Understanding cross-cultural differences is an important application of natural language understanding. This problem is difficult due to the relativism between cultures. We present a computational approach to learn cultural models that encode the general opinions and values of cultures from multi-lingual Wikipedia. Specifically, we assume a language is a symbol of a culture and different languages represent different cultures. Our model can automatically identify statements that potentially reflect cultural differences. Experiments on English and Chinese languages show that on a held out set of diverse topics, including *marriage, gun control, democracy*, etc., our model achieves high correlation with human judgments regarding within-culture values and cultural differences.

1 Introduction

Languages and cultures have radical correlations as individuals communicate with each other by language, which carries the aspects of their culture: their prior beliefs, attitudes and values (Khaslavsky, 1998). Understanding cross-cultural differences plays a vital role in understanding expectations of people from different national and cultural backgrounds. For example, a proper cultural model enables better downstream applications such as more pragmatic machine translation or cultural-aware sentiment analysis.

However, as beliefs/values are tacitly shared within a culture, cultural knowledge is usually concealed in written texts. While there is increasing interest in teaching machines commonsense knowledge (Bosselut et al., 2019; Sap et al., 2019a; Huang et al., 2019; Sap et al., 2019b), little light has

*The research was conducted when the author interned at USC/ISI.

†Equal contribution.

English	<i>In contrast, unsafe abortions (those performed by unskilled individuals, with hazardous equipment, or in unsanitary facilities) cause 47,000 deaths and 5 million hospital admissions each year.</i>
Chinese	据不完全统计中国由于以往进行的一孩政策（每对夫妇只容许有一个子女），每年人工流产至少1300万例，位居世界第一。... 人们对堕胎的反对较受基督教教义影响的西方国家低，在中国计划生育政策的背景下，很多人采取自愿堕胎，部分是婴儿的性别因素（重男轻女）也有人被计生部门实行强制堕胎（如残疾等）。
Translated English Version	<i>Due to the one-child policy in China (each couple only allowed one child), China has at least 13 million abortions per year, ranking first in the world, ... Opposition to abortion is lower than in western countries influenced by Christianity. In the context of China's family planning policy, many people have abortions of their own free will, partly because of the gender of the baby (preference for sons) and partly because of forced abortions by family planning authorities (disability, etc.).</i>

Table 1: Wikipedia excerpts from English and Chinese Wikipedia page of Abortion exemplifying cultural difference on the same topic. Note that either selected sentence is unique in its own languages and non-existent in the page of the other language.

been shed on systematic approaches to represent knowledge about cultural differences. Recently, Lin et al. (2018) present a distributional approach to compute cross-cultural differences (or similarities) between two aligned terms. However, the approach focuses on word-level (named entities in particular) semantics and is hard to be generalized to other forms.

We propose to identify cultural models by recognizing *statements* that are culturally disagreed upon. Cultural models shape the social identities of those that ascribe to them (Geertz, 1973), and help members of that culture to interpret the world in a shared manner. Cultural models can inform a number of different categories of daily life, in-

cluding which foods are edible, how to form social ties, and which values are agreeable. In this work, we use natural language statements to represent cultural models. Statements such as “*Abortion should be made illegal*” can help to reflect cultural beliefs. This statement is viewed very differently by English-speaking and Chinese-speaking cultures. Despite controversiality within English-speaking cultures, support for this statement is much greater in English-speaking cultures than in Chinese-speaking ones where there has been a “wide awareness and acceptance of abortion [...] since antiquity” (Tien, 1987). Table 1 shows an example on this topic, where the Chinese and English Wikipedia pages reflect the underlying cultural differences between the two. Learning cultural models through text will enable researchers to understand differences between cultures with fineness. This is especially necessary with ad-hoc cultures emerging from interactions in daily life such as internet cultures.

Considering the fact that culture is intrinsic to language, we take different languages to represent different cultures.¹ We leverage Wikipedia to learn cultural models through text. Wikipedia has 301 languages, presents a diverse array of topics, and hence can serve as an ideal source to understand cultural differences. Despite Wikipedia’s overall goal of objectivity, it embeds latent cultural bias.

We summarize our contributions as follows:

- We propose Cultural Difference Identifier (CDI), a model that automatically learns cultural differences based on Wikipedia articles. Towards this, we develop a novel procedure for algorithmically generating negative samples (introduced in 3.1.2) based on Wikipedia.
- We design an evaluation framework to systematically study the efficiency of the proposed approach by testing our models on self-labeled opinions ranging through diverse topics including *Food, Cuisine, Savings, Festivals, Marriage, Corruption, Terrorism, Democracy* and *Privacy*.
- Comprehensive quantitative and qualitative studies show that our model outperforms multiple well-crafted baselines and achieves strong correlation with human judgements.

¹We acknowledge this is a simplified categorization; there are other subtleties. The core of our argument that, at a minimum, a language provides a coarse-grained representation of a culture.

2 Task Definition

In this paper, we focus on identifying statements that provoke different opinions from different cultures. As is discussed in the previous section, we take different languages as different cultures. We further focus on Chinese and English as our target languages and use them as examples throughout the paper, although the algorithm we propose is generally applicable to any culture pairs. We now formally define the terminologies we will be using throughout the paper.

Statement. A statement, s_i , is a sentence that expresses an fact or opinion towards a certain topic. We start with a list of multi-lingual statement pairs $\mathcal{S} = \{\mathcal{S}_{en}, \mathcal{S}_{cn}\}$, where $\mathcal{S}_{en} = \{s_{en1}, s_{en2}, \dots, s_{en_n}\}$ is a set of English statements, and $\mathcal{S}_{cn} = \{s_{cn1}, s_{cn2}, \dots, s_{cn_n}\}$ is a set of Chinese statements. Each Chinese statement s_{cn_i} ($i = 1, 2, \dots, n$) is the exact translation of its corresponding English statement s_{en_i} .

Cultural Model. Cultural model is a probabilistic model that gives a score to a statement. For s_i in each statement pair $\{s_{en_i}, s_{cn_i}\}$, a machine-generated cultural score $MP(s_i) \in [0, 1]$ is assigned to estimate s_i ’s probability of being accepted by its corresponding culture. Similarly, a human-annotated cultural acceptance score pair $HA(s_i) \in [0, 1]$ is assigned and considered as the ground truth of the extent to which s_i is accepted by its corresponding cultural background.

Cultural Difference. Finally, we define “cultural difference.” Let $\mathcal{D}_{model} \in [-1, 1]$ be the quantity of cultural difference predicted by machines, where

$$\mathcal{D}_{model_i} = MP(s_{en_i}) - MP(s_{cn_i}). \quad (1)$$

A positive \mathcal{D}_{model_i} indicates that the English model agrees more with the statement s_i than the Chinese model. Similarly, we denote $\mathcal{D}_{human} \in [-1, 1]$ as the quantity of cultural difference reported by human annotators:

$$\mathcal{D}_{human_i} = HA(s_{en_i}) - HA(s_{cn_i}). \quad (2)$$

3 Data Preparation

In this section, we describe the procedure of collecting and composing our cultural dataset from multi-lingual Wikipedia articles to train the cultural models. We then introduce a test dataset with statements containing opinions about a wide range of topics.

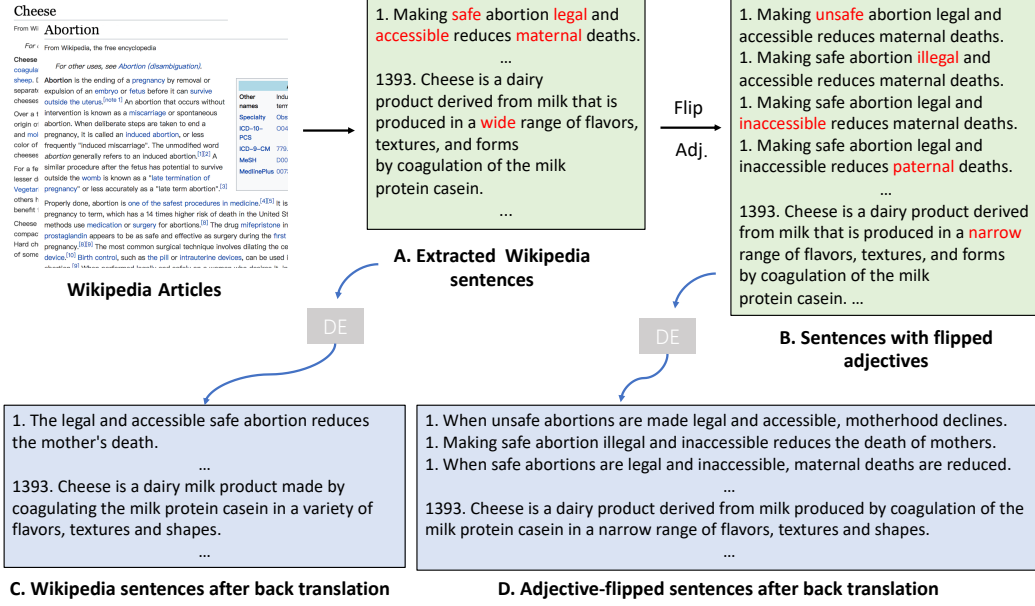


Figure 1: An illustration of the creation of the English training data by first extracting sentences from the retrieved Wikipedia articles to form the positive samples, and then replacing adjectives with their antonyms as negative samples. *Back-translation* (discussed in 4.2) is also used to resolve pattern bias among negative samples. For the Chinese language, the procedure is the same except that the pivot language we use in *back-translation* is Japanese.

3.1 Training Data

3.1.1 Topic Selection

We leverage the category hierarchy provided by Wikipedia to retrieve a list of child topics that belong to a few parent categories, including *Politics* (政治), *Foods* (饮食), *Sport* (体育运动), *Music* (音乐), *Literature*, *History* (历史), and *Social issues* (社会问题). The selected root categories in English and Chinese are aligned entities that are obtained from Wikipedia language links, and their sub-tree structures are only partially aligned. In this way, our obtained sub-topics in English and Chinese have considerable overlap but are not exactly the same and hence we have less subtopics and consequently lesser sentences from Chinese Wikipedia compared to English as seen in Table 2.

We then retrieve all the articles linked to the captured subtopics in English² and Chinese³ separately, so that and different opinions between the two cultures would be included.

3.1.2 Training Dataset Creation

Positive Examples. Upon our observation on the abortion page (in Table 1) and many other similar examples, we form our fundamental assumption that each sentence extracted from Wikipedia,

whether in English or Chinese, represents its corresponding cultural ideology. Therefore, we label each sentence extracted from the Wikipedia articles (obtained in section 3.1.1) as positive, as illustrated in part A of Figure 1.

Negative Examples. Although positive examples mirror their corresponding cultural ideology, we also need to compose negative samples, the statements that the corresponding cultures will disagree with. A natural approach is to flip the semantic meaning of positive samples. This could be achieved by replacing the adjectives in a sentence with their antonyms. As can be seen in Figure 1.A.1, there are four adjectives in the original text: ‘*Making **safe** abortion **legal** and **accessible** reduces **maternal** deaths.*’ Thus, we can obtain four negative sentences by replacing each of the adjectives with its antonym: ***unsafe**, **illegal**, **inaccessible** and **maternal***. Each of the four fabricated negative samples (in Figure 1.B) are ideal because it expresses conflicting standpoints against the original text.

However, certain collocations such as *New York* and *legal systems* are also converted into *Old York* and *illegal systems*, respectively. A basic statistical n-gram model can easily spot the poorly constructed sentences as negative, because bigrams such as *Old York* and *illegal systems* seldom appear in real sentences. To further improve the quality of our fabricated negative samples, we prohibit

²https://en.wikipedia.org/wiki/Wikipedia:Category_tree

³<https://zh.wikipedia.org/w/index.php?title=Special:分类树&target>

	Topics	Sentences	Positive Samples	Negative Samples
English	4,245	617,000	292,444	292,444
Chinese	1,563	64,020	57,904	57,904

Table 2: Statistics of Our Training Dataset. We deliberately balance the number of positive and negative samples so that no priori probability will intervene with the learning step.

common collocations that contain adjectives from being converted. We also neglect temporal phrases such as *in the early (late) 1850s* because they are considered less meaningful in our scenario.

So far, we have obtained all the data needed to train the cultural models. The number of topics, retrieved sentences, and training samples are listed in Table 2.

3.2 Out-of-domain Test Data

While training and testing on the same Wikipedia data is a possible choice, a more ideal scenario is to test on different data to see if the cultural representation learned by the model generalizes to other datasets, and is not a mere representation of the style of Wikipedia.

Selecting a good held out set for testing the performance of our models is hence very important. People often talk about their respective cultures on social media and in that process many opinionated claims or statements have been included. In previous works, Chakrabarty et al. (2019) collected a distant supervision-labeled corpus of 5.5 million opinionated claims covering a wide range of topics using sentences containing the acronyms IMO (in my opinion) or IMHO (in my humble opinion) from Reddit. Table 3 shows two examples from the IMO dataset that may reveal cultural difference. The only caveat is that this dataset is only in English. To get scores from the Chinese cultural model, we translate each sentence into Chinese using the Youdao api.

4 Methodology

In this section, we present the procedure of training our Cultural Difference Identifier (CDI) model. CDI is composed of two culture-sensitive classifiers: one for English and the other for Chinese. We then raise the issue of pattern bias in negative samples and provide our corresponding solution. Lastly, we introduce the inference process.

IMHO , what I find strange, and this is totally, some Chinese people have dogs as both pets and as dinner. ⁴
IMO , in an utopia Communism is the best system to live by.

Table 3: Sentence from the IMO dataset expressing opinions about which differ between cultures.

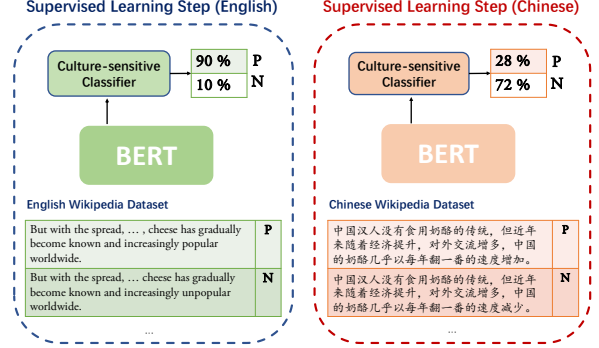


Figure 2: Diagram of Training Stage

4.1 Training Process

In the training stage, we leverage the pre-trained multilingual BERT (Devlin et al., 2018) and fine-tune it for the culture-specific classification task on the labeled data that is obtained in 3.1.2.

Figure 2 illustrates the training procedure. To enable the whole system to capture as much cultural discrepancy as possible, we fine-tune the English and Chinese systems on separate BERT models despite the multilingualism of BERT. In other words, the learning steps of English and Chinese systems have the exactly the same structure but are completely isolated from each other in terms of training data and model parameters.

4.2 Pattern Bias in Negative Samples and Targeted Improvements

While flipping adjectives to create negative samples appears as an obvious approach, it ends up introducing certain style biases. Since the placeholders for adjectives is the only difference between positive and negative samples in training data, most classifier would be able to identify this.

Niven and Kao (2019) show that high performance obtained from pre-trained language models such as BERT (Devlin et al., 2018) is often achieved by exploiting spurious statistical cues in the dataset. We faced a similar problem in our preliminary study when evaluating on a test set from a different domain. While the quantitative results of our models trained on Wikipedia data are extremely

⁴This does not reflect the opinion of the authors of this paper.

Testing \ Training	None Back-translated		Only Negative		Both Back-translated	
	Negative	Positive	Negative	Positive	Negative	Positive
None Back-translated	85.15	88.74	65.23	72.55	67.82	64.61
Only Negative	79.78	87.11	92.06	94.53	77.26	76.31
Both Back-translated	92.56	94.88	92.17	95.69	87.10	91.92

Table 4: Table showing F1 scores of Positive and Negative class when trained on Positive and Negative Samples from Wikipedia under three different settings and tested on the same.

high, we observe a huge drop when testing on out of domain data. This motivate us to mitigate any statistical cues in our data that our models might have learned.

Inspired by *back-translation* (Hoang et al., 2018), we generate paraphrases of our training data by introducing a pivot language and then translate the sentences back. In this way we are able to retain the semantics of the statements while removing existing stylistic biases. We back-translated both original Wikipedia sentences (i.e., positive samples) and the fabricated ones (i.e., negative samples). For English sentences, we use large BPE based transformer models fine-tuned for WMT19 News Translation Task (Ng et al., 2019)⁵ while for Chinese sentences we use Youdao api⁶. For English we use German as the pivot language while for Chinese we use Japanese as the pivot language. Part C and D of figure 1 show the back-translated version of our positive and negative samples respectively.

4.3 Inference Process

The framework of our inference stage is similar to the training procedure illustrated in Figure 2, except that we also test on out-of-domain data. For each statement s_i in test data, a model prediction pair $\{MP(s_{en_i}), MP(s_{cn_i})\}$ is generated. We then compute the cultural difference of s_i based on equation 1. We compute the correlation between model-predicted scores and human annotations.

5 Experimental Setup

We binarize the ground truth ($> 0.5 = 1$) of our experiments for the easiness of data collection. Here 0 represents that a culture tends to disagree with the statement, while 1 indicates a culture tends to agree with the statement. For Wikipedia sentences, which we use for training and in-domain evaluation as shown in Section 5.1, the sentences originally selected from Wikipedia are Positive (1) while the one we modified algorithmically are Negative (0).

⁵github.com/pytorch/fairseq/tree/master/examples/wmt19

⁶<https://ai.youdao.com/>

5.1 Hyper-Parameter Settings

For both the Chinese and English classifiers, we start the sentence representations with BERT-base model, which will be fine-tuned during the training process. The sequence length and number of training epochs for both languages are set to be 128 and 3. The batch size was set to 64 and learning rate $2e-5$.

We first study the efficiency of back-translation on reducing stylistic biases. We trained BERT (Devlin et al., 2018) models using data from 3 different settings:

1. No back-translate.
2. Back-translate only negative samples.
3. Back-translate both positive and negative samples.

5.2 Test Data Selection

Considering the fact that the IMO/IMHO dataset indiscriminately includes every sentence containing IMO/IMHO, it may contain many out-of-context statements. Hence, careful selection is needed before we can send the opinionated claims to human annotators.

Concretely, we first automatically extract statements that contain certain topical keywords, such as *privacy*, *democracy*, *noodles* and *cheese*, and then remove the extracted candidates which suffer from unREFERRED pronouns. Then we ask the English and Chinese volunteers to jointly select high-quality statements. A total of 128 high-quality statements are finally selected out of over 2000 candidates from the IMO/IMHO dataset spanning topics namely *Food*, *Cuisine*, *Savings*, *Festivals*, *Marriage*, *Corruption*, *Terrorism*, *Democracy* and *Privacy* for human annotation.

In parallel, we select 800 sentences from 16 different topics (50 sentences per topic). A few of the selected topics are mentioned above, along with additional topics such as *Abortion*, *Baseball*, and *Racism*. We do not collect human annotations for these sentences, but use them for qualitative analysis and visualization purposes detailed in Section 7.

	Pearson Correlation	Spearman Correlation
English Annotator	0.44 (0.35)	0.43 (0.35)
Chinese Annotator	0.48 (0.26)	0.47 (0.30)
Cultural Difference	0.29 (0.31)	0.30 (0.30)

Table 5: Inter-rater agreement, in the format of *corr* (*p-value*)

5.3 Human Evaluation

Recall that our goal is to achieve a higher correlation between machine-prediction and human-annotation. For each of the 128 claims, we collect 20 annotations from the United States using the Amazon Mechanical Turk (MTurk) platform. We collect another 20 annotations from Chinese netizens using the SurveyHero⁷ platform because MTurk is less known in China. The annotations are binarized, with 1 indicating an overall agreement, and 0 indicating an overall disagreement. For a given statement s_{en_i} , if 13 out of 20 English annotators give scores of 1, and rest 7 score them as 0, we then assume that the human-annotated score $HA(s_{en_i})$ is: $HA(s_{en_i}) = 13/20 = 0.65$. In this way, we ensure that the human annotations are of exactly the same scale and meaning as the model predictions, and thus prove the validity of using the correlation between model predictions and human annotations as a measurement of effectiveness.

6 Experimental Results

6.1 The Effects of Back-translation

Table 4 shows that the model trained on data that is back translated for both positive and negative Wikipedia samples outperforms the other models. It gives significant improvement in results when tested under the 3 possible settings and hence is ideal to be used for inference in other domains. It also shows that models trained on *None Back-translated* and *Only Negative* data, while working well under their own respective setting, doesn't transfer well to other settings.

6.2 Inter-annotator agreement

To show how the annotators within a culture agree with each other, we calculate the inter-annotator agreement using both Spearman and Pearson Correlations. We leverage attention questions and the inter-annotator correlation to remove irresponsible annotators. The final correlations of valid annotators are listed in Table 5.

⁷<https://www.surveyhero.com/>

Model	Correlation Type	English Annotator	Chinese Annotator	Cultural Difference
Random	Pearson	0.00 (0.50)	0.00 (0.50)	0.00 (0.50)
	Spearman	0.00 (0.50)	0.00 (0.50)	0.00 (0.50)
XLNet	Pearson	0.17 (0.05)	0.07 (0.42)	0.11 (0.23)
	Spearman	0.16 (0.08)	0.08 (0.35)	0.09 (0.30)
Weak CDI	Pearson	0.22 (0.01)	0.19 (0.03)	0.03 (0.73)
	Spearman	0.11 (0.23)	0.13 (0.14)	0.07 (0.42)
CDI	Pearson	0.37 (1e-5)	0.41 (1e-6)	0.25 (4e-3)
	Spearman	0.32 (2e-4)	0.34 (5e-4)	0.21 (0.01)

Table 6: Agreement between model predictions and human annotations, numbers in the format of *corr* (*p-value*)

For either Pearson (product moment) correlation or Spearman (rank-order) correlation, the coefficient within a culture is above 0.4, meaning that the annotators are moderately correlated within a culture.

6.3 Agreement between model prediction and human annotation

We compare our proposed Cultural Difference Identifier (CDI) with the following three baselines:

- **Random:** Monte Carlo method. Random numbers ranging from $[0, 1]$ are generated to simulate model predictions of English and Chinese cultural classifiers.
- **XLNet:** Competitive language model. We regard the average of word-level log probability (sentence log probability divided by length) generated by XLNet (Yang et al., 2019) and Chinese XLNet⁸ as model predictions. We then use min-max method to normalize the log probabilities.
- **Weak CDI:** Our proposed Cultural Difference Identifier, trained on Wikipedia sentences without *back-translation*.

Based on the correlations reported in Table 6, we observe that the Random method does not capture any cultural representations at all. A competitive language model such as XLNet can bring significant improvements over Random because it is trained on a very large NLP corpus (including English Wikipedia), where culture is implicitly included. Moreover, the performance of Weak CDI is partially better than language models, but still rather limited, probably due to the issue of style bias in negative samples. Finally, we can

⁸<https://github.com/ymcui/Chinese-PreTrained-XLNet>

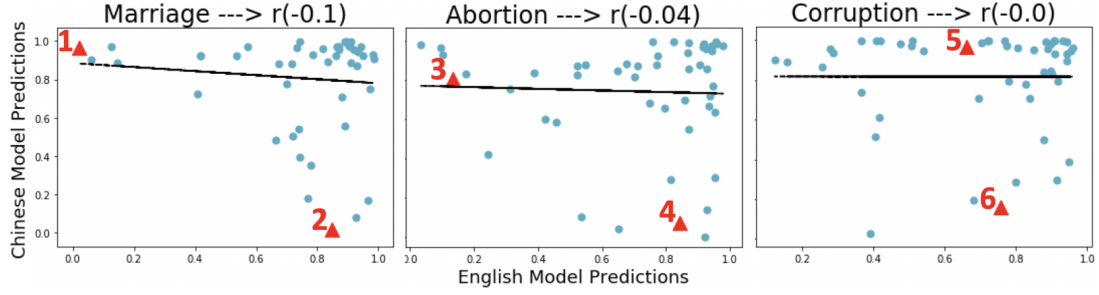


Figure 3: Plots of English Chinese model predictions on *Marriage*, *Abortion* and *Corruption*. Each dot (or triangle) represents one of the 50 statements randomly selected from IMHO, with x-axis representing the English model prediction and y-axis representing the Chinese model prediction. The triangular points in red are example sentences: **1.** *Marriage is not about meeting someone you connect to, but both people being matured, and in the same headspace.* **2.** *If he cannot share his concerns with her, he is poor marriage material.* **3.** *The only moral choice, is to make abortions legal and educate the populace so they are as rare as possible.* **4.** *There 's nothing wrong with keeping the baby, and doing so is always preferable to abortion.* **5.** *Corruption runs rampant everywhere, and having only two parties in the first place is a horrible way to conduct politics.* **6.** *With that said i feel everything is much worse now violence, corruption, war... the luxury of technology is the only thing that has changed.* Sentence 1, 3, and 5 are closer to the Chinese culture, while the English speakers tend to agree more with sentence 2, 4, and 6.

Model	English	Chinese	Cultural Difference
Random	0.50	0.50	0.50
LM (XLNet)	0.60	0.50	0.53
Weak CDI	0.69	0.55	0.45
CDI	0.73	0.61	0.58

Table 7: Binary Accuracy

find that CDI consistently outperforms all its competitors, and obtains 0.1050, 0.2493 and 0.1447 performance gains over the second best model for English, Chinese, and Cultural Difference respectively. The higher performance on Pearson Correlation indicates that the linear correlation is larger than the rank correlation.

Last, we want to point out that unlike many other NLP tasks, the inter-rater agreement (human performance) should not be viewed as golden in our evaluation. The listed inter-rater agreement is just an indicator of how unanimous the annotators are, and therefore machine-human correlation can reasonably be higher than within-human correlation.

6.4 Binary Accuracy

First, we calculate the number of instances that prediction and ground truths match with each other in a binary sense. For scores within a culture, the threshold is set to 0.5, to classify continuous scores into binary scores. For cross-culture scores ($\in [-1, 1]$), the threshold is set to 0, to classify continuous scores into binary scores. The results are shown in Table 7. Again, our CDI model achieved the best performance in all the three aspects: English/Chinese cultures and cultural differences.

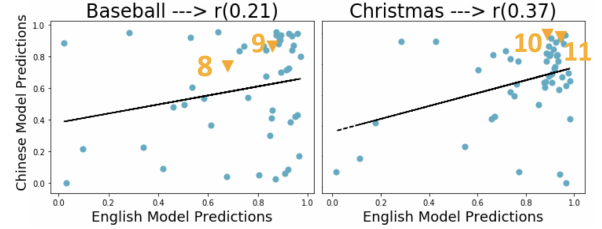


Figure 4: Plots of model predictions on *Baseball* and *Christmas*. The meaning of dots is the same as Figure 3. In addition, the orange triangles represent culturally agreed sentences and they are: **8.** *Cricket is as fun to play as baseball if you limit the "innings" or overs.* **9.** *Things like basketball, baseball, tennis, golf, etc are far more popular globally.* **10.** *Christmas, even minus the religious meanings, has good attributes in theory but has been too commercialized.* **11.** *Oh i believe in giving gifts to kids because, Christmas is for children.*

7 Qualitative Analysis

While Table 6 shows quantitative results and correlation with human judgments on chosen sentences, we want to further our understanding on the advantages of our model. To this end, we selected 50 statements from five particular topics: *Marriage*, *Abortion*, *Corruption*, *Christmas* and *Baseball*. We then obtained our model predictions on these sentences.

The visualization for each sentence score pairs can be found in Figure 3 and 4. For each visualization, those blue dots that fall along the diagonals are agreed by both models. On the contrary, the dots that fall on the upper left or the lower right part are disagreed across the English and Chinese model. We select representative examples in each region and list them in the captions.

First, we can observe that the English and Chinese models have zero or negative correlation on three topics: *Marriage*, *Abortion* and *Corruption*, meaning that perhaps Chinese speakers and English speakers view these topics very differently. Second, both cultures hold similar views on other topics, such as *Christmas* and *Baseball*. This is again consistent with our commonsense. For example, Christmas, which is not a traditional holiday in China, is adopted directly from the western world. Hence, Chinese people view Christmas very similarly to English speakers. The same reason applies to baseball.

8 Related Work

Online Disagreement Most work on social media about disagreement focus on a single culture or language (Sridhar et al., 2015; Wang and Yang, 2015; Sridhar et al., 2015; Rosenthal and McKown, 2015), so the differences in stance or agreement are restricted to a single group. While these works try to computationally model online disagreement or stance in debates, they are not targeted at finding cultural differences. We, on the other hand, want to understand cultural disagreement through different groups relying on their respective languages.

Cross Cultural Study in Blogs or Social Media Nakasaki et al. (2009) presented a framework to visualize the cross-cultural differences in concerns in multilingual blogs collected with a topic keyword. Elahi and Monachesi (2012) show that using emotion terms as culture features is effective to analyze cross-cultural difference in social media data, however it is only restricted to a single topic (love and relationship). We on the other hand use Wikipedia to study cross cultural difference on a much larger scale and do not restrict ourselves to one single topic.

Cross Cultural Difference in Word Usage Garimella et al. (2018) investigate the cross-cultural differences in word usages between Australian and American English through *socio-linguistic features* in a supervised way. Garimella et al. (2016) used social network structures and user interactions, to study how to quantify the controversy of topics within a culture and language. Gutiérrez et al. (2016) detect differences of word usage in the cross-lingual topics of multilingual topic modeling results. Lin et al. (2018)

present distributional approaches to compute cross-cultural differences (or similarities) between two terms from different cultures focusing primarily on named entities. Our work is not limited to word usage or any particular topic only. Instead, we focus on understanding cross-cultural difference at a sentence level.

Argumentation Chen et al. (2019) release a dataset of claims, perspectives and evidence and propose the task of substantiated perspective discovery where, given a claim, a system is expected to discover a diverse set of well-corroborated perspectives that take a stance with respect to the claim. Different interests, cultural backgrounds, and socializations make people disagree on taking a certain course of action. In argumentation, *Framing* is used to emphasize a specific aspect of a controversial topic. Ajjour et al. (2019) introduce frame identification, which is the task of splitting a set of arguments into non-overlapping frames. While both these work deal with different perspectives or frames about arguments (again in only English), our work focuses on them from a cultural point of view. We are interested in understanding the difference in cultural perspectives or framing.

9 Conclusion

We present CDI, a computational method to compute cross-cultural differences and evaluate CDI with human judgements. Through detailed experiments, we show that the proposed lightweight yet effective method outperforms a number of baselines. The general model of cultural difference identifier can be useful in translation applications as well as cross-cultural studies in computational social science. Furthermore, we showed how to take advantage of Wikipedia in multiple languages to understand cross cultural differences over other discussion forums like *Reddit* or *Weibo*.

There are several future directions for this work. Most notably, we wish to extend this pipeline to cultural understanding where the cultures are not defined by the language they speak (as is often the case in internet cultures). Our model could do this provided with explicit cultural labels, such as those provided for free in the case of subreddits. Other future directions include informing downstream NLP tasks, such as machine translation and sentiment analysis.

References

- Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2922–2932.
- Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. **COMET: Commonsense transformers for automatic knowledge graph construction**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.
- Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. **IMHO fine-tuning improves claim detection**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *arXiv preprint arXiv:1906.03538*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- 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 *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 4080–4086, Istanbul, Turkey. European Languages Resources Association (ELRA).
- Aparna Garimella, Rada Mihalcea, and James Pennebaker. 2016. **Identifying cross-cultural differences in word usage**. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 674–683, Osaka, Japan. The COLING 2016 Organizing Committee.
- Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):3.
- Clifford Geertz. 1973. *The interpretation of cultures*, volume 5019. Basic books.
- E Dario Gutiérrez, Ekaterina Shutova, Patricia Lichtenstein, Gerard de Melo, and Luca Gilardi. 2016. Detecting cross-cultural differences using a multilingual topic model. *Transactions of the Association for Computational Linguistics*, 4:47–60.
- Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. **Iterative back-translation for neural machine translation**. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.
- Julie Khaslavsky. 1998. Integrating culture into interface design. In *CHI 98 conference summary on Human factors in computing systems*, pages 365–366. ACM.
- Bill Yuchen Lin, Frank F. Xu, Kenny Zhu, and Seungwon Hwang. 2018. **Mining cross-cultural differences and similarities in social media**. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 709–719, Melbourne, Australia. Association for Computational Linguistics.
- Hiroyuki Nakasaki, Mariko Kawaba, Sayuri Yamazaki, Takehito Utsuro, and Tomohiro Fukuhara. 2009. Visualizing cross-lingual/cross-cultural differences in concerns in multilingual blogs. In *Third International AAAI Conference on Weblogs and Social Media*.
- Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair’s wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*.
- Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.
- 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 *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

- Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. [Joint models of disagreement and stance in online debate](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 116–125, Beijing, China. Association for Computational Linguistics.
- H. Yuan Tien. 1987. [Abortion in china: Incidence and implications](#). *Modern China*, 13(4):441–468.
- 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 *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.