The 27th International Conference on Computational Linguistics

COLING 2018

Author Response

Title: Mining Cross-Cultural Differences and Similarities in Social Media

Authors: Yuchen Lin, Frank Xu, Kenny Zhu and Seung-won Hwang

Instructions

You may now leave comments for chairs, regarding your reviews. Note that the reviewers may not see your responses, and will not change their reviews or scores from this point onwards. Your response only affects the area chairs' interpretation of reviews.

If you want to respond to the points raised in the reviews, you may do so in the boxes provided below. Please note: *you are not obligated to respond to the reviews*.

For reference, you may see the review form that reviewers used to evaluate your submission. If you do not see some of the filled-in fields in the reviews below, it means that they were intended to be seen only by the committee. See the review form HERE.

Review #1

Relevance (1-5):	5
Readability/clarity (1-5):	3
Clarity (a) - Hypothesis:	Yes, it is stated directly
Clarity (b) - Hypothesis tested:	Yes
Originality (experiment) (1-5):	4
Technical correctness/soundness (1-5):	3
Soundness (a):	The relationship between hypothesis and results is partially clear
Soundness (b):	No, it is not explained
Soundness (c):	Datasets are clearly described and appropriate
Reproducibility (1-5):	5
Data/code availability (1-5):	5
Error analysis (1-5):	1
Meaningful comparison (1-5):	4
Substance (1-5):	4

Detailed Comments

Summary: This paper investigates the question of cross-lingual differences and similarities between words. It presents a new method for estimating bilingual word representations, called SocVec, which uses only monolingual corpora, monolingual “social lexicons” (lists of “social words”), and a bilingual lexicon (list of word translations).

Strengths: The paper is very clearly written and easy to follow. The problem is well motivated and interesting, and the proposed work is clearly related to the original question. Extensive comparison, includes many related systems and multiple tasks. The SocVec model performs very well, and is also appealing for its simplicity and potential interpretability, since each feature in the new space corresponds to a Chinese and/or English word in the Bilingual Social Lexicon.

Based on the model description and the results, this could potentially be an influential work.

Weaknesses: The presentation of the SocVec model obscures its relation to previous work. Section 2.1 for example seems to frame the approach of projecting monolingual vectors into a bilingual space as an original concept, derived from first principles: “our intuition is thus to project English and Chinese word vectors to a common third space, known as SocVec.” But previous work, including some papers in the comparison, have already established this idea. Ideally, some similar original work, not just survey papers, should be cited early on to help make this point. But even just clarifying that although this general projection approach is done elsewhere, this work’s particular method is new, would be an improvement.

The paper provides extensive quantitative results, including good ablations on one task, but little to no analysis or explanation of those results. No error analysis of different models is included, and little understanding of the advantages of the different models, or discussion of the aspects of the proposed model that lead to its success.

For example: BLex is described as essentially an ablated version of SocVec that does not include the social lexicon information and uses bilingual lexicon entries in place of the BSL. The relation between BLex and SocVec seems much like that between e.g. MultiCluster-BL and MultiCluster-BSL. But it’s not presented that way in Section 4.2 or Table 4, where BLex is set apart from the various SocVec models and variations, and not mentioned as having its similarity method tuned like the SocVec models—have I misunderstood? In any case, the following points apply equally well to the SocVec:noun and similar models which are directly stated to be ablations of SocVec that don't use social information.

On its own, BLex outperforms all previous methods except that of Duong et al. and comes close to them. This is an intriguing result, since the motivation for the SocVec model was the use of *social* information and the BLex model contains none. What about the BLex model causes it to perform so well? Adding social information to create the SocVec model then adds another *7 points of MAP* and equally large improvements on the other metrics. Again, granted that the use of the social lexicon is important, but why does the use of social information help the BLex/SocVec model so dramatically, while having only minor positive effects (1-2 points MAP) on most other models? These questions should be at least addressed in a basic discussion.

The analysis of the second task is equally unhelpful. "Our method (SV) outperforms baselines by directly using the distances in our proposed bilingual embeddings" is no explanation, since the MultiCCA, MultiCluster, and Duong methods also directly use distances in a bilingual embedding space but perform comparably to the online translation baselines. Discussion and analysis of the results is needed to make this paper more than a system report.

The goal of the first task is somewhat vague. “Cross-cultural differences in concerns” plus the example in the intro suggest something like difference in sentiment, but it’s not obvious to me that lists of terms most frequently co-occurring with a named entity really captures that. It’s not clear that the topic-relatedness of the entities in the two corpora is a good measure of cross-cultural differences in views about the named entity.

Minor points (mostly about citations):

- The sentence “Dimensionality is tuned from {50, 100, 150, 200} in all methods.” appears in the MultiCCA paragraph in 4.2, but it applies (I think?) to all the methods in the section. It should go at the beginning or end of the section. Similarly with the use of both the BL and BSL.
- The sentence “Note that our method is more efficient because it requires no re-training on original corpora.” is misleading; almost all the models in the comparison share this property, only Duong et al. and MultiCluster excepted. This point should be clarified.
- MultiCCA is used in the comparison and Ammar et al. 2016 cited; however the original crosslingual CCA of Faruqui and Dyer 2014 was

only very slightly changed by MultiCCA. The original projects two sets of multilingual word vectors into an intermediate space. The later paper, cited in this work, projects the vectors of one language into the vector space of the other, in order to allow three or more languages to be projected into the same space using a bilingual projection method, a scenario not relevant to this work. Faruqui and Dyer might be a slightly better comparison for this method due to its symmetry vs. MultiCCA's asymmetric projection, but should be cited in either case.

- Artexe et al. 2016, by contrast, are a relevant update to crosslingual CCA.
- Ruder 2017 should be cited as Ruder et al. 2017—Ivan Vulić and Anders Søgaard were coauthors.
- Citations in general are somewhat inconsistently formatted, e.g. ACL vs Proc. of ACL.

Questions: Has a method similar to the SocVec model, in which new vectors are constructed based on the similarities of each word in a vocabulary to the vectors of a set of task related words, been used in a monolingual setting? This technique seems familiar, though I could not find a citation. If so, this should be clearly cited, as it is a significant part of the novelty of the model. If not, perhaps this point should be emphasized, since this method of transforming a feature space seems applicable to many tasks, both multi- and mono-lingual.

Conclusion: This is potentially a strong paper, limited by its lack of discussion and error analysis of the results. I think the model looks very interesting and would like to see it published but I'm doubtful about the current presentation.

Review #2

Relevance (1-5):	5
Readability/clarity (1-5):	4
Clarity (a) - Hypothesis:	Yes, it is stated directly
Clarity (b) - Hypothesis tested:	Yes
Originality (experiment) (1-5):	2
Technical correctness/soundness (1-5):	4
Soundness (a):	Hypothesis and results clearly relate
Soundness (b):	Yes, it is explained
Soundness (c):	Datasets are clearly described and appropriate
Reproducibility (1-5):	4
Data/code availability (1-5):	5
Error analysis (1-5):	4
Meaningful comparison (1-5):	2
Substance (1-5):	3

Detailed Comments

This paper takes on two tasks: mining cross cultural difference of named entities, and finding similar terms for slang across languages. They seek to be able to answer questions like: how did people from different cultures feel about X, and what's the english equivalent for a chinese slang term. Figure 1 is very small and the font is difficult to read. The authors propose a novel word-vector method called Social Vector, which is like a word vector. It attempts to deal with the gap that there are no parallel social media texts to do the above analysis. They list 3 contributions of the paper. The end of the introduction is disorganized, difficult to tell what each section is about and how the paper is structured.

Section 2: never begin a section without text. You need to add a paragraph of text between Section 2 and Section 2.1. Footnote #2 reads very loosely, using

an odd style of writing. Perhaps you could sharpen the language to be more precise. Instead of "Thanks to the salient cross-cultural differences.." use some other terms that are not as slang.

The authors describe why a simple translation won't work: slang is often OOV, similarity does not capture cross-cultural differences in language use, want to preserve context. This paragraph has an extra full-stop and does not make sense. Also remember that it takes 3 sentences to form a paragraph. SO if you have a paragraph of 1 sentence, like you do, then it is not a paragraph. The authors describe their categorical notations, though that seems to be misplaced in terms of the context of the paragraphs.

They build a BSL - bilingual social lexicon. The text in figure 3 is very small and hard to read. Not sure what the chinese term "toady" refers to. What is the meaning/context? Also it's not good practice to stack footnotes. Instead of putting two footnotes onto a sentence, just make 1 footnote. Footnotes 5 and 6 belong inside the main text.

section 4.1 - incorrect method of citing (Harris, 1954). wrong format

It's hard to read Table 2 - does it go across or down? This needs to be fixed.

The results are interesting, but what is your baseline to compare to? You need a baseline.

For task 2: are you trying to find a slang equivalent of a foreign slang term? Or are you trying to translate a slang term? What are you doing exactly?

In the conclusion, you have a sentence that is not actually a paragraph. As written, the entire conclusion section should be a single paragraph. A paragraph consists of at least 3 sentences.

Related work should go before experiments, having it right before the conclusion makes this paper *extremely* difficult to read and interpret. Also you should provide an overview of the structure of the paper in the introduction.

Review #3

Relevance (1-5):	5
Readability/clarity (1-5):	3
Clarity (a) - Hypothesis:	No
Clarity (b) - Hypothesis tested:	No
Originality (experiment) (1-5):	3
Technical correctness/soundness (1-5):	3
Soundness (a):	The relationship between hypothesis and results is partially clear
Soundness (b):	Yes, it is explained
Soundness (c):	Datasets lack important description or relevance
Reproducibility (1-5):	4
Data/code availability (1-5):	4
Error analysis (1-5):	1
Meaningful comparison (1-5):	3
Substance (1-5):	3

Detailed Comments

This paper presents a simple but comparatively effective method for 1) discovering cross-lingual differences in the perception of named entities (NE) and 2) the translation of slang words.

The authors address two non-standard tasks that are interesting and deserve attention. I think this is a strong point of the paper.

However, due to this, no standard evaluation sets could be used for evaluation. I think the evaluation is a weak point of this paper. The evaluation sets on which the performance of the methods is measured are constructed by the authors themselves. The first evaluation set constructed for evaluating the cross-lingual differences in the perception of NE is not well described. It is not clear to me what the human judges had to compare exactly. I guess it is the cooccurrences for the NE in the Chinese data against the cooccurrences for that same NE in the English data. I guess the same corpora used for building the models are used for finding the coocurrences. I wonder if this does not result in biases. It would be better to use a part of the corpus that is held out from the vector-based models. If all models in comparison use the same corpora (I assume they do?), this bias is the same for all, so comparisons still hold, but in absolute terms, these results are not that telling. The second evaluation set is selected from independent gold standards, but the selection of words from the definitions is done manually and the authors might have been involved in this. The paper does not say they are not. Why not just keep all words in order to avoid any bias? I think that an extrinsic evaluation would make this paper much stronger and the authors mention this under future work.

The authors do compare their results with a large number of systems and show the effect of using the 'social lexicon' instead of the general lexicon for these methods. This is a again a merit point for this paper. These methods should, however, be discussed in the related work section, so that is is clearer how the methods differ from the one proposed in this paper. The related work section now only reviews work that is related in the types of task it addresses. It should also review related work on the basis of the method used.

The difference between the method by Duong and the current method is not large and I wonder if the difference is significant given the small test set in the second task.

The performance figures do not tell us much about how good the system is. It just tells us it outperforms the systems in comparison. It would be useful to show an upper bound by comparing to human output. For example (average) cosine sim between human generated definitions and gold standard for task 2.

Also, the authors should make clear what hypotheses they are testing exactly with their experiments. Is it the sue of 'social lexicons' for these tasks or other aspects of the method for mapping the two spaces for the two languages?

Smaller points:

The method based on WordNet is not a distributional method. There are no distributions involved.

Please provide a motivation for generating pseudowords in 2.3

I find the term social word vocabulary strange. Why not use sentiment lexicon?

Writing:

The authors often omit determiners in their writings and some sentences are hard to follow. The description of E-BL-JS needs to be rephrased.

When references are part of a sentence, use ‘author names (year)’ and not ‘(author names, year)’.

Submit Response to Reviewers

Use the following boxes to enter your response to the reviews. Please limit the total amount of words in your comments to 400 words (longer responses will not be accepted by the system).

Response to Review #1:

Response to Review #2:

Response to Review #3:

Submit

START Conference Manager (V2.61.0 - Rev. 5257)