RepL4NLP 2020 Reviews for Submission #12

Title: Word Embeddings as Tuples of Feature Probabilities
Authors: Siddharth Bhat, Alok Debnath, Souvik Banerjee and Manish Shrivastava

REVIEWER #1

Reviewer's Scores

Appropriateness (1-5)	5
Clarity (1-5)	4
Originality / Innovativeness (1-5)	3
Soundness / Correctness (1-5)	4
Meaningful Comparison (1-5)	4
Thoroughness (1-5)	4
Recommendation (1-5)	4
Reviewer Confidence (1-5)	4

Detailed Comments

The paper "Word Embeddings as Tuples of Feature Probabilities" discusses the idea normalizing word2vec-like word embeddings vocabulary-wise across each individual dimension and fuzzy-set-theoretical interpretation of such normalization as well as some empirical evaluations.

I would say that the proposed interpretation and framework for doing "arithmetic" operations on vectors is rather creative, and although static word embeddings are recently being outperformed by contextualized approaches, I believe there is still interest in such work.

The paper is overall well written, although feels a bit rushed: there are, for instance, incomplete sentences like in line 228.

More substantial issues to improve:

Authors claim the the column normalization turns each dimension of embeddings into "feature probabilities" - I could not fully follow why it is so.
Authors claim that dimensions of word2vec models are not interpretable, however Levi at all showed that those are factorized PPMI of word concurrences.
In any case, the proposed method being build on top of word2vec/ GloVe can not be more interpretable than the original!

Word examples in included tables seem to be randomly sampled, I assume to show that they are not hand-picked in favor of author's implementation.
This, however, makes for pretty weird examples. First of all, why not use same pairs of words for both models, (table 2), and in general sample from more frequent "normal" words,
if not hand-pick coupe of examples?

"Google" analogy is not well balanced and does not cover many linguistic categories. Consider adding BATS to evaluation.

REVIEWER #2

Reviewer's Scores

Appropriateness (1-5)	5
Clarity (1-5)	4
Originality / Innovativeness (1-5)	4
Soundness / Correctness (1-5)	3
Meaningful Comparison (1-5)	3
Thoroughness (1-5)	4
Recommendation (1-5)	4
Reviewer Confidence (1-5)	4

Detailed Comments

This paper proposes normalization of word vectors trained using the skip-gram model across the vocabulary and reinterprets the resulting vectors as a collection of features. Authors use fuzzy set theoretic operations over these representations to reinterpret the notions of difference, union, intersection of two words. Moreover, notions of inclusion, entropy and KL-divergence are also defined. All the above notions are backed by qualitative analyses, whereas quantitative evaluations are provided for the similarity and analogy tasks.
The premise of the paper is very interesting and provides a fresh treatment to a widely studied topic. The evaluation is exhaustive and thorough. Some of the qualitative experiments, however, are not immediately convincing. For Table 1, some unions are not interpretable. Even though in terms of qualitative results the paper doesn't seem to improve the performance by a lot and many of the evaluations are qualitative, the paper provides an interesting view of interpreting the word vectors backed by sufficient analyses and might be of interest to members of the community.

Some typos:
Line 298; subscript should be symbol for union not intersection
In conclusion; "We and performed" -> "We performed"
5.1.1 "ration" -> "ratio"

REVIEWER #3

Reviewer's Scores

Appropriateness (1-5)	5
Clarity (1-5)	3
Originality / Innovativeness (1-5)	3
Soundness / Correctness (1-5)	3
Meaningful Comparison (1-5)	3
Thoroughness (1-5)	3
Recommendation (1-5)	4
Reviewer Confidence (1-5)	5

Detailed Comments

This paper presents a way to reinterpret dimensions in the non-contextualized word embeddings (i.e., Word2Vec) as the probability for the words to express the features corresponding to the dimensions. Based on a simple column-wise normalization of the vectors, the paper employ the operations in fuzzy sets to capture the typical operations in the word vectors, including the asymmetric similarity, analogy, function word detection, and compositionality. The paper provides intuitive examples about the use of the fuzzy operations and shows promising results on the intrinsic evaluations of the embeddings.

This paper is interesting as it introduces a different way to view the non-contextualized word embeddings. Although the current works seem to suggest contextualized embeddings as the standards for downstream applications, this paper still provides some insights that might be helpful for the future work on word embeddings.

As the authors noticed, the proposed method seems to work better for lower-dimension embeddings. The paper explains this a bit, but it is interesting to provide some examples to provide more intuition about this. Also, the paper only shows the performance of the methods with the embeddings up to 200 dimensions. As 300 dimensions seem to be the popular version being used, a natural question is why the performance with 300 dimensions is not shown? Will the performance be worse than orignal word2vec in this case?

Tables 1, 2, and 3 would be more helpful if some additional explanations are provided. For instance, what is the connection between the words mentioned in the captions and the words in the tables?

==========
REVIEWER #4
==========

Reviewer's Scores

Appropriateness (1-5): 4
Clarity (1-5): 3
Originality / Innovativeness (1-5): 3
Soundness / Correctness (1-5): 4
Meaningful Comparison (1-5): 4
Thoroughness (1-5): 3
Recommendation (1-5): 3
Reviewer Confidence (1-5): 3

Detailed Comments

This paper takes the word embeddings as a collection of features and investigates using fuzzy-sets operations (like intersection, union, complement, etc.) to measure the relationships between different words. Experimental results on many tasks and datasets (including MSR Word Relatedness, GoogleNews, etc) shows that using the fuzzy-sets viewing of word embeddings can improve the performance of using word embeddings (especially for small dims such as 50, 100).

However, one important question is not enough discussed, i.e., what is the impact of word embedding size on the performance. I found that if we keep increasing the word embedding size, the performances (baseline vs fuzzy-sets) are becoming closer to each other. Because in many real-world applications, the embedding siz

e is usually not that small (like 50), we should better discuss what is the pote
ntial application of our fuzzy-sets view.

Besides, the analysis of this paper should be strengthened, those scores and sho
wcases can not help me easily understand why does the fuzzy-sets view help. Mayb
e a visualization of different word-embedding views can make the paper more conv
incing.
