SI 630 - Winter 2020
Reading Response 3

Yirui Gao

February 4, 2020

Overall summary

This paper, *Evaluation methods for unsupervised word embeddings*, propose several comparative studies on evaluating multiple existing word embedding methods, like **CBOW**, **GloVe**, **C&W**, etc. Different from traditional evaluation methods that rely on absolute evaluations, which mostly depend on aggregate evaluations, the authors use comparative evaluation methods by taking some query inventories and put them inside several embedding blackboxes, then ask for human judgement. This method can directly reflect the human references and is relative compared to absolute methods. It shows the results for intrinsic tasks that word frequency determines the performance of different embeddings. Comparing the comparative results with the absolute results, the authors find the results are similar, but absolute metrics are less principled and insightful.

The authors also conduct a novel coherence task to measure the coherence of neighborhoods in different embedding space, and they find that better intrinsic performances of the embedding does not lead to better extrinsic performances that apply those embedding methods into real NLP tasks like noun phrase chunking and sentiment classification. In the discussion part, the authors point out potential issues when they use cosine similarity as the distance metric resulted from the word frequency information in the embedding space.

Main contributions

The main contributions for this paper are:

- The authors introduce a novel evaluation framework by comparatively comparing the difference of different embeddings and present in human judgement that makes the results more convincing and relevant.
- They also perform extrinsic studies that apply comparative studies on different specific NLP tasks to show the comparisons between embeddings.
- They perform studies and discover that the different performance of embeddings are highly related to the word frequency in the embedding space.

Fill Out Three of the Following Responses (Your Choice)

3.1 What would I have done differently?

For me, since the authors point out in the paper that the cosine similarity might be affected due to the difference between word frequency, an alternative comparative study might be conducted using a differnt similarity metric like word mover's distance proposed in 2015 as well. Another thing is that for the chunking task, we can see in the paper that all the embeddings achieve quite high performances with the F1 score ranging from 0.9418 to 0.9453, which is a very tiny difference. I would consider finding a way to enlarge the difference of different embeddings since a very close result may not be too convincing to indicate which one is better.

3.2 What did I learn?

For me, the biggest finding in this paper, is that the word frequency can highly influence the performances of those embeddings, which is due to the difference of original training corpuses. What's more, from this paper, I have a more general about how different embeddings work and how they perform in some specific tasks, which helps me with my embedding choice in future NLP studies. The comparative evaluation procedure using query inventories is also a good point to learn in the future study if an evaluation needs it since it can provide more direct and convincing results than using a simple aggregation to calculate something.

3.3 What was I confused about?

This paper involves studies on the six common embedding methods, but for me, most of them are unfamiliar to me, so a deeper understanding on what they are and how they work is needed. Also, one of my confusions is how to ensure the credibility of selected users from Amazon Mechanical Turk since I have seen several studies using this kind of "volunteer-involved" way to conduct an experiment, I'm wondering how to ensure their recognition and incentive can fulfill the researchers' requirement.

References

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. pages 298–307.