Review of Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank (Socher et al., 2013)

Teresa Martín Soeder

This paper introduces the Stanford Sentiment Treebank with sentiment analysis annotations and the Recursive Neural Tensor Network (RNTN), designed to predict the compositional semantic effects present in the corpus. Furthermore, they provide evidence that shows that the RNTN has greater accuracy than a standard Recursive Neural Network (RNN) and a Matrix-Vector RNN (MV-RNN) when classifying the sentences in the corpus into a fine-grained classification of 5 labels, into a binary classification, in subsets consisting of positive or negative sentences and their negation, and in a subset for the analysis of the conjunction *but*.

The corpus, which consists of phrases extracted from movie reviews, was parsed with the Stanford Parsed and annotated by 3 turkers on a 25-point scale. Then, based on the annotators' use of the labels, these 25 points were reduced to 5 labels for the experiments. This is partially reminiscent of more recent studies like that of Pavlick and Kwiatkowski (2019), but it does not hold the test of time since the number of annotators is minimal and there is no consideration of possible inherent disagreement, a phenomenon very common in sentiment analysis (Uma et al., 2021), or even an inter-annotator agreement score. As to the mapping from 25 to 5 labels, it seems plausible given the graphics in Figure 2, but, since for full sentences there appears to be a more even distribution along the whole scale, it would have been good to see a discussion of the limitations of this set of labels for full sentences and longer texts. Furthermore, a graphic clearly showing the mapping from 25 to 5 labels and a clarification on how many labels the available corpus has, would have been helpful.

Regarding the RNTN, the detailed explanation about how the computations are made and the comparisons to its closed relatives RNN and MV-RNN is really good and helps in understanding the value of the RNTN. Furthermore, the consideration of cognitive and linguistic plausibility by implementing an efficient method for semantic compositionality is very interesting and one wonders what Natural Language Explanation Techniques would reflect if applied to each model. In regards to the experiments, they seemed to support the combination of linguistic and cognitive plausibility characteristic of the RNTN, but given that they only provide accuracy as an evaluation metric and that we ignore the exact corpus distribution among the 5

labels, it appears that more evidence is needed to certify such improvement. Lastly, it would have been good to provide a discussion of the limits of the model especially since semantic compositionality has long been known to have problems with, for example, idioms like *At the end, the movie went down in flames*.

In conclusion, it is possible to say that this study holds to the standards of the time it was published, but it is short on information on both of the corpus and the evaluation metrics.

References

Pavlick, E. and T. Kwiatkowski (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics 7*, 677–694.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Uma, A. N., T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research 72*, 1385–1470.