

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. Specifically, the corpus consists of phrases with their fully labeled parse trees which enables their analysis in terms of their semantic compositionality and allows a more fine-grained study of sentiment analysis. The neural network model is distinguished from standard Recursive Neural Networks (RNNs) in that the vectors representing each node in a parse tree are explicitly combined. From the Matrix-Vector RNN (MV-RNN) the difference lies in the composition function: for the MV-RNN we have one composition function at each interaction and for the RNTN there is one for the whole dataset. Experimental results showed that the corpus increases the accuracy of the previous models and that the RNTN outperforms the other models in the whole dataset, in subsets consisting of positive or negative sentences and their negation, and in a subset for the analysis of the conjunction *but*.

Regarding the corpus, its creation is explained with enough detail but the details of the finalized corpus and whether is distinct from the one used for the experiments it is not so clear. Although it is true that the authors refer to additional materials, which are not published with the paper and were not found after a quick search, some important information could have been provided within the article. For example, the mapping from the 25 labels to the 5 classes is not completely clear. In Figure 2 the mapping is roughly shown, but there are other two categories (Very negative and Very positive). Furthermore, it is not clarified whether the corpus made available has the 25 labels or just the 5 categories.

Two other relevant features about the corpus that could have been provided are a measure of inter-annotator agreement and a distribution of the categories. It is true that with so many labels and just 3 annotators an inter-annotator agreement score could be problematic, but it is difficult to assess the quality of the annotations without an evaluation of how the annotators influenced the annotations. Also, given that unbalanced datasets in terms of the categories can influence modeling, it would have been good to see a simple table that shows the frequency of each of them.

As to 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, given the simple logic and lack of linguistic information behind bag-of-words corpora and models, it is not completely surprising that accuracy improves with a corpus informed with semantic compositionality and with a model that explicitly considers it. Although the experiments presented clearly show this improvement, a more thorough discussion is needed.

The experiments proved that the RNTN performs better in short texts, but it would have been good to discuss the generalizability of such a model. Specifically when having long documents and where compositional semantics fails to properly represent an utterance. In the first case, I do not expect problems for the RNTN, it likely performs better than bag-of-words models, but in the second case, expectations are not so clear. Would relying on the annotators' labels would be enough to correctly predict the sentiment of a phrase like *At the end, the movie went down in flames?*

References

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.