Representations of Meaning, 7.5 HEC Distributional Semantics for Textual Entailment

Simon Dobnik

simon.dobnik@gu.se

June 20, 2016

Investigate to what degree distributional semantic similarity between premises and conclusion can be used as measure for textual entailment.

You will need some code to estimate semantic similarity between sentences and paragraphs.

Gensim has `models.doc2vec` which you can find here. The standard `models.word2vec` can be found here which also involves methods to deal with phrases (which may or may not be useful).

Describe in your report what this code does. The code implements the following papers which would be worth looking at.

Le, Quoc, and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1188–1196

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. ArXiv preprint arXiv:1301.3781

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119

Another implementation which is based on [1] can be found here.

When you are ready to estimate similarities between sentences/paragraph choose a couple of examples of each category (yes, no, don't know) from the RTE task (5 at least) and estimate the similarity between the premises and the conclusion. Discuss what this method tells us about textual entailment.

The report and the code can be best presented as an iPython notebook.