Presentation by the LTC Sponsor

SAMSUNG

Filip Graliński, Krzysztof Wilkosz, Mikołaj Wypych

Samsung Poland R&D Institute
Pl. Europejski 1, Warszawa, Poland
f.gralinski@partner.samsung.com, {k.wilkosz,m.wypych}@samsung.com

Word Embeddings: From Parlour Tricks through Linguistic Insights to Practical Applications

Despite being known earlier, word embeddings really caught on when the word2vec tool (skip-gram or CBOW model with negative sampling) was made available in 2013. That was the moment when everybody could (pre-)train dense word representations on a very large corpus using readily available hardware. The original word2vec forked and morphed into a number of tools, such as GloVe, Gensim, bivec, multivec, fastText (the last one could be viewed as a combination of word2vec and another simple, yet powerful tool - VowpalWabbit). A number of theoretical advances have led to further improvements: subword or character embeddings, phrase embeddings, handling polysemy, representing words with probability distributions rather than simple vectors.

It was noticed, by the authors of word2vec themselves, that word embeddings are capable of resolving word analogy tasks (A is to B as C is to what?) just by adding and subtracting vectors. The fact that a word2vec model is able to guess that "queen" is in the same relation to "woman" as "king" to "man" is not just as a curiosity and it can bring real linguistic insights. We will show this on the example of temporal word analogies induced from a large diachronic corpus.

Word embeddings are indispensable for historical linguists and NLP practitioners alike. In this talk we will argue for their usefulness especially in the context of multilinguality and low-resource domains and how practical NLU (Natural Language Understanding) applications can be developed with them rapidly.

Balancing train sets with categorical features for improved classification

Imbalanced datasets pose a significant challenge to training of classifiers. For independent, continuous features there exists a range of methods (e.g. majority class undersampling, SMOTE, ADASYN) but there seem to be no known methods to reduce imbalance for datasets with inherently categorical and potentially dependent features. Such a heuristical method applicable for large imbalanced datasets consisting of hundreds of (potentially interdependent) features will be presented. This dataset balancing approach proved successful in reducing classification error for Deep Neural Network classifiers for sequence labelling task but their applicability is not limited to only this problem.