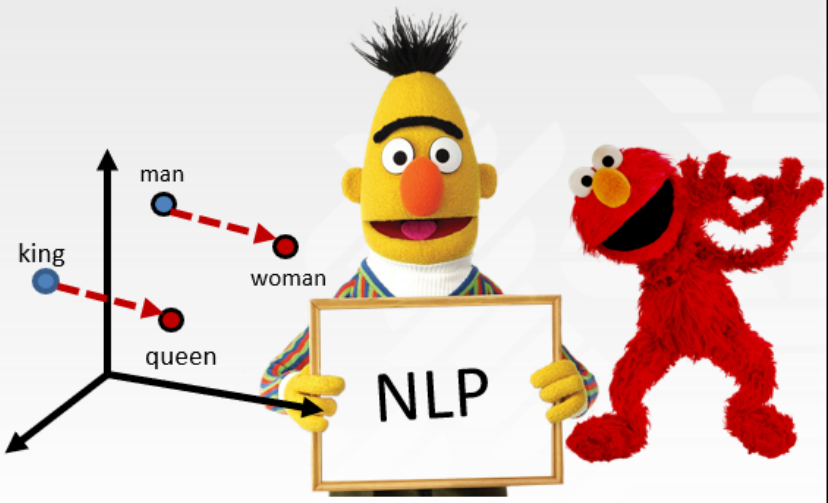
Learning Vector Representations of Words, Phrases, and Documents

Background

Vector representations of words, better known as word embeddings, are the groundwork for most natural language processing tasks. The goal is to embed words in a vector space such that words with similar semantical meaning end up close to each other in this space. Word2Vec is one of the most popular strategies to learn word embeddings using shallow neural network (Mikolov et al. 2013). However, deep learning approaches for contextualized word embeddings such as BERT (Devlin et al. 2018) or EMLO (Peters et al. 2018) are gaining more popularity as well. We would like to explore the ability of these methods to capture the semantical meaning of not just words but long-text documents such as Wikipedia articles.



Goal

- Analyze the strengths and weaknesses of individual embedding methods for the task of finding semantical related phrases and documents.

Tasks

- Review state-of-the-art approaches to learn vector representations for words and how they can be extended to derive representations for phrases and documents.
- Evaluate different approaches on their ability to find related Wikipedia articles and coreferential phrases.

Malte Ostendorff
malte.ostendorff@dfki.de

Anastasia Zhukova
zhukova@uni-wuppertal.de