Abstract

Empirical analysis of word vectors for ambiguous words

Máté Gedeon

supervised by
András Kornai, D.Sc.

Natural Language Processing (NLP) has gained significant attention in recent years with the appearance of state-of-the-art models and techniques. In the thesis, we focus on word embeddings, which aim to represent a word using a vector capturing as much semantic and morphological information as possible. We explore the differences between static embeddings, like Word2Vec (Mikolov et al., 2013) and dynamic ones, as well as techniques for extracting word embeddings from Transformer-based models, like huBERT (Nemeskey, 2021). The Transformer architecture is also presented along with key concepts like tokenization and attention.

We put the embeddings extracted from huBERT to the test on a set of analogy questions by Makrai (2015). The goal was to surpass the performance of a Word2Vec embedding used as a benchmark, which achieved an accuracy of 42.4%. With the best model performing at 43.2%, the attempt proved to be successful. We also investigated how Principal Component Analysis (PCA) affects the performance on these analogy tasks and found that as the dimensions reduced, the performance improved initially, followed by a gradual decrease and a rapid decline at the end. This suggests that PCA is a helpful tool to improve performance by decreasing dimensionality, but there is a limit to by how much we can reduce the dimensionality efficiently.

Finally, we focused on ambiguous words and their treatment in NLP. We obtained separate embeddings for ten Hungarian ambiguous words and their two senses. We explored whether the embedding generated for the word disregarding the sense can be produced as a linear combination of the different embeddings for the different senses. The idea originated from the Linearity Assertion presented by Arora et al. (2016). We found that the assertion didn't hold in our case, likely due to the use of a Transformer-based model. This assertion could be investigated further with more ambiguous words and alternative embedding techniques.

2023