

Word (Actually Token) Embeddings

- Idea: Map each token to a dense vector in continuous space.
- Result: 300–2000 dimensions instead of 1–2M.
 - The dimensions have no clear interpretation.
- The “embedding” is the mapping.
 - Technically, the first layer of NNs for NLP is the matrix that maps 1-hot input to the first layer.
- Embeddings are trained for each particular task.
 - Sentence classification (sentiment analysis, etc.)
 - Neural language modelling.
 - The famous word2vec (Mikolov et al., 2013):
 - CBOW: Predict the word from its four neighbours.
 - Skip-gram: Predict likely neighbours given the word.
 - End-to-end neural MT.