WIKIPEDIA
The Free Encyclopedia

Word embedding

In natural language processing (NLP), a **word embedding** is a representation of a word. The embedding is used in text analysis. Typically, the representation is a real-valued vector that encodes the meaning of the word in such a way that the words that are closer in the vector space are expected to be similar in meaning.[1] Word embeddings can be obtained using language modeling and feature learning techniques, where words or phrases from the vocabulary are mapped to vectors of real numbers.

Methods to generate this mapping include neural networks,[2] dimensionality reduction on the word co-occurrence matrix,[3][4][5] probabilistic models,[6] explainable knowledge base method,[7] and explicit representation in terms of the context in which words appear.[8]

Word and phrase embeddings, when used as the underlying input representation, have been shown to boost the performance in NLP tasks such as syntactic parsing[9] and sentiment analysis.[10]

Development and history of the approach

In distributional semantics, a quantitative methodological approach to understanding meaning in observed language, word embeddings or semantic feature space models have been used as a knowledge representation for some time.[11] Such models aim to quantify and categorize semantic similarities between linguistic items based on their distributional properties in large samples of language data. The underlying idea that "a word is characterized by the company it keeps" was proposed in a 1957 article by John Rupert Firth,[12] but also has roots in the contemporaneous work on search systems[13] and in cognitive psychology.[14]

The notion of a semantic space with lexical items (words or multi-word terms) represented as vectors or embeddings is based on the computational challenges of capturing distributional characteristics and using them for practical application to measure similarity between words, phrases, or entire documents. The first generation of semantic space models is the vector space model for information retrieval.[15][16][17] Such vector space models for words and their distributional data implemented in their simplest form results in a very sparse vector space of high dimensionality (cf. curse of dimensionality). Reducing the number of dimensions using linear algebraic methods such as singular value decomposition then led to the introduction of latent semantic analysis in the late 1980s and the random indexing approach for collecting word cooccurrence contexts.[18][19][20][21] In 2000, Bengio et al. provided in a series of papers titled "Neural probabilistic language models" to reduce the high dimensionality of word representations in contexts by "learning a distributed representation for words".[22][23][24]

A study published in NeurIPS (NIPS) 2002 introduced the use of both word and document embeddings applying the method of kernel CCA to bilingual (and multi-lingual) corpora, also providing an early example of self-supervised learning of word embeddings [25]

Word embeddings come in two different styles, one in which words are expressed as vectors of co-occurring words, and another in which words are expressed as vectors of linguistic contexts in which the words occur; these different styles are studied in Lavelli et al., 2004.[26] Roweis and Saul