

Word embedding

From Wikipedia, the free encyclopedia

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Conceptually it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension.

Methods to generate this mapping include neural networks,^[1] dimensionality reduction on the word co-occurrence matrix,^{[2][3][4]} probabilistic models,^[5] and explicit representation in terms of the context in which words appear.^[6]

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^[7] and sentiment analysis.^[8]

Contents

- 1 Development of technique
- 2 For biological sequences: BioVectors
- 3 Thought vectors
- 4 Software
- 5 See also
- 6 References

Development of technique

In linguistics word embeddings were discussed in the research area of distributional semantics. It aims 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 popularized by Firth.^[9]

The word embedding technique began development in 2000. Bengio et al. provided in a series of papers the "Neural probabilistic language models" to reduce the high dimensionality of words representations in contexts by "learning a distributed representation for words". (Bengio et al, 2003).^[10] Roweis and Saul published in *Science* how to use "locally linear embedding" (LLE) to discover representations of high dimensional data structure.^[11] The area developed gradually and really took off after 2010, partly because important advances had been made since then on the quality of vectors and the training speed of the model.

There are many branches and many research groups working on word embeddings. In 2013, a team at Google led by Tomas Mikolov created word2vec, a word embedding toolkit which can train vector space models faster than the previous approaches.^[12] Most new word embedding techniques rely on a neural network architecture instead of more traditional n-gram models and unsupervised learning.^[13]

For biological sequences: BioVectors

Word embeddings for n-grams in biological sequences (e.g. DNA, RNA, and Proteins) for bioinformatics applications have been proposed by Asgari and Mofrad.^[14] Named bio-vectors (BioVec) to refer to biological sequences in general with protein-vectors (ProtVec) for proteins (amino-acid sequences) and gene-vectors

(GeneVec) for gene sequences, this representation can be widely used in applications of deep learning in proteomics and genomics. The results presented by^[14] suggest that BioVectors can characterize biological sequences in terms of biochemical and biophysical interpretations of the underlying patterns.

Thought vectors

Thought vectors are an extension of word embeddings to entire sentences or even documents. Some researchers hope that these can improve the quality of machine translation.^{[15] [16]}



Software


Software for training and using word embeddings includes Tomas Mikolov's Word2vec, Stanford University's GloVe,^[17] Gensim^[18] and Deeplearning4j. Principal Component Analysis (PCA) and T-Distributed Stochastic Neighbour Embedding (t-SNE) are both used to reduce the dimensionality of word vector spaces and visualize word embeddings and clusters.^[19]

See also

- Latent semantic analysis
- Brown clustering

References

- Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg; Dean, Jeffrey (2013). "Distributed Representations of Words and Phrases and their Compositionality". arXiv:1310.4546 (<https://arxiv.org/abs/1310.4546>)  [cs.CL (<https://arxiv.org/archive/cs.CL>)].
- Lebret, Rémi; Collobert, Ronan (2013). "Word Emdeddings through Hellinger PCA". *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. **2014**. arXiv:1312.5542 (<https://arxiv.org/abs/1312.5542>) .
- Levy, Omer; Goldberg, Yoav (2014). *Neural Word Embedding as Implicit Matrix Factorization* (<http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.pdf>) (PDF). NIPS.
- Li, Yitan; Xu, Linli (2015). *Word Embedding Revisited: A New Representation Learning and Explicit Matrix Factorization Perspective* (<http://ijcai.org/papers15/Papers/IJCAI15-513.pdf>) (PDF). Int'l J. Conf. on Artificial Intelligence (IJCAI).
- Globerson, Amir (2007). "Euclidean Embedding of Co-occurrence Data" (<http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/34951.pdf>) (PDF). *Journal of Machine learning research*.
- Levy, Omer; Goldberg, Yoav (2014). *Linguistic Regularities in Sparse and Explicit Word Representations* (<https://levyomer.files.wordpress.com/2014/04/linguistic-regularities-in-sparse-and-explicit-word-representations-conll-2014.pdf>) (PDF). CoNLL. pp. 171–180.
- Socher, Richard; Bauer, John; Manning, Christopher; Ng, Andrew (2013). *Parsing with compositional vector grammars* (http://www.socher.org/uploads/Main/SocherBauerManningNg_ACL2013.pdf) (PDF). Proc. ACL Conf.
- Socher, Richard; Perelygin, Alex; Wu, Jean; Chuang, Jason; Manning, Chris; Ng, Andrew; Potts, Chris (2013). *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank* (http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf) (PDF). EMNLP.
- Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955". *Studies in Linguistic Analysis*. Oxford: Philological Society: 1–32. Reprinted in F.R. Palmer, ed. (1968). *Selected Papers of J.R. Firth 1952-1959*. London: Longman.
- "A Neural Probabilistic Language Model" (http://link.springer.com/chapter/10.1007/3-540-33486-6_6). doi:10.1007/3-540-33486-6_6 (https://doi.org/10.1007%2F3-540-33486-6_6).

11. Roweis, Sam T.; Saul, Lawrence K. (2000). "Nonlinear Dimensionality Reduction by Locally Linear Embedding" (<http://science.sciencemag.org/content/290/5500/2323>). *Science*. **290** (5500): 2323. Bibcode:2000Sci...290.2323R (<http://adsabs.harvard.edu/abs/2000Sci...290.2323R>). PMID 11125150 (<http://www.ncbi.nlm.nih.gov/pubmed/11125150>). doi:10.1126/science.290.5500.2323 (<https://doi.org/10.1126/2Fscience.290.5500.2323>).
12. word2vec (<https://code.google.com/archive/p/word2vec/>)
13. "A Scalable Hierarchical Distributed Language Model" (<http://papers.nips.cc/paper/3583-a-scalable-hierarchical-distributed-language-model>).
14. Asgari, Ehsaneddin; Mofrad, Mohammad R.K. (2015). "Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics" (<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141287>). *PloS one*. **10** (11): e0141287. Bibcode:2015PLoSO..1041287A (<http://adsabs.harvard.edu/abs/2015PLoSO..1041287A>). doi:10.1371/journal.pone.0141287 (<https://doi.org/10.1371%2Fjournal.pone.0141287>).
15. Kiros, Ryan; Zhu, Yukun; Salakhutdinov, Ruslan; Zemel, Richard S.; Torralba, Antonio; Urtasun, Raquel; Fidler, Sanja (2015). "skip-thought vectors". arXiv:1506.06726 (<https://arxiv.org/abs/1506.06726>)  [cs.CL (<https://arxiv.org/archive/cs.CL>)].
16. "thoughtvectors" (<http://deeplearning4j.org/thoughtvectors>).
17. "GloVe" (<http://nlp.stanford.edu/projects/glove/>).
18. "Gensim" (<http://radimrehurek.com/gensim/>).
19. Ghassemi, Mohammad; Mark, Roger; Nemati, Shamim (2015). "A Visualization of Evolving Clinical Sentiment Using Vector Representations of Clinical Notes" (<http://www.cinc.org/archives/2015/pdf/0629.pdf>) (PDF). *Computing in Cardiology*.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Word_embedding&oldid=790665823"

-
- This page was last edited on 15 July 2017, at 07:26.
 - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.