Funded by the European Union

Word embeddings

Vitor Rocio

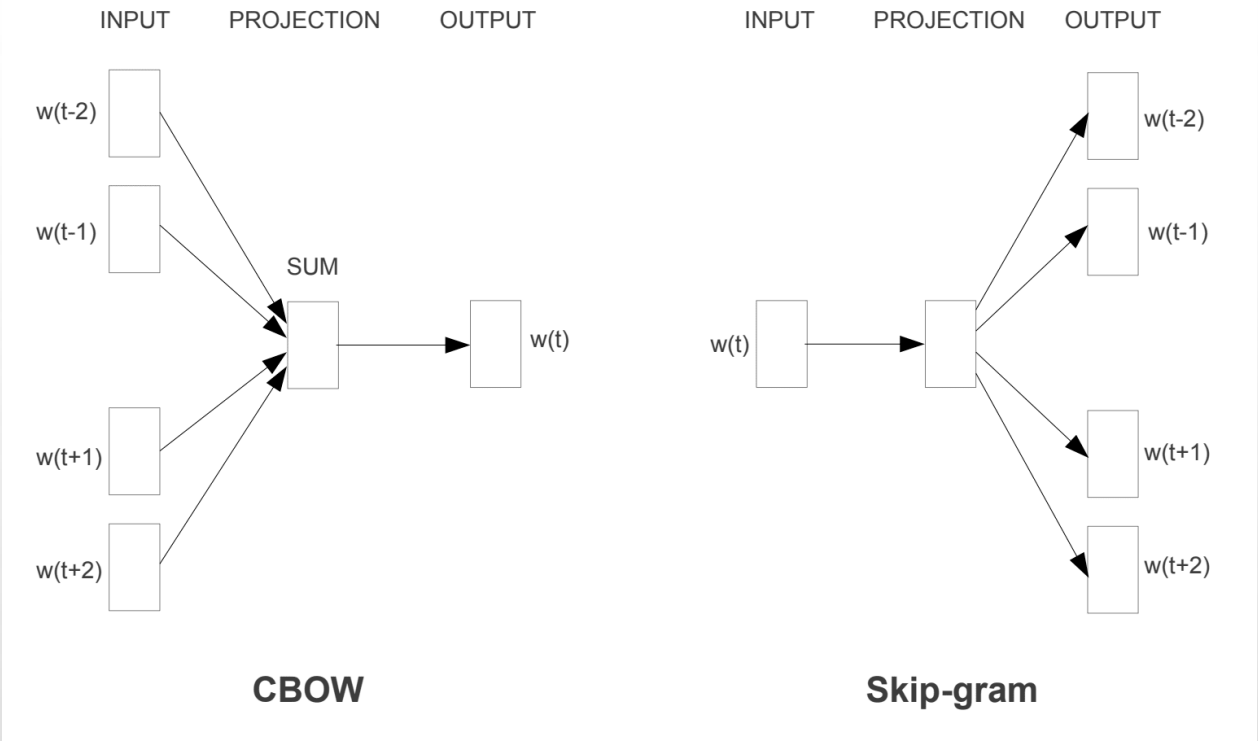
(vitor.rocio@uab.pt)

What is Word embedding?

Source: https://www.enjoyalgorithms.com/blog/word-vector-encoding-in-nlp

- Text representations where words are described by numeric vectors: proximity in vector space corresponds to proximity in meaning
- “A word is caracterized by the company it keeps” (Firth, 1957)
- Key-feature: vector low dimensionality
- Instead of bag-of-words, where words are undifferentiated, word embeddings capture meaning.

Word2Vec



Source: Tomas Mikolov, 2013

Word2Vec algorithm

```
sentences = ...

model = Word2Vec(sentences, …)
```

- Parameters:
  - size – no. of dimensions
  - window – no. of words to the left and to the right
  - min_count – minimum frequency to consider
  - workers – no. of threads used for training
  - sg – CBOW (0) or skip gram (1)

Applications

- Topic extraction and summarization
- Recommender systems
- Trend identification, sentiment analysis
- Content analysis
- Document classification
- Named entity recognition (NER)
- Identification of similiarities and analogies among words

References

- Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John (ed.). "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993
- Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781
- Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeff (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. arXiv:1310.4546
- Word Vector Encoding in NLP (enjoyalgorithms.com) *https://www.enjoyalgorithms.com/blog/word-vector-encoding-in-nlp*

Funded by
the European Union

Thank you for your attention