Information Retrieval & Text Mining

Word Embeddings an Introduction

Tutorial by Farooq Zaman
Information Technology University

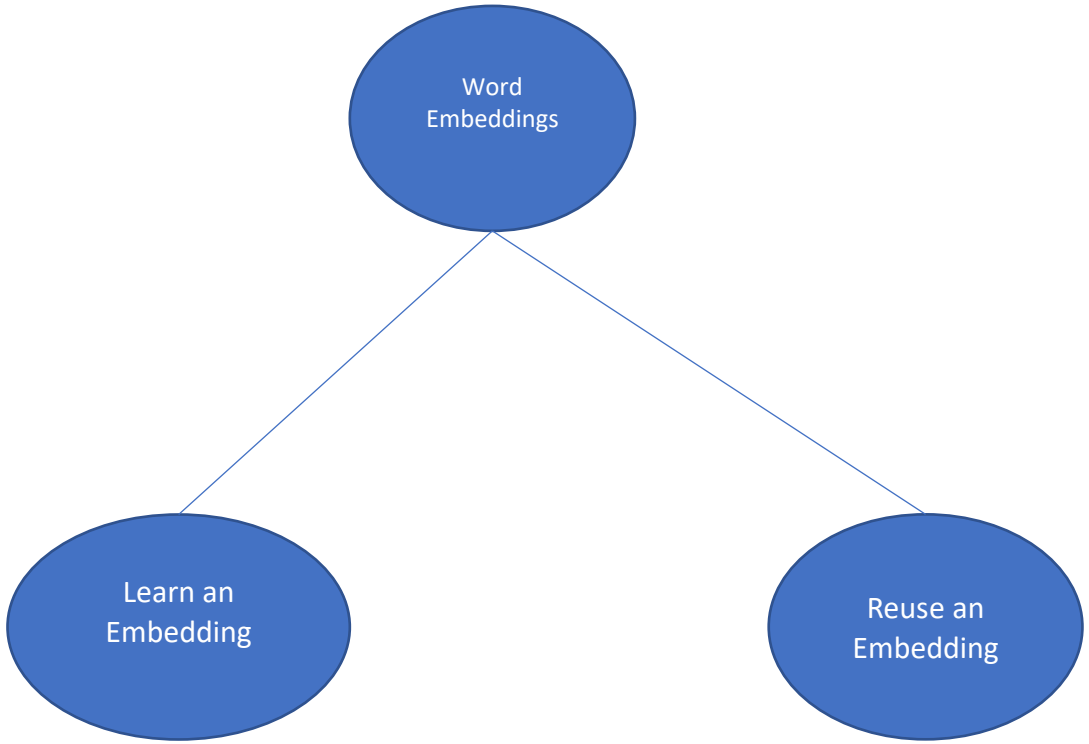
Embeddings

- What are word Embeddings:
  - Type of word representation that allows words with similar meaning to have a similar representation
  - A distributed representation for text
  - The impressive performance of deep learning methods on challenging natural language processing problems is due to Embeddings

Pre-trained Embeddings

- Word2Vec (Tomas Mikolov, et al Google 2013)
  - Word2Vec is a statistical method for efficiently learning a standalone word embedding from a text corpus

- GloVe (Pennington, et al. at Stanford 2014)
  - Global Vectors for Word Representation (GloVe)
  - Is an extension to the word2vec method
  - Global log-bilinear regression model
  - Perform well on word analogy, word similarity, and named entity recognitior tasks


Word Embeddings
Learn an Embedding
Reuse an Embedding

Learn Embedding

- Require a large amount of text data (millions or billions of words)
- Learn it Standalone
  - Train and learn the embedding
  - Saved and used as a part of another model for your task later
  - Good if you require to use the same embedding in multiple models
- Learn Jointly
  - Learned as part of a large task-specific model
  - Good if you are interested in using the embedding on one task only

Reuse Embedding

- Pre-trained word embeddings
- Word2vec and GloVe word embeddings are available
- Can be static only seed the model with
- Can be updated with the new data

Demo session

- Follow the shared notebook

Further Readings

1. How to Develop Word Embeddings in Python with Gensim
2. How to Use Word Embedding Layers for Deep Learning with Keras
3. How to Develop a Deep CNN for Sentiment Analysis (Text Classification)
4. Word embedding on Wikipedia
5. Word2vec on Wikipedia
6. GloVe on Wikipedia
7. An overview of word embeddings and their connection to distributional semantic models, 2016.
8. Deep Learning, NLP, and Representations, 2014
9. Distributional structure, 1956.
10. A Neural Probabilistic Language Model, 2003.
11. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning, 2008.
12. Continuous space language models, 2007.
13. Efficient Estimation of Word Representations in Vector Space, 2013
14. Distributed Representations of Words and Phrases and their Compositionality, 2013.
15. GloVe: Global Vectors for Word Representation, 2014.