VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Algorithms for Big Data

Basics of Text Processing using DNN

Jan Platoš

November 2, 2020

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Representation of the meaning of a words

Representation of the meaning of a words

- The **meaning** means:
  - the idea that is represented by a word, phrase, etc.
  - the idea that a person wants to express by using words, signs, etc.
  - the idea that is expressed in a work of writing, art, etc.
- A WordNet is a great resource of meaning:
  - A complex network of words made by human.
  - A list of synonyms, hypernyms (generalization), antonyms, etc.
  - A word category with dictionary-like description of a meaning.
  - A new meaning are missing in a database.
  - Some meaning and synonyms are valid only in some contexts.

Representation of the meaning of a word

- The standard representation is called **one-hot** vector.

$$\begin{aligned} \textit{motel} &= [00000000100] \\ \textit{hotel} &= [00000100000] \end{aligned}$$

- Vector dimension = number of word in a corpus
- Vectors are orthogonal $\textit{motel} \cdot \textit{hotel} = 0$
- Similarity cannot be defined on one-hot vector representation.
- WordNet may be used to extract synonyms for each word that will be used as similarity function, but ist too complicated approach.

Representation of the meaning of a word

A word’s meaning is given by the words that frequently appear close-by

Representation of the meaning of a word

A word's meaning is given by the words that frequently appear close-by

Example:

...reasonable and to prevent the	network	trips from swamping out the execution...
...distance between nodes;	network	traffic or bandwidth constraints; ...
...beyond your control (i.e.	network	outage, hardware failure) or the latency ...
...experience was a temporarily-high	network	load which caused a timeout...
...is removed (i.e. temporary	network	disconnection resolved) then ...
...see their involvement with the	network	and its digital properties expand ...
...but cant get mobile	network	connection to work. Basically ...

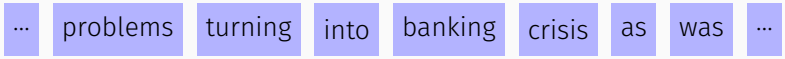
Representation of the meaning of a word - Word2Vec

Word2vec is a framework for learning word vectors.

- We have a large corpus of text.
- Every word in a fixed vocabulary is represented by a vector.
- Go through each position *t* in the text, which has a center word *c* and context words *o*.
- Use the similarity of the word vectors for *c* and *o* to calculate the probability of *o* given *c*.
- Keep adjusting the word vectors to maximize the probability.

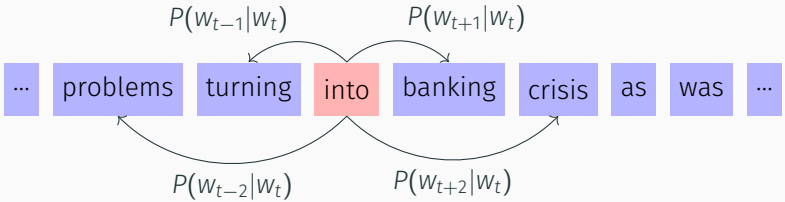
Representation of the meaning of a word - Word2Vec

Word2vec principle.



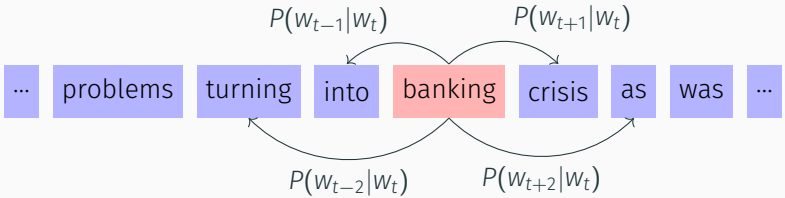
Representation of the meaning of a word - Word2Vec

Word2vec principle.



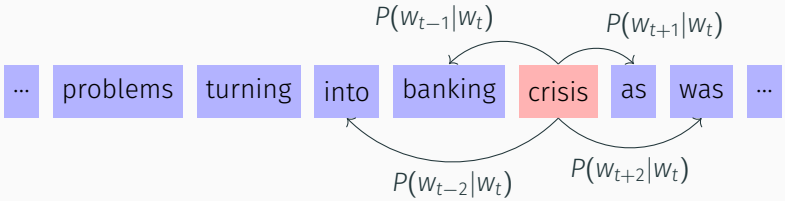
Representation of the meaning of a word - Word2Vec

Word2vec principle.



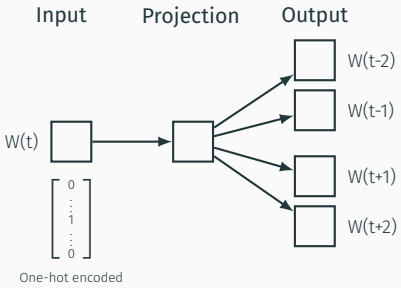
Representation of the meaning of a word - Word2Vec

Word2vec principle.

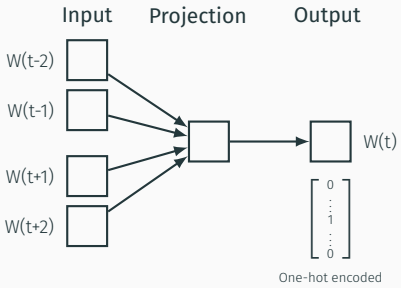


Representation of the meaning of a word - Word2Vec Variants

Skip-Gram (SG) where the contexts predicts words given the center word independently on position.



Continuous Bag of Words (CBOW) where the center word is predicted from context words.



Representation of the meaning of a word - GloVe

GloVe: Global Vectors for Word Representation

- Combines both Skip-gram and C-Bow methods
- Fast training, scalable to huge corpora but works even on small ones.

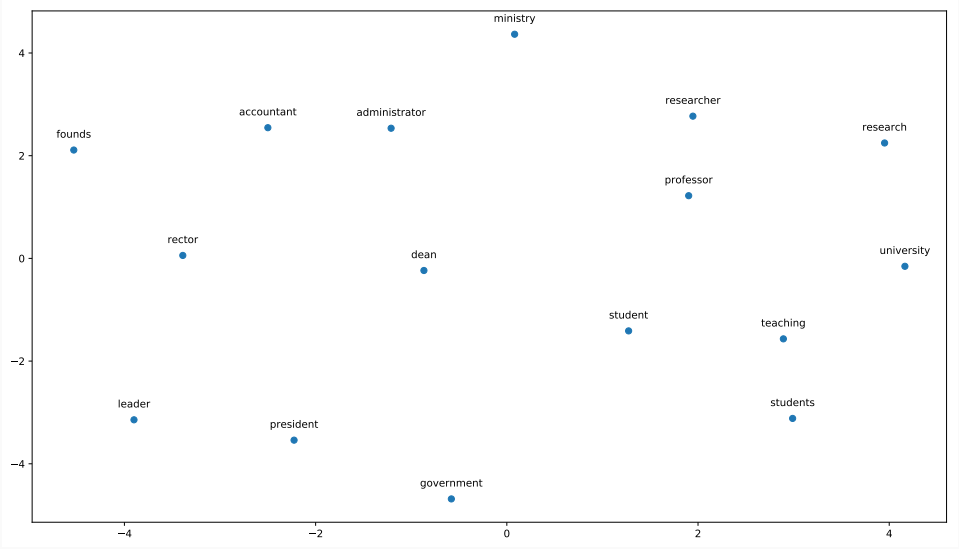
Representation of the meaning of a word - Glove: Examples

Most similar words in a corpora Glove6B[1] using Euclidean distance

man		woman		queen		university		learning	
2.60	woman	2.43	girl	3.01	princess	3.23	college	2.65	teaching
2.81	another	2.60	man	3.16	lady	3.29	harvard	2.93	practical
2.81	boy	2.79	mother	3.30	elizabeth	3.42	graduate	2.93	experience
2.97	one	2.88	boy	3.39	prince	3.51	institute	3.10	knowledge
3.02	old	3.08	her	3.44	coronation	3.54	yale	3.10	lessons
3.04	turned	3.14	she	3.48	king	3.58	professor	3.14	skills
3.07	whose	3.17	herself	3.57	consort	3.72	faculty	3.15	instruction
3.15	himself	3.38	victim	3.62	victoria	3.74	school	3.16	classes
3.15	who	3.38	child	3.67	crown	3.83	graduated	3.17	learn
3.24	friend	3.45	husband	3.69	bride	3.86	academy	3.18	studying
3.24	him	3.47	old	3.73	majesty	3.90	princeton	3.19	teach

[1] https://nlp.stanford.edu/projects/glove/

Representation of the meaning of a word - Glove: Examples



Representation of the meaning of a word

Word2Vec as well as GloVe generates context-independent representation.

Representation of the meaning of a word

Word2Vec as well as GloVe generates context-independent representation.

Context dependent embedding may be generated using ELMo and other approaches.

Tensorflow for Text Processing

Tensorflow for Text Processing

- tf.keras.layers.experimental.preprocessing.**TextVectorization**

max_tokens is the maximum tokens produced by the Vectorizer

standardize is the standardization option applied on the strings. Default is LOWER_AND_STRIP_PUNCTUATION = convert to lower case and remove punctuation.

split - splitting strategy on input strings, default is SPLIT_ON_WHITESPACE

ngrams defines if the strings are combined into ngrams with their length.

output_mode defines the type of the output, possible options are:

- **INT** - outputs integer indices, one per unique tokens
- **BINARY** - outputs single int array per batch, where 1s are in place of words.
- **COUNT** - output real count instead of 1s
- **TF-IDF** - similar to binary but includes tf-idf values.

Tensorflow for Text Processing

- tf.keras.layers.**Embedding**

  input_dim is the size of the vocabulary, i.e. maximum integer index + 1.

  output_dim is the dimension of the dense embedding.

  embeddings_initializer is the initializer.

  embeddings_regularizer is the regularizer function applied to the embeddings matrix.

  embeddings_constraint is the constraint function applied to the embeddings matrix.

  mask_zero defines whether or not the input value 0 is a special ”padding” value that should be masked out.

  input_length defines the shape when flatten layers is used. Not used when variable length is used.

Tensorflow for Text Processing

```
embedding_dim=16
vocab_size = 10000
sequence_length = 100

vectorize_layer = TextVectorization(
    standardize=LOWER_AND_STRIP_PUNCTUATION,
    max_tokens=vocab_size,
    output_mode='int',
    output_sequence_length=sequence_length)

vectorize_layer.adapt(dataset) #match the vectorizer to dataset

model = Sequential([
  vectorize_layer,
  Embedding(vocab_size, embedding_dim, name="embedding"),# learning embedding
  GlobalAveragePooling1D(),
  Dense(16, activation='relu'),
  Dense(1)
])
```

Questions?