Chapter NLP:VIII

VIII. Text Representation Models

Representation of semantic properties

Word Embeddings

Classical semantic representation with DTM's

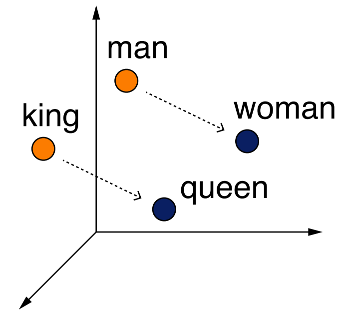
- Can be very large and unhandy
- Example $\rightarrow$ Sentences: 26,142,898, Types: 5,876,655 Term-Term-Matrix of Dimension 5,876,655 $\times$ 5,876,655, **3.8TB of data with 32 Bit Integer**
- How to compare term similarity?
- How to find word context efficiently?
- **Idea: Dimension reduction of semantic space!**

	dog	plays	children	playing	sun	...	shining	burning	fire	moon	candle	agree	fact
dog	0	1	0	0	0		0	0	0	0	0	0	0
plays	1	0	0	0	0		0	0	0	0	0	0	0
children	0	0	0	1	0		0	0	0	0	0	0	0
playing	0	0	1	0	0		0	0	0	0	0	0	0
sun	0	0	0	0	0		1	1	0	0	0	0	0
...													
shining	0	0	0	0	1		0	1	0	1	1	0	0
burning	0	0	0	0	1		1	0	1	0	1	0	0
fire	0	0	0	0	0		0	1	0	0	0	0	0
moon	0	0	0	0	0		1	0	0	0	0	0	0
candle	0	0	0	0	0		1	1	0	0	0	0	0
agree	0	0	0	0	0		0	0	0	0	0	0	1
fact	0	0	0	0	0		0	0	0	0	0	1	0

Representation of semantic properties
Word Embeddings

Extension of the distributional idea

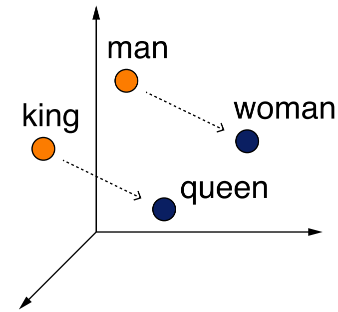
- Representation of a word by the context it occurs in.
- To do so, words are mapped to an *embedding space* where contextually related words are similar.



Representation of semantic properties

Word Embeddings

Extension of the distributional idea



- Representation of a word by the context it occurs in.
- To do so, words are mapped to an *embedding space* where contextually related words are similar.

Word embedding (aka word vector)

- A real-valued vector that represents the *distributional semantics* of a particular word in the embedding space.

$$\text{“}king\text{''} \quad \rightarrow \quad v_{king} = (0.13, 0.02, 0.1, 0.4, \ldots, 0.22)$$

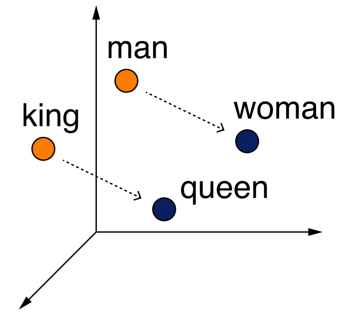
- The longer the vector, the more variance is kept (typical: 100–500).

Representation of semantic properties

Word Embeddings

Extension of the distributional idea

- Representation of a word by the context it occurs in.
- To do so, words are mapped to an *embedding space* where contextually related words are similar.



Word embedding (aka word vector)

- A real-valued vector that represents the *distributional semantics* of a particular word in the embedding space.

$$\text{“}king\text{''} \quad \rightarrow \quad v_{king} = (0.13, 0.02, 0.1, 0.4, \ldots, 0.22)$$

- The longer the vector, the more variance is kept (typical: 100–500).

Some properties of embedding spaces

- Similar context results in similar embeddings. projector.tensorflow.org
- Analogies are arithmetically represented. turbomaze.github.io/word2vecjson

$$v_{king} - v_{man} + v_{woman} \approx v_{queen} \qquad v_{france} - v_{paris} + v_{berlin} \approx v_{germany}$$

Representation of semantic properties

Embedding Models

Word embedding model

- A function that maps each known word to its word embedding.
- Such mappings are created unsupervised based on huge corpora, capturing the likelihood of words occurring in sequence.

  The technical details are beyond the scope of this course.

Several software libraries and pre-trained models exist

- Libraries. Glove, word2vec, Fasttext, Flair, Bert, ...
- Models. GoogleNews-vectors, ConceptNet Numberbatch, ...

Representation of semantic properties

Embedding Models

Word embedding model

- A function that maps each known word to its word embedding.
- Such mappings are created unsupervised based on huge corpora, capturing the likelihood of words occurring in sequence.

  The technical details are beyond the scope of this course.

Several software libraries and pre-trained models exist

- Libraries. Glove, word2vec, Fasttext, Flair, Bert, ...
- Models. GoogleNews-vectors, ConceptNet Numberbatch, ...

From word embeddings to text embeddings

- Simple. Average the embeddings of each word in a text.
- More sophisticated. Learn embeddings for sentences or similar.
- In general, the longer the text, the harder it is to capture its semantics in an embedding.

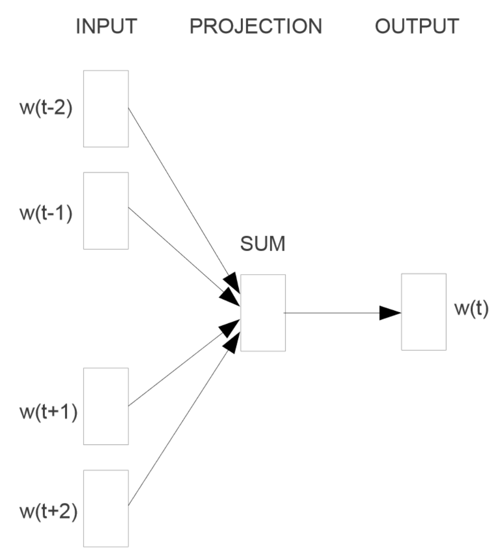
Representation of semantic properties

Word2Vec Example

Main Idea [Mikolov et al. 2013]

- Shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.
- The projection layer transforms the input to the output: **Similar words or contexts need a similar weight vector** in order to be transformed to the same target.

- **Continuous bag-of-words architecture (CBOW)**: the model predicts the current word from a window of surrounding context words.
- CBOW is faster while skip-gram is slower but does a better job for infrequent words



CBOW

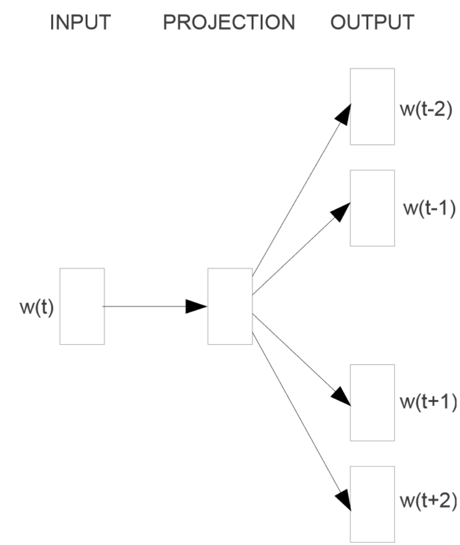
Representation of semantic properties

Word2Vec Example

Main Idea [Mikolov et al. 2013]

- Shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words.
- The projection layer transforms the input to the output: **Similar words or contexts need a similar weight vector** in order to be transformed to the same target.

- **Continuous skip-gram architecture (SKIPGRAM)**: the model uses the current word to predict the surrounding window of context words.



Skip-gram

Representation of semantic properties

Word Vectors as Feature

Main Idea

- Replacement of vectors among vocabulary by semantic embedding vectors
- Better capturing of semantic properties, composition and ambiguities
- Modern language model embeddings (Bert, Elmo, Flair, FastText) use character sequence embeddings → no (O)ut (O)f (V)ocabulary Problem in prediction, robust to typos

Example: Enriching Word Vectors with Subword Information [Bojanowski et al. 2017]