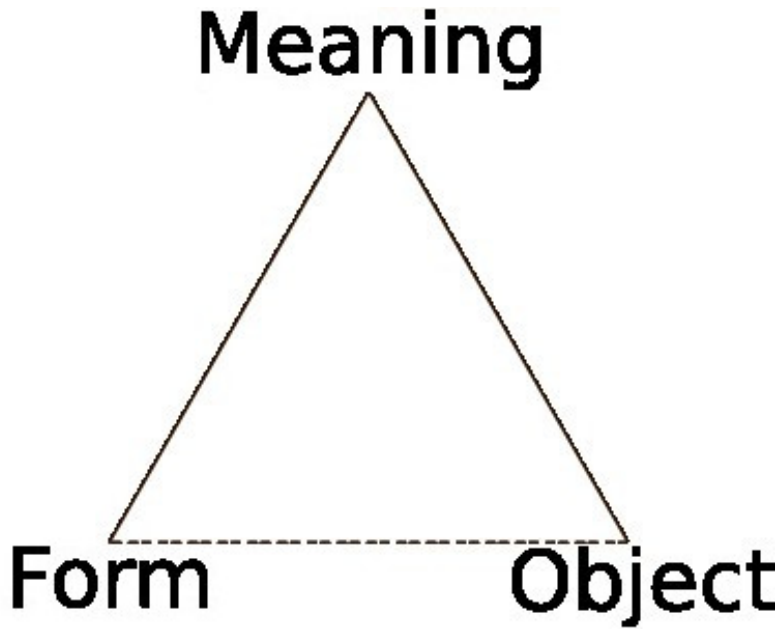
Przetwarzanie Języka Naturalnego
Lab 9 – Word Embedding

Aleksander Smywiński-Pohl

Wydział IEiT
Katedra Informatyki

16.05.2017

Trójkąt semiotyczny



Jak opisać znaczenie słowa

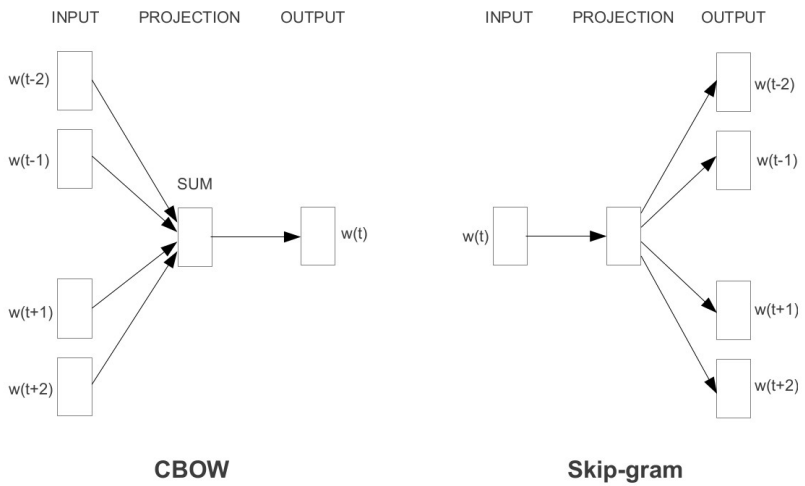
- definicja
- konteksty (“a word is characterized by the company it keeps”)
- relacje paradygmatyczne
- relacje syntagmatyczne

Konwersja tekstu na wektor

- BOW, tf-idf itp.
- LSA (zliczanie)
- word embedding (przewidywanie)

Word embedding

- Continous Bag of Words
- Skip-gram



Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”

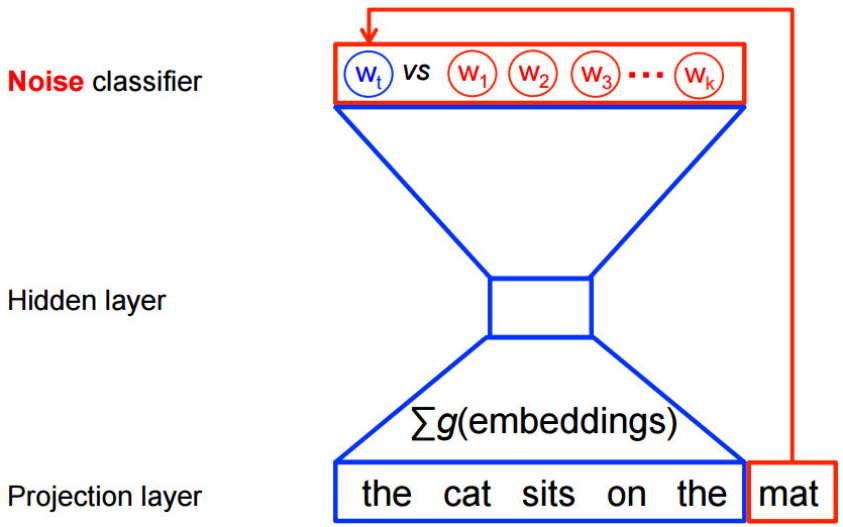
Softmax

$$z = (z_1, z_2, ..., z_k)$$
$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{i=1}^{k} e^{z_i}}$$

Algorytm

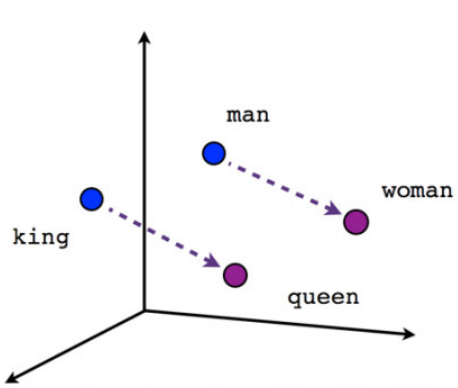
$$P(w_t|h) = softmax(score(w_t, h))$$

Algorytm c.d.

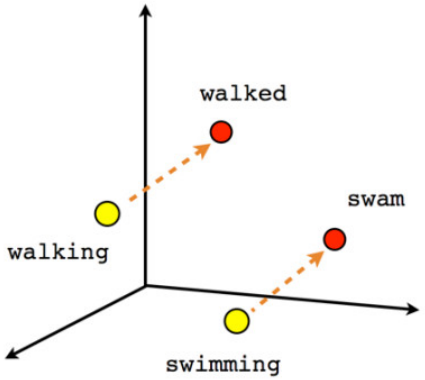


https://www.tensorflow.org/tutorials/word2vec

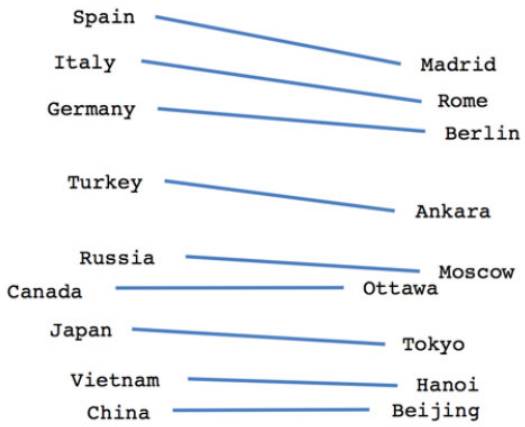
Algorytm



Male-Female



Verb tense



Country-Capital

https://www.tensorflow.org/tutorials/word2vec

Zadania

1. wytrenować model word2vec na korpusie zawierającym minimum 100 mln słów w j. polskim (1 pkt)
2. wybrać trzy relacje morfosyntaktyczne dla różnych części mowy (np. dopełniacz–mianownik dla rzeczownika, czas teraźniejszy–czas przeszły dla czasownika) i przetestować ich zachowanie w wytrenowanym modelu (1 pkt)
3. wybrać trzy relacje semantyczne (np. państwo–stolica, część–całość) i przetestować ich zachowanie w wytrenowanym modelu (1 pkt)

Materiały

- https://www.tensorflow.org/tutorials/word2vec
- Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”
- Mikolov et al. “Distributed Representations of Words and Phrases and their Compositionality”