Hasso Plattner Institut

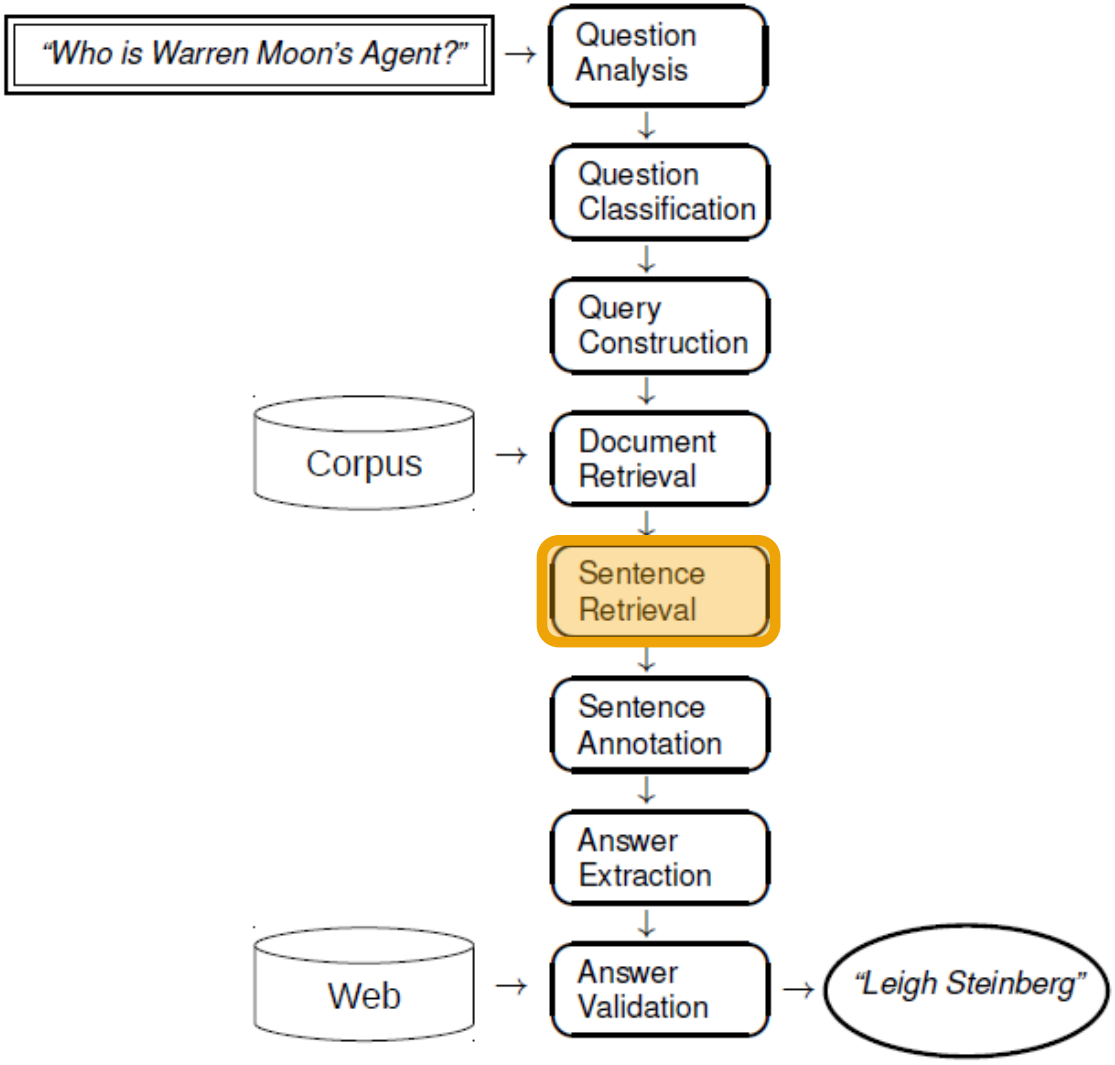
IT Systems Engineering | Universität Potsdam

Word Relationships in Sentence Retrieval

Philipp Langer

January 2nd, 2012

Where are we at?



term mismatch problem

Given

- a query
- documents relevant to the query

Find

- sentences relevant to the query

Problem

Query: “The Mount of Olives is just east of which city?“

Sentence: “The Mount of Olives is located near the town of Jerusalem.“

solution: a class-based LM

using *term clustering*

idea:

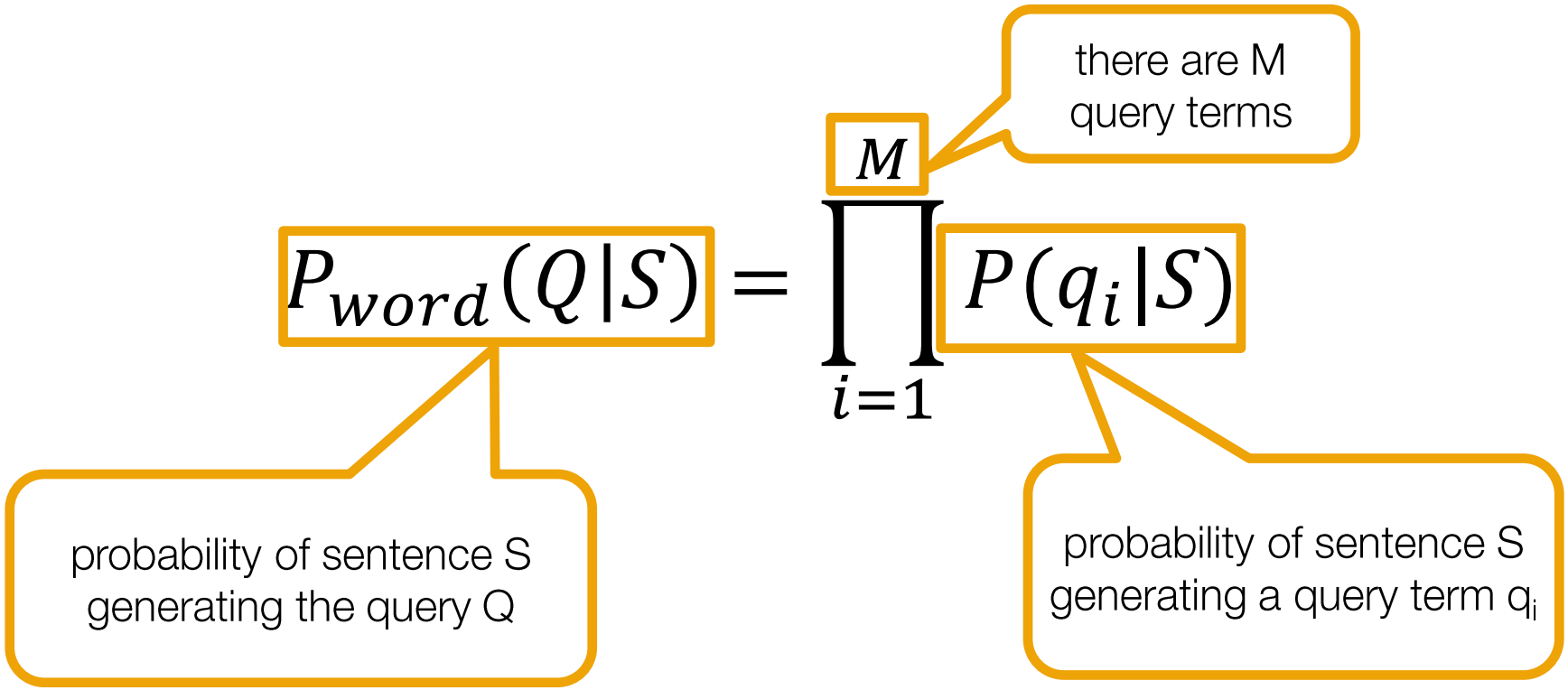
- relax the “exact match criterion“ using word clusters
- word clusters contain related words
- now, the LM is defined on clusters of words, not on single words

example cluster:

- { city, town }

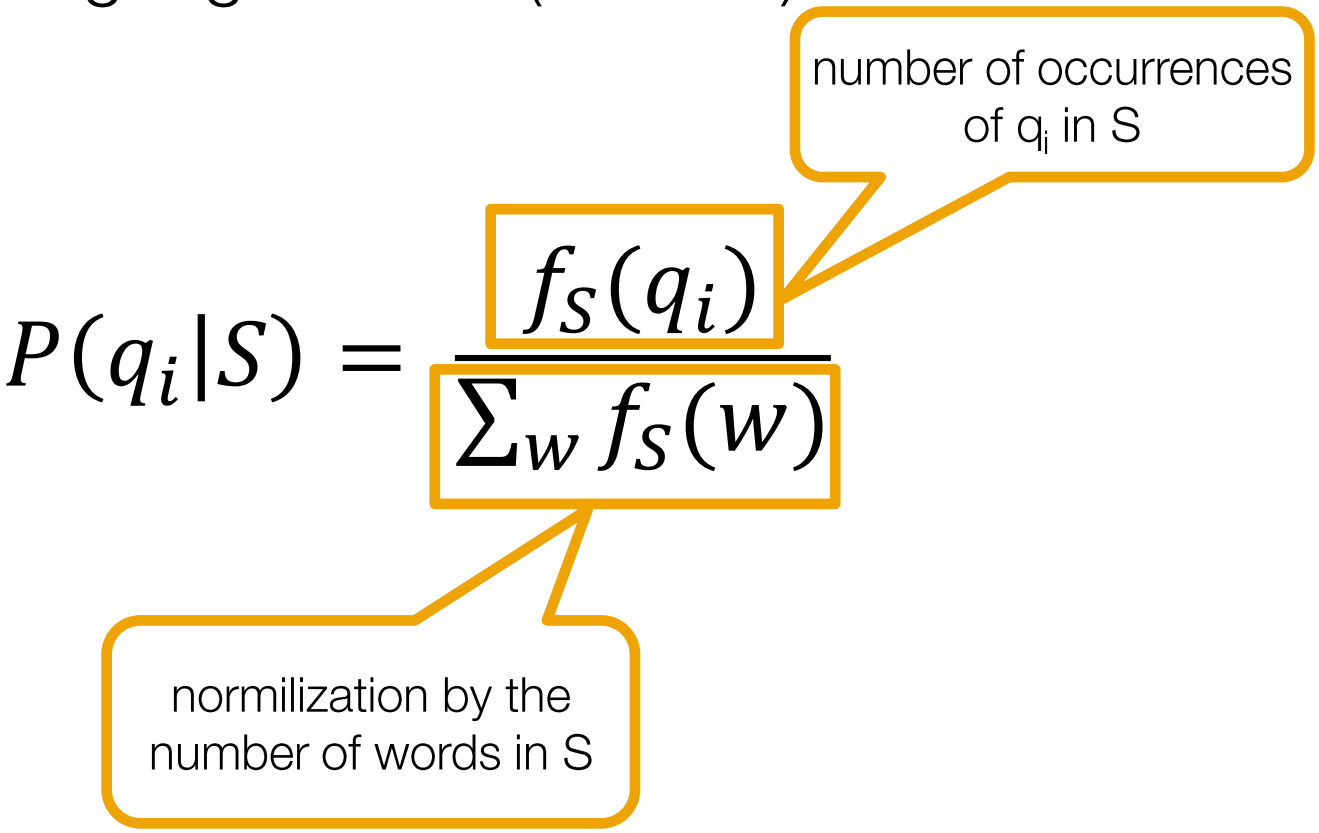
definition: the word-based LM

word-based language model (revision)



definition: the word-based LM

word-based language model (revision)



defintion: the class-based LM

the class-based language model (1)

$$\prod_{i=1}^{M} P(C_{q_i}|S)$$

$$P(C_{q_i}|S) = \frac{f_S(C_{q_i})}{\sum_w f_S(w)}$$

number of occurrences in S of all words that are in the same cluster as q_i

defintion: the class-based LM

the class-based language model (2)

$$P_{class}(Q|S) = \prod_{i=1}^{M} \underbrace{P(q_i|C_{q_i}, S)}_{\text{emission probability}} P(C_{q_i}|S)$$

emission probability

Cluster C_{q_i} , cluster words $t_k \in C_{q_i}$, sentence words s_l
emission probabilities:
$s_1: (t_1: 0.1), (t_2: 0.3), \dots, (q_i: 0.4), \dots$
$s_2: (t_1: 0.1), (t_2: 0.2), \dots, (q_i: 0.5), \dots$
$s_3: (t_1: 0.1), (t_2: 0.4), \dots, (q_i: 0.2), \dots$

building clusters

Brown word clustering algorithm

- input: words from a vocabulary, designated number of clusters

idea:

- put each word into one cluster
- greedily merge clusters with minimal loss of mutual information until predefined number of clusters is reached

- needs a common notion of mutual information of clusters

average mutual information

$$AMI(C_w, C_{w'}) = \sum_{C_w, C_{w'}} f(C_w, C_{w'}) log \frac{f(C_w, C_{w'})}{f(C_w) f(C_{w'})}$$

number of times that words in the cluster C_w occur in the same context as $C_{w'}$

number of times that words from the cluster C_w occur in the corpus

word co-occurrence

document-wise co-occurrence

sentence-wise co-occurrence

word co-occurrence

document-wise co-occurrence

sentence-wise co-occurrence

co-occurrence in a window of text (bigram)

- “There are no lectures *on* Sunday.“
- “QA takes place *on* Monday.“

co-occurrence in a syntactic relationship

other approaches

translate sentence terms to query terms

lexicon (thesaurus)

- almost no effect on the results

WordNet

- better than thesauri, but still little effect

English-Arabic / Arabic-English lexicons

- best of these approaches

references

clustering approach and the definition of the class-based and word-based language models:

Saeedeh Momtazi, Classification in Question Answering Systems, PhD Thesis, 2010 (Chapter 5 + 8.3.1)

“other approaches“:

Vanessa Murdock, W. Bruce Croft, A Translation Model for Sentence Retrieval, EMNLP Conference, 2005