Czech Information Retrieval with Syntax-based Language Models

Jana Straková a Pavel Pecina

Institute of Formal and Applied Linguistics

Charles University in Prague

How can we improve information retrieval?

(Especially for morphologically rich languages with considerable free word order and long distance relations between words?)

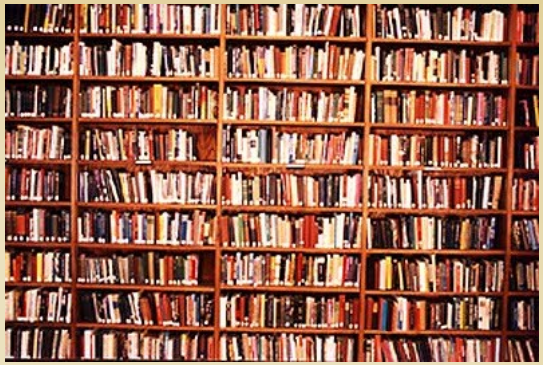
Outline

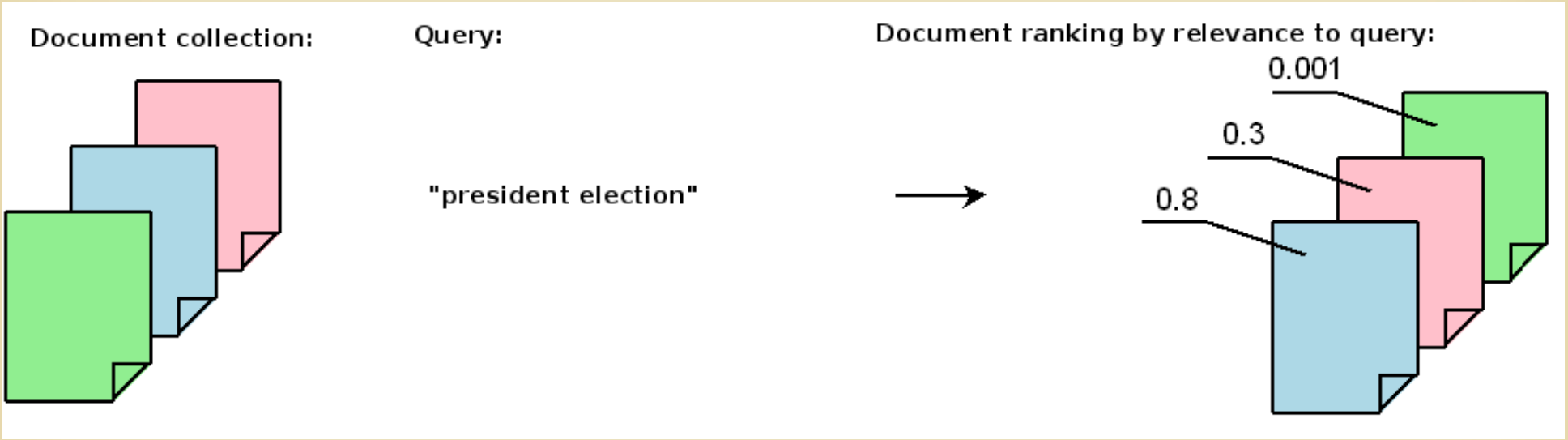
- Motivation
- The Task
- Test Collection
- The Model
- Experimental Setup
- Results and discussion
- Conclusions

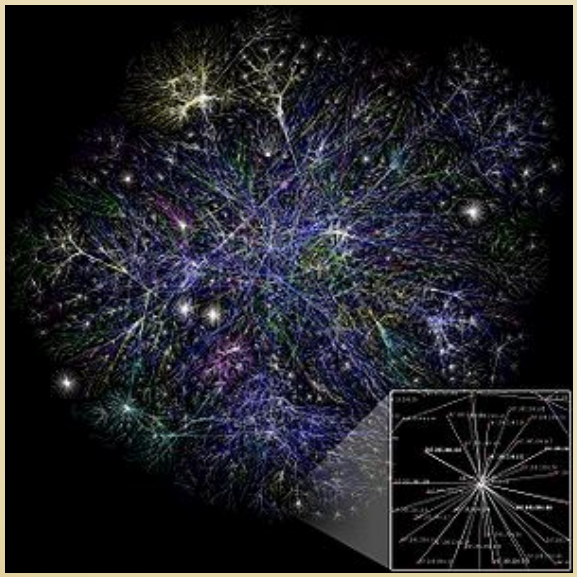
Outline

- ~~Motivation~~
- The Task
- Test Collection
- The Model
- Experimental Setup
- Results and discussion
- Conclusions

The Task







For given document collection and given query, rank documents with relevance to the query.

Test Collection

- Czech collection from Cross Language Evaluation (CLEF) Forum 2007 Ad-Hoc Track
- 81,735 documents, 50 topics
- average document length: 349.46 words
- 15.24 documents in average assessed as relevant to each topic

Test Collection

- Czech collection from Cross Language Evaluation (CLEF) Forum 2007 Ad-Hoc Track
- 81,735 documents, 50 topics
- average document length: 349.46 words
- 15.24 documents in average assessed as releavant to each topic
- Results on this shared task published in Nunzio et al., 2008:
  - MAP: 35.68%, 34.84%, 32.04%
  - best known MAP: 42.42% (Dolamic, Savoy (2008))

Nunzi, Ferro, Mandl (2008):
CLEF 2007: Ad Hoc Track Overview

Dolamic, Savoy (2008):
Stemming Approaches
for East European Languages

Topics

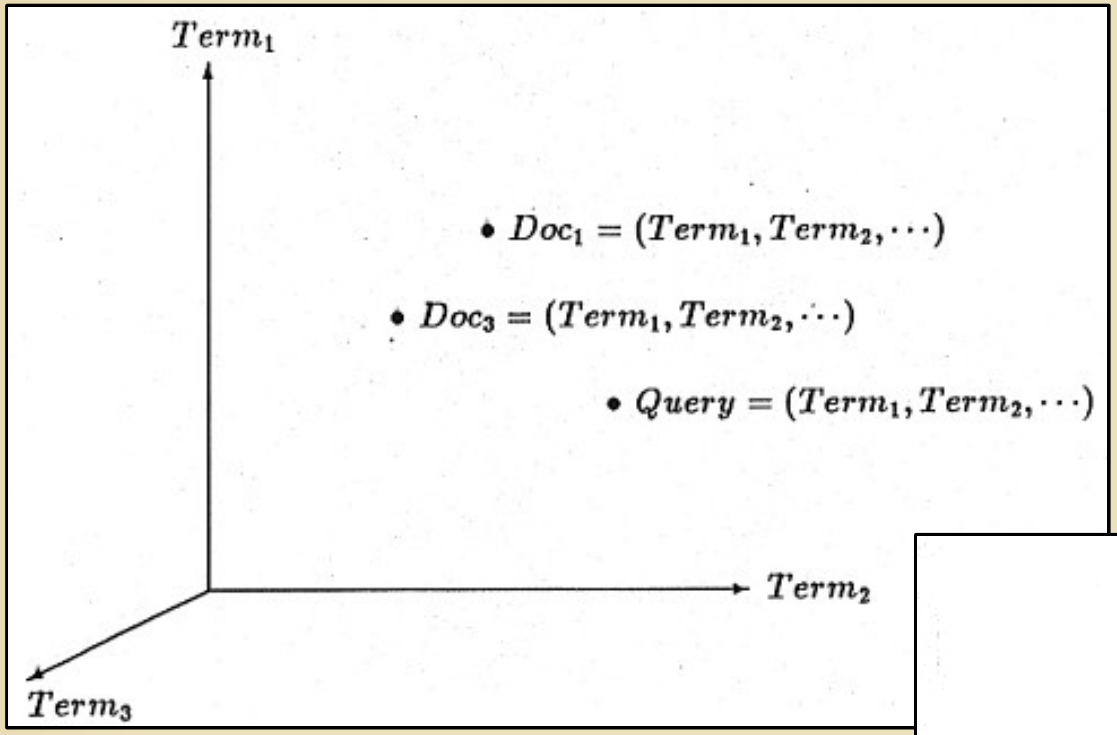
- Queries describing „information need“ in natural language.
- TREC format: a structure of three fields
  - title: keyword query
  - desc: more detail (one sentence)
  - narr: detailed description of relevant documents
- Randomly divided into a development set of 10 topics and test set of 40 topics.

Topic Example

```
<title>

Inflace Eura

</title>

<desc>

Najděte dokumenty o růstech cen po zavedení
Eura.

</desc>

<narr>

Relevantní jsou jakékoli dokumenty, které
poskytují informace o růstu cen v jakékoli zemi,
v níž byla zavedena společná evropská měna.

</narr>
```

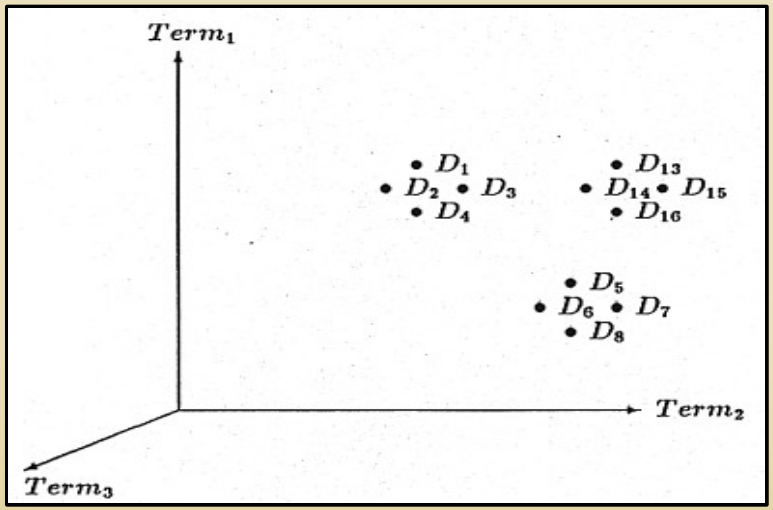
Vector space model for IR

Salton, Buckley (1987):
Term Weighting Approaches
in Automatic Text Retrieval



$$sim(A, B) = \frac{A \bullet B}{|A|\,|B|}$$

$$\text{Cosine}(\text{doc}_i, \text{doc}_j) = \frac{\sum_{k=1}^{t} TERM_{ik} * TERM_{jk}}{\sqrt{\sum_{k=1}^{t} TERM_{ik}^2 * \sum_{k=1}^{t} TERM_{jk}^2}}$$



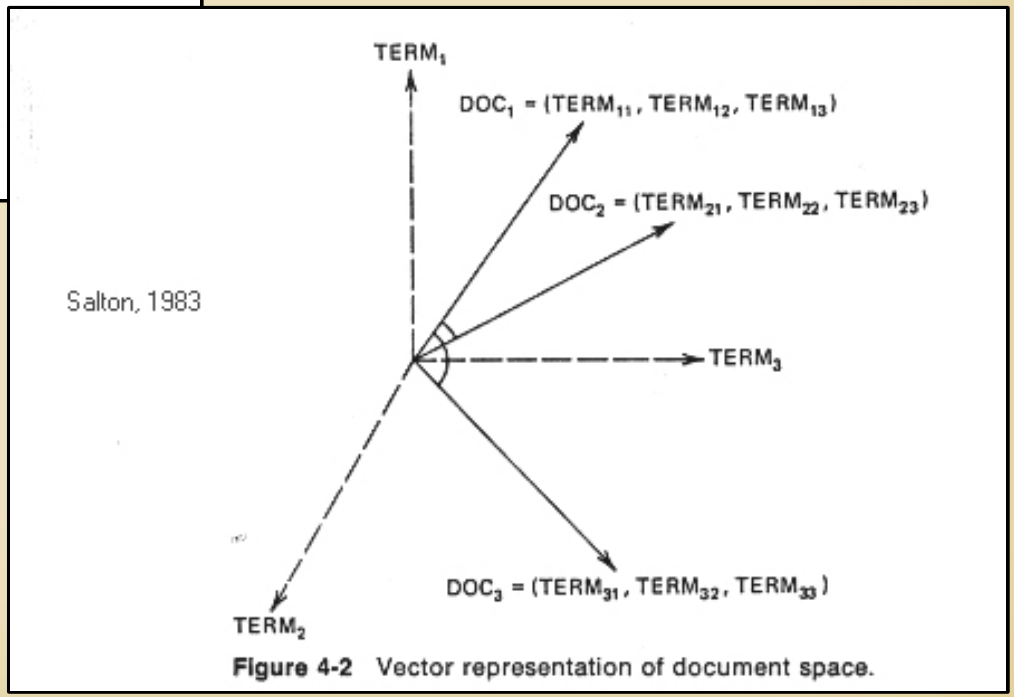


Figure 4-2 Vector representation of document space.

Language modeling in IR

- Notation:
  - document: D
  - collection of documents: C
  - query: $Q = q_1, q_2, \ldots, q_n$
  - surface bigram: (q_i, q_{i+1})
  - dependency bigram: $(p(q_i), q_i)$
- Documents D are ranked by probability P(D|Q) of being (independently) generated from queries Q.
- From Bayes, we consider „reverted“ probability P(Q|D).

Ponte, Croft (1998):
A language modeling approach to information retrieval

Manning, Raghavan, Schutze (2008):
Introduction to Information Retrieval

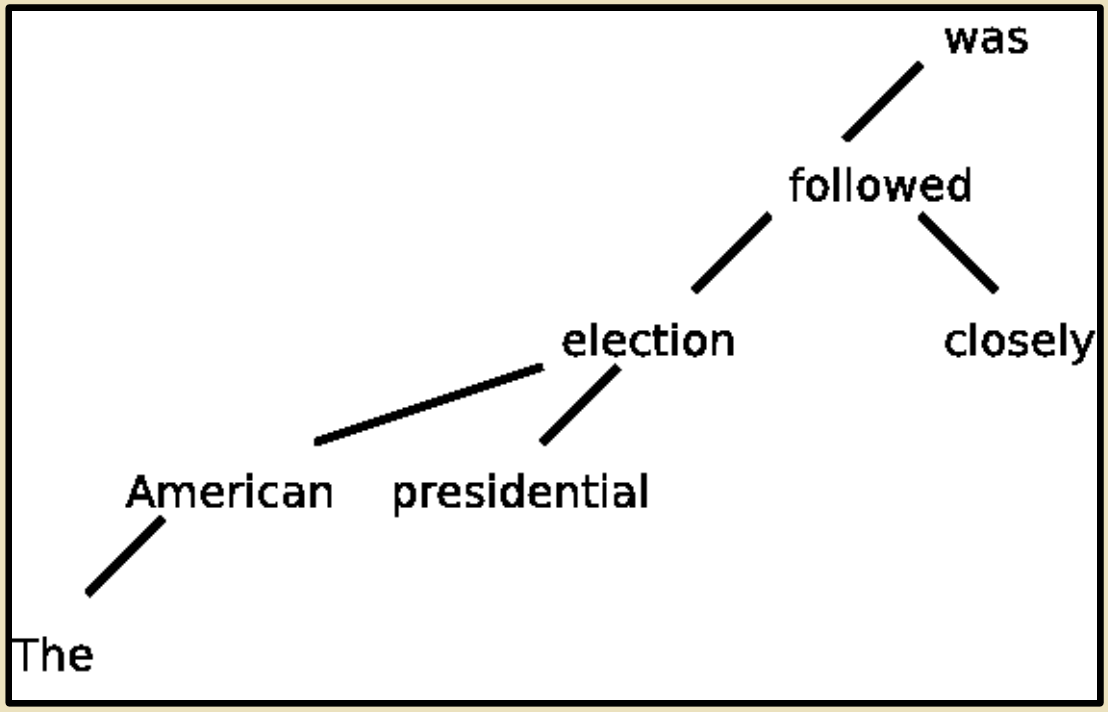
Manning, Schutze (1999):
Foundations of Statistical Natural Language Processing

Language models

Dolamic, Savoy (2008):
Stemming Approaches
for East European Languages

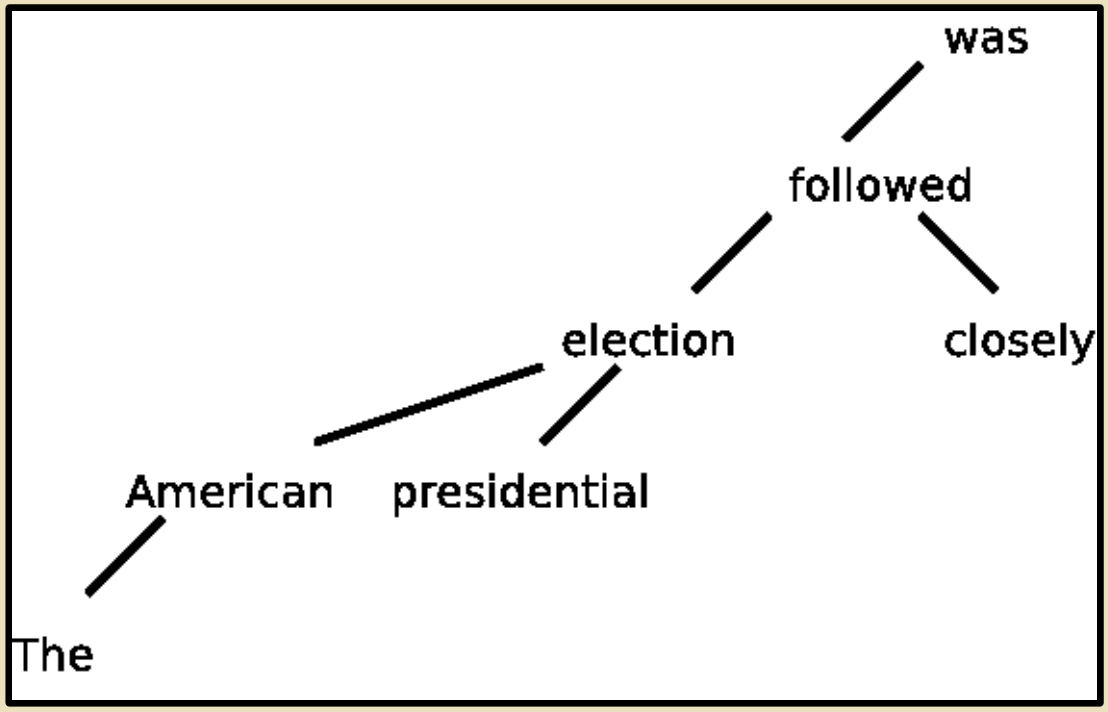
- Unigram model
  - $P_D(Q) = \prod P_D(q_i) = \prod \frac{C_D(q_i)}{|D|}$
  - Where $P_D(Q)$ stands for P(D|Q) and $C_D(q_i)$ is the raw count of word q_i in document D
- Bigram (surface) model
  - $P_D(Q) = \prod P_D(q_i, q_{i+1}) = \prod \frac{C_D(q_i, q_{i+1})}{|D|}$

Dependency tree



Dependency tree for sentence „The American presidential election was followed closely.“

Dependency bigram model



$$P_D(Q) = \prod_{q_i : \exists\, p(q_i)} P_D(p(q_i), q_i)$$

Experimental Setup

- baseline: plain unigram model
- comparison: surface vs. dependency bigram model

Experimental Setup

- baseline: plain unigram model
- comparison: surface vs. dependency bigram model
- lemmatization (= linguistically motivated means of stemming)
- smoothing: Jelinek-Mercer

Experimental Setup

- baseline: plain unigram model
- comparison: surface vs. dependency bigram model
- lemmatization (= linguistically motivated means of stemming)
- smoothing: Jelinek-Mercer
- combination of all models by simple linear interpolation
  - coefficients fitted by simple grid search using development data
- Stopwords: 256 words from UniNE

Experimental Setup II (Tools)

- lemmatization: Hajič, 2004
- parsing: McDonald et al., 2005
- evaluation: MAP with trec_eval
- morphological and syntax analysis performed in TectoMT framework (Žabokrtský et al., 2008)

Results

model	MAP
unigram-surface-form	0.3116
unigram-surface-lemma	0.3731
bigram-surface-form	0.1775
bigram-surface-lemma	0.2023
bigram-dependency-form	0.1826
bigram-dependency-lemma	0.2447
combination	0.3890

Results

model	MAP
unigram-surface-form	0.3116
unigram-surface-lemma	0.3731
bigram-surface-form	0.1775
bigram-surface-lemma	0.2023
bigram-dependency-form	0.1826
bigram-dependency-lemma	0.2447
combination	0.3890

Results

model	MAP
unigram-surface-form	0.3116
unigram-surface-lemma	0.3731
bigram-surface-form	0.1775
bigram-surface-lemma	0.2023
bigram-dependency-form	0.1826
bigram-dependency-lemma	0.2447
combination	0.3890

Results

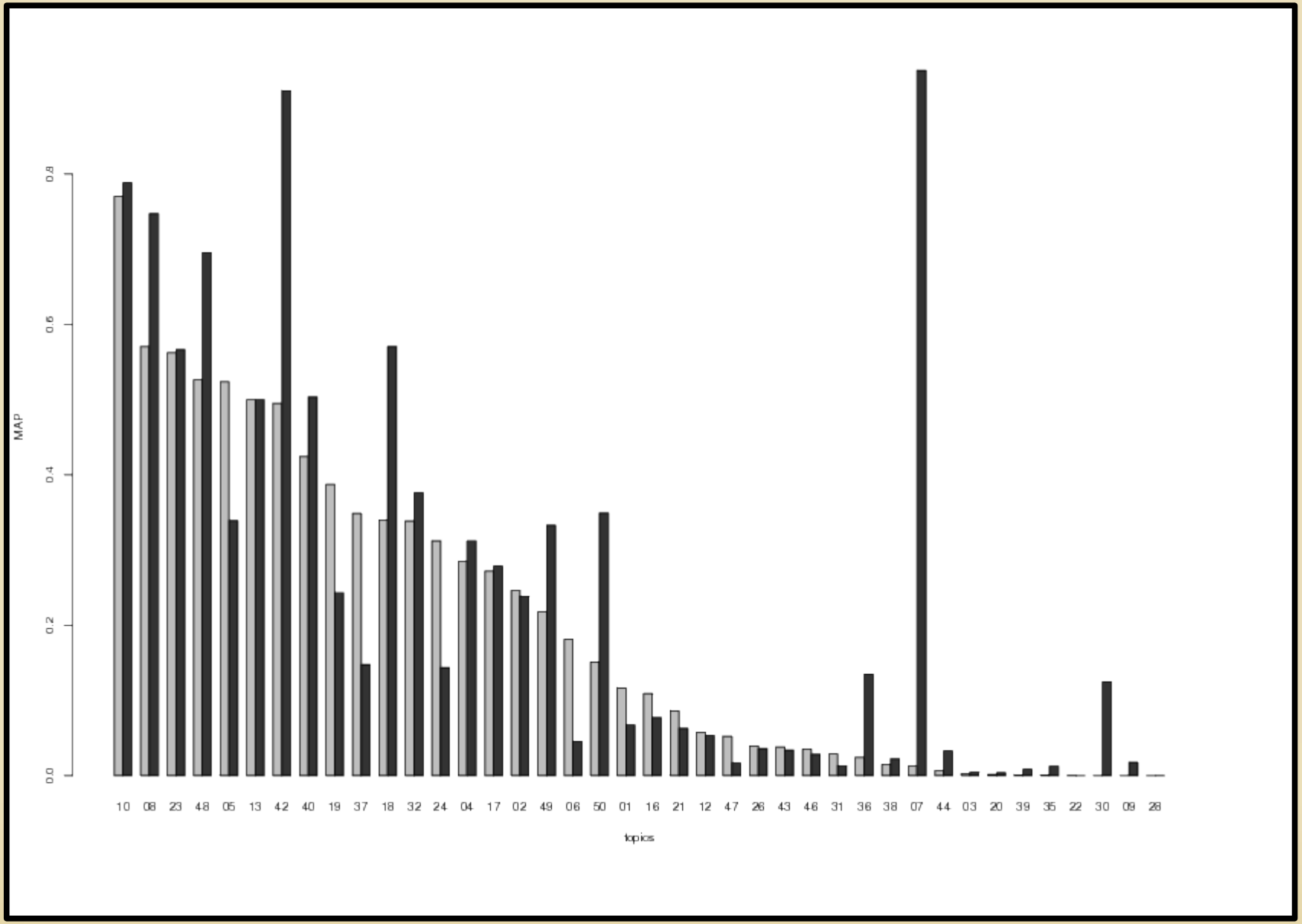
model	MAP
unigram-surface-form	0.3116
unigram-surface-lemma	0.3731
bigram-surface-form	0.1775
bigram-surface-lemma	0.2023
bigram-dependency-form	0.1826
bigram-dependency-lemma	0.2447
combination	0.3890

(all 50 topics MAP: 41.02)

Results

unigram-surface-lemma	0.3731
bigram-surface-form	0.1775
bigram-surface-lemma	0.2023
bigram-dependency-form	0.1826
bigram-dependency-lemma	0.2447
combination	0.3890

Bigram surface (20.23) vs. bigram dependency (24.47)



Conclusions

- We have presented a simple dependency bigram language model for information retrieval.
- With this model, we have outperformed most of the results published in Nunzio et al., 2008.
- Finally, we have found examples, where syntax model performs significantly better than surface bigram model.

Thank you!