Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval

Christof Müller and Iryna Gurevych

Domain-Specific Track

- Monolingual tasks
  - English
  - German
  - Russian

- Bilingual tasks
  - English topics – German documents

Objectives

- Overcome vocabulary gap between queries and documents

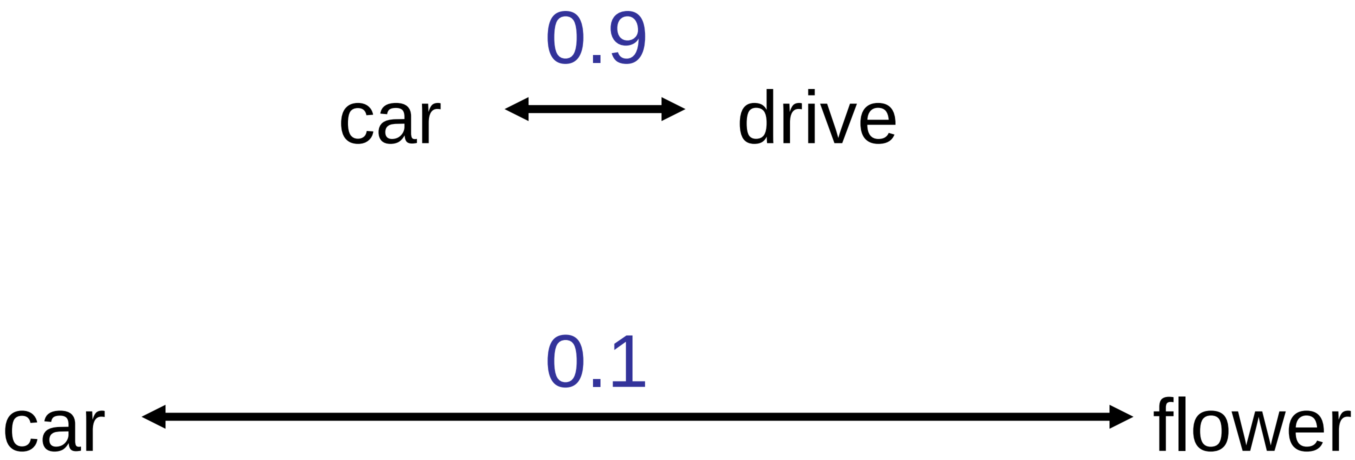
- Methods
  - Query Expansion
  - (Blind) Relevance Feedback

- Our approach:
  - **Semantic Relatedness (SR)** of query and document terms

Outline

- IR models based on SR
- Preprocessing of queries and documents
- Results for monolingual runs: English, German, Russian
- Approach for bilingual IR
- Results for bilingual runs: English – German
- Additional experiments concerning efficiency
- Summary and future work

Semantic Relatedness Measure



car ↔ drive: 0.9

car ↔ flower: 0.1

SR Measures and Knowledge Bases

- Various SR measures
  - Path based, information content based, dictionary based
- Using linguistic knowledge bases like WordNet

- Our approach:
  - Using **collaborative knowledge bases** like Wikipedia and Wiktionary
    - +: coverage, domain-specific terms, up-to-dateness
    - -: quality, accessibility

  - **Concept vector based** SR measure
    - Represent terms and texts as concept vectors
    - Compare concept vectors to compute SR
    - **Explicit Semantic Analysis** (Gabrilovich & Markovitch, 2007)

Constructing Concept Vectors from Knowledge Sources

Knowledge Sources

Concepts

Article Titles

Textual Representation

Article Text

Representing Words as Concept Vectors

taxicab

0.8	automobile
0.7	drive
0.6	fast
0.8	hire
0.6	New York
0.8	passenger
0.1	SUV
0.9	taxi
0.8	transport
0.8	yellow

In some countries, taxicabs are commonly yellow. This practice began in Chicago, where taxi entrepreneur John Hertz painted his taxis yellow based on a University of Chicago study alleging that yellow is the color most easily seen at a distance.

Concept Vector Measure

taxicab

automobile	0.8
drive	0.7
fast	0.6
hire	0.8
New York	0.6
passenger	0.8
SUV	0.1
taxi	0.9
transport	0.8
yellow	0.8

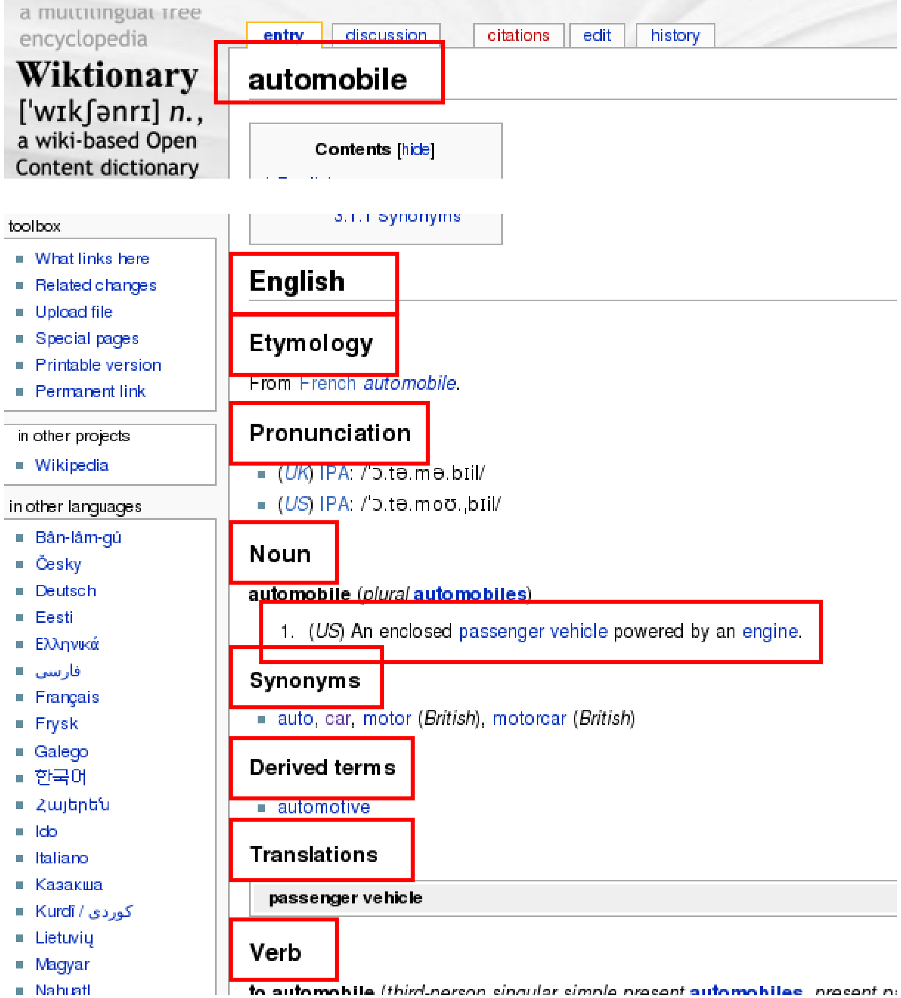
cosine similarity

0.7	automobile
0.8	drive
0.2	fast
0.1	hire
0.0	New York
0.1	passenger
0.0	SUV
0.0	taxi
0.9	transport
0.1	yellow

truck

$$\overline{V}_{taxicab} \times \overline{V}_{truck} = \text{SR score}$$

New Resource for IR:
Wiktionary – Wikipedia‘s lexical companion



a multilingual free encyclopedia

Wiktionary ['wɪkʃənrɪ] *n.*, a wiki-based Open Content dictionary

entry | discussion | citations | edit | history

automobile

Contents [hide]

3.1.1 Synonyms

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

in other projects

- Wikipedia

in other languages

- Bân-lâm-gú
- Česky
- Deutsch
- Eesti
- Ελληνικά
- فارسی
- Français
- Frysk
- Galego
- 한국어
- Հայերեն
- Ido
- Italiano
- Қазақша
- Kurdî / كوردی
- Lietuvių
- Magyar
- Nahuatl

English

Etymology

From French *automobile*.

Pronunciation

- (*UK*) IPA: /'ɔ.tə.mə.bɪil/
- (*US*) IPA: /'ɔ.tə.moʊ.ˌbɪil/

Noun

automobile (*plural* **automobiles**)

1. (*US*) An enclosed passenger vehicle powered by an engine.

Synonyms

- auto, car, motor (*British*), motorcar (*British*)

Derived terms

- automotive

Translations

passenger vehicle

Verb

to automobile (*third-person singular simple present* **automobiles**, *present p*

- Language
- Etymology
- Pronunciation
- Part-of-speech
- Word senses
- Synonyms
- Derived Terms
- Translations

- Abbreviations, Antonyms, Categories, Collocations, Examples, Glosses, Hypernyms, Hyponyms, Morphology, Quotations, Related terms, Troponyms

Constructing Concept Vectors from Knowledge Sources

Knowledge Sources	WIKIPEDIA The Free Encyclopedia	Wiktionary ['wɪkʃənrɪ] *n.*, a wiki-based Open Content dictionary
Concepts	Article Titles	Entry Titles
Textual Representation	Article Text	Entry Information

(Zesch et al., 2008)

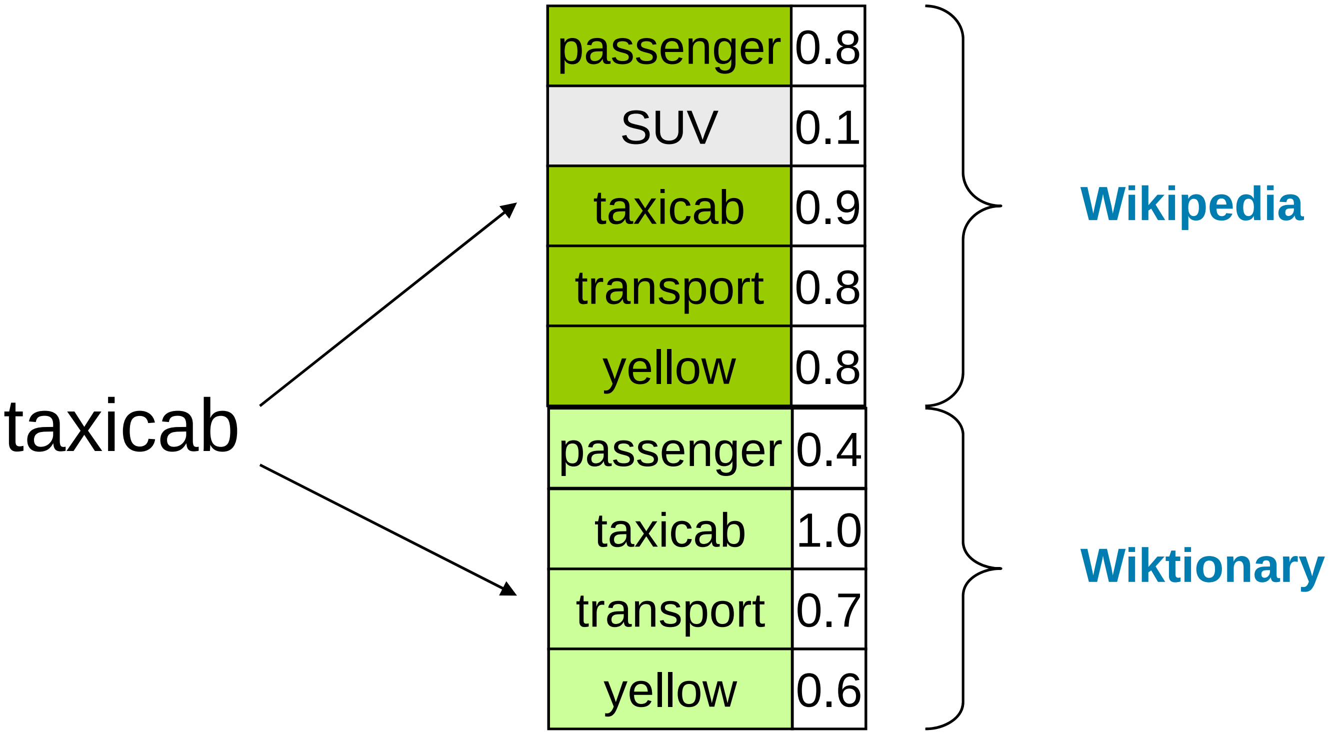
Constructing Concept Vectors from Knowledge Sources

TECHNISCHE UNIVERSITÄT DARMSTADT

Knowledge Sources	WIKIPEDIA The Free Encyclopedia	+	Wiktionary [ˈwɪkʃənrɪ] *n.*, a wiki-based Open Content dictionary
Concepts	Article Titles	+	Entry Titles
Textual Representation	Article Text		Entry Information

(Zesch et al., 2008)

Combination of Wikipedia and Wiktionary



taxicab

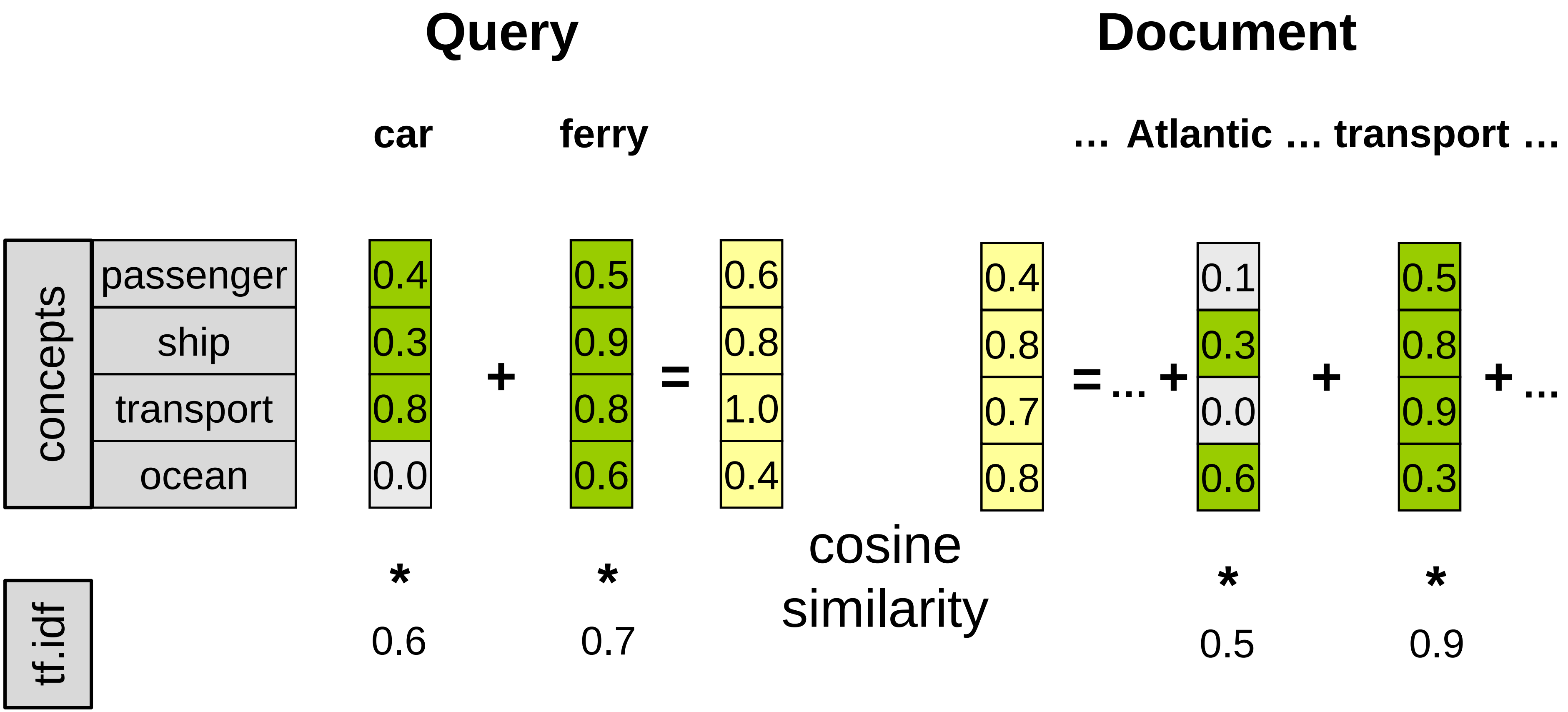
passenger	0.8	Wikipedia
SUV	0.1	
taxicab	0.9	
transport	0.8	
yellow	0.8	
passenger	0.4	Wiktionary
taxicab	1.0	
transport	0.7	
yellow	0.6	

$\overline{V}_{taxicab}$

SR Based IR Models

- **SR-Text** (Gabrilovich & Markovitch, 2007)
  - Compare concept vectors of query and document

SR-Text



Query

Document

car ferry

… Atlantic … transport …

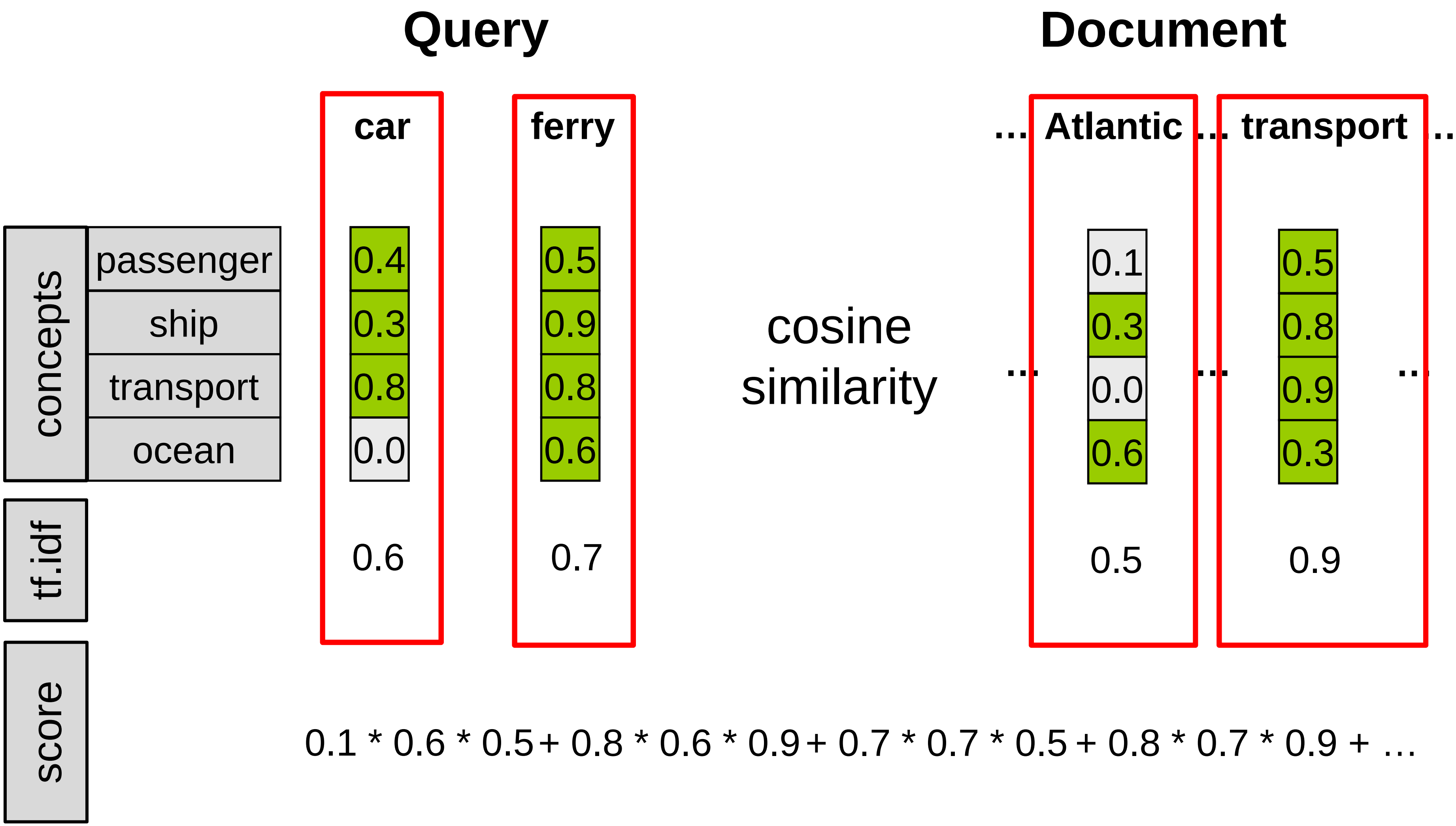
concepts	car	+	ferry	=	Query	cosine similarity	Document	= … +	Atlantic	+	transport	+ …
passenger	0.4		0.5		0.6		0.4		0.1		0.5	
ship	0.3		0.9		0.8		0.8		0.3		0.8	
transport	0.8		0.8		1.0		0.7		0.0		0.9	
ocean	0.0		0.6		0.4		0.8		0.6		0.3	
tf.idf	* 0.6		* 0.7						* 0.5		* 0.9	

SR Based IR Models

- **SR-Text** (Gabrilovich & Markovitch, 2007)
  - Compare concept vectors of query and document

- **SR-Word** (Müller & Gurevych, 2006)
  - Compare concept vectors of query and document term pairs

SR-Word



		Query			Document	
		car	ferry	…	Atlantic …	transport …
concepts	passenger	0.4	0.5		0.1	0.5
	ship	0.3	0.9		0.3	0.8
	transport	0.8	0.8	cosine similarity	0.0	0.9
	ocean	0.0	0.6		0.6	0.3
tf.idf		0.6	0.7		0.5	0.9

score

0.1 * 0.6 * 0.5+ 0.8 * 0.6 * 0.9+ 0.7 * 0.7 * 0.5 + 0.8 * 0.7 * 0.9 + …

SR-Word: Threshold

		Query		Document				
		car	ferry	…	Atlantic	…	transport	…
concepts	passenger	0.4	0.5		0.1		0.5	
	ship	0.3	0.9	…	0.3	…	0.8	…
	transport	0.8	0.8		0.1		1.0	
	ocean	0.0	0.6		0.0		0.3	
tf.idf		0.6	0.7		0.5		0.9	

threshold for SR values: 0.2

score

~~$\boxed{0.1}$ * 0.6 * 0.5~~ + 0.8 * 0.6 * 0.9 + 0.7 * 0.7 * 0.5 + 0.8 * 0.7 * 0.9 + …

SR-Word: Term Not Semantically Related

Query Document

car ferry … Atlantic … transport …

		Query: car	Query: ferry		Document: Atlantic		Document: transport	
concepts	passenger	0.4	0.5	…	0.1	…	0.5	…
	ship	0.3	0.9		0.3		0.8	
	transport	0.8	0.8		0.1		1.0	
	ocean	0.0	0.6		0.0		0.3	
tf.idf		0.6	0.7		0.5		0.9	

score

sr(car, Atlantic) < threshold **sr(car, transport) < threshold …**

decrease

SR-Word: No Match on String-Level

		Query		Document				
		car	ferry	~~car~~ …	Atlantic	…	transport	…
concepts	passenger	0.4	0.5		0.1		0.5	
	ship	0.3	0.9	…	0.3	…	0.8	…
	transport	0.8	0.8		0.1		1.0	
	ocean	0.0	0.6		0.0		0.3	
tf.idf		0.6	0.7		0.5		0.9	

score

query term not contained in document

decrease

Retrieval Models

- Semantic Relatedness
  - SR-Text (SRT)
  - SR-Word (SRW)

- Baseline: Vector Space Model
  - Apache Lucene (LUC)

- Combination using CombSUM (Fox & Shaw, 1994)
  - For each document the similarity scores of the models are normalized and added up

Outline

- IR models based on SR
- **Preprocessing of queries and documents**
- Results for monolingual runs: English, German, Russian
- Approach for bilingual IR
- Results for bilingual runs: English – German
- Additional experiments concerning efficiency
- Summary and future work

Preprocessing

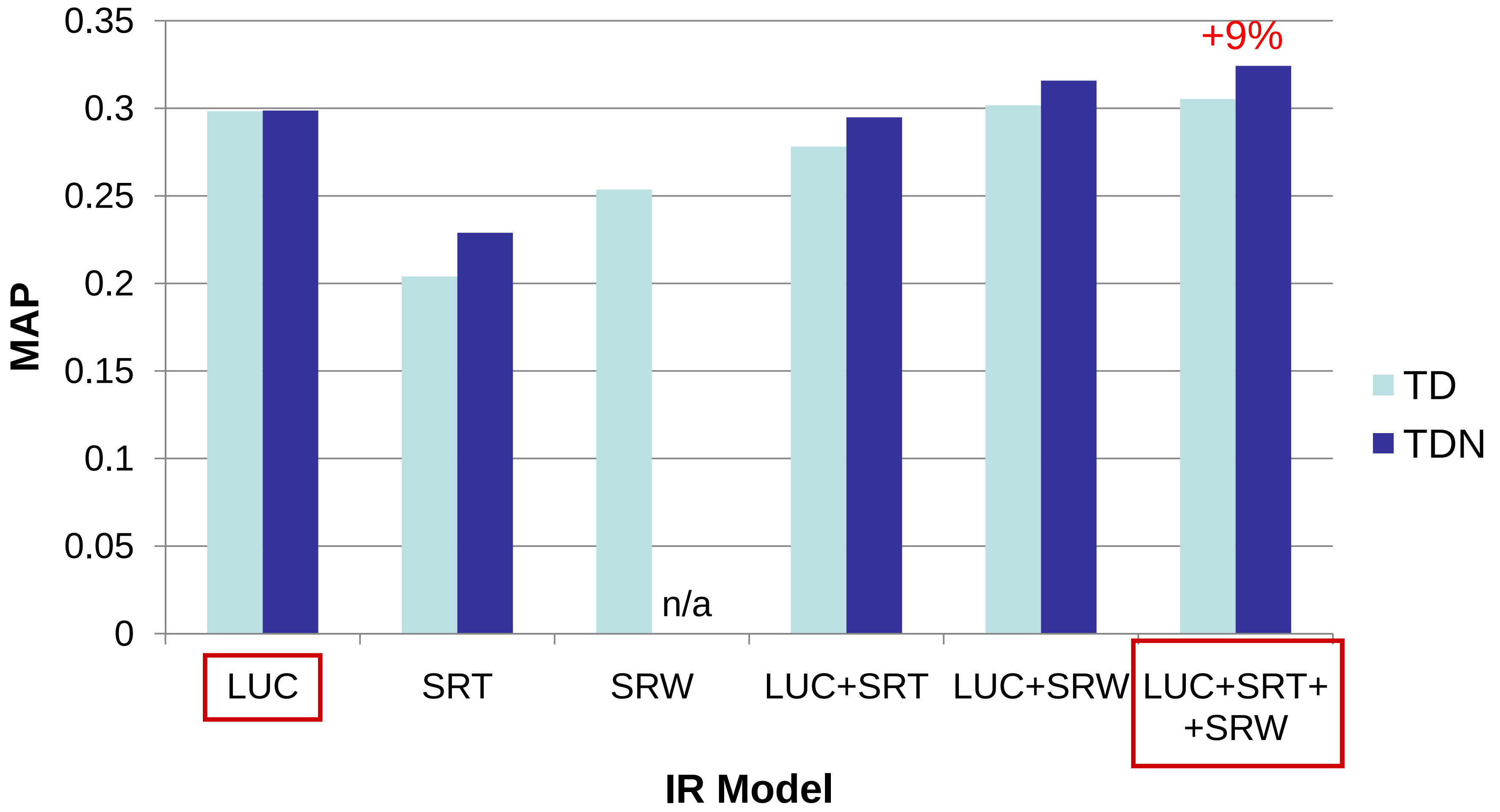
- Lemmatization
  - Using the probabilistic part-of-speech tagging system TreeTagger (Schmid, 1994)
  - English, German, Russian

- Decompounding (Langer, 1998)
  - German
  - Using compound words and their elements

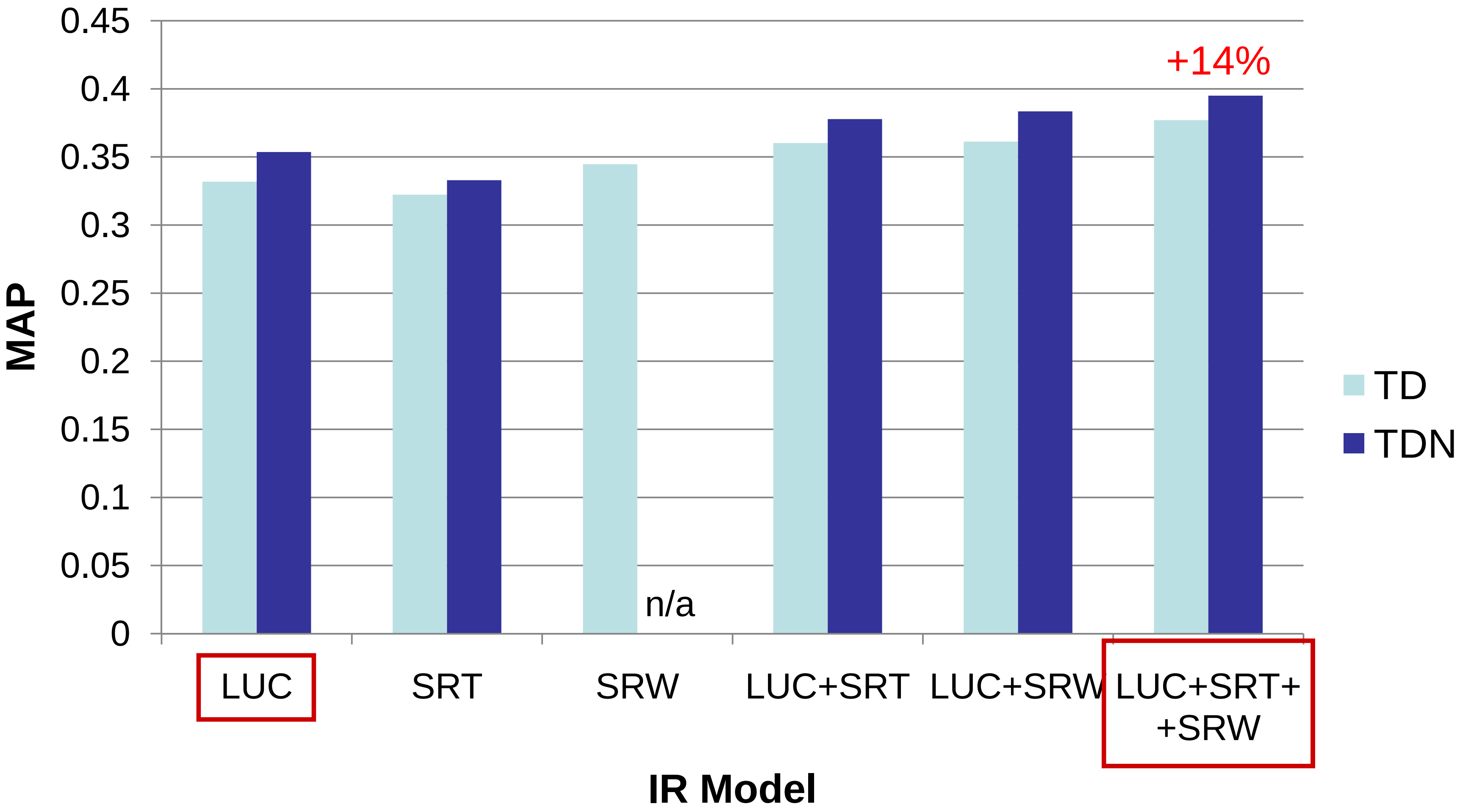
Outline

- IR models based on SR
- Preprocessing of queries and documents
- **Results for monolingual runs: English, German, Russian**
- Approach for bilingual IR
- Results for bilingual runs: English – German
- Additional experiments concerning efficiency
- Summary and future work

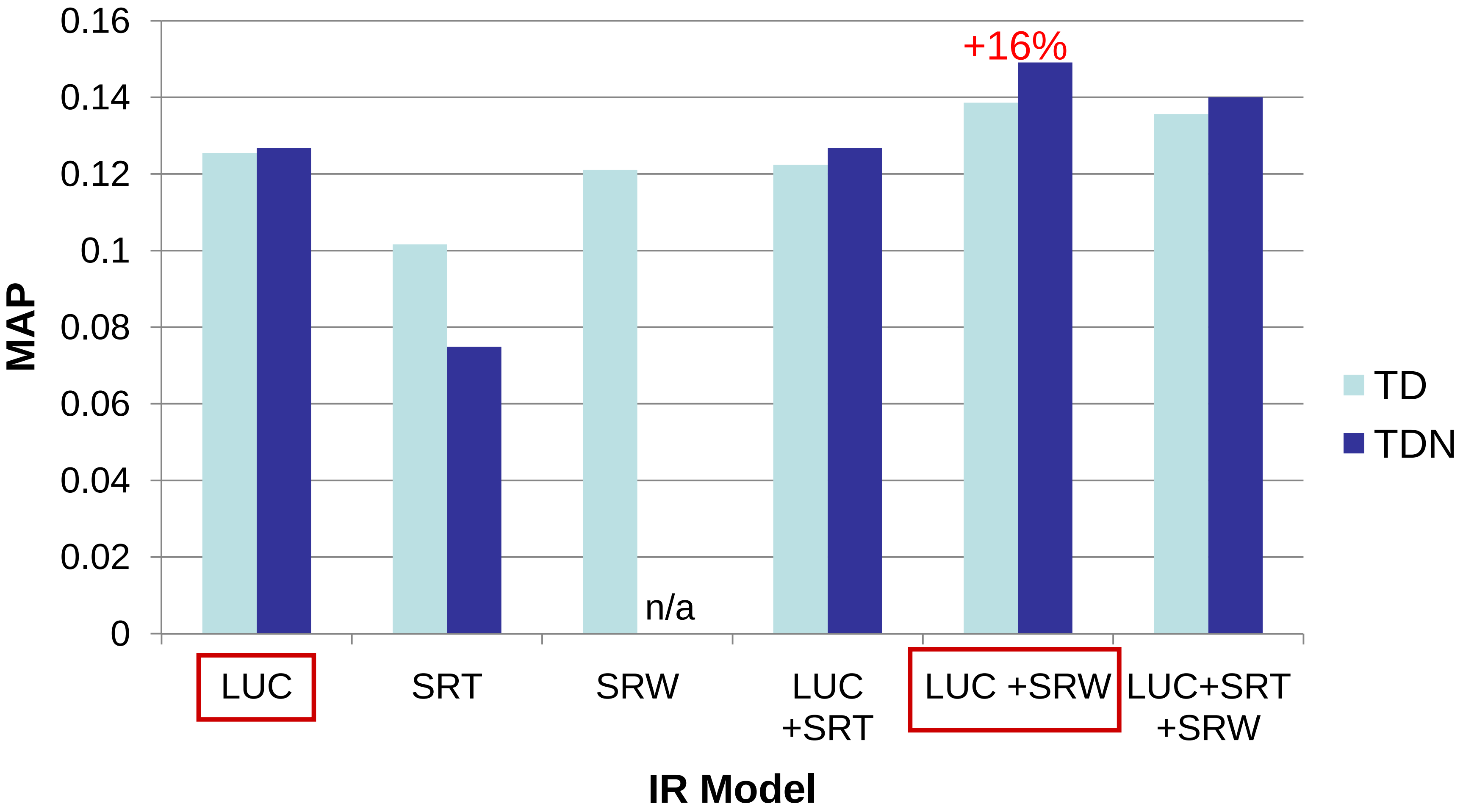
Monolingual Results: English
IR Models and Query Types



Monolingual Results: German
IR Models and Query Types



Monolingual Results: Russian
IR Models and Query Types



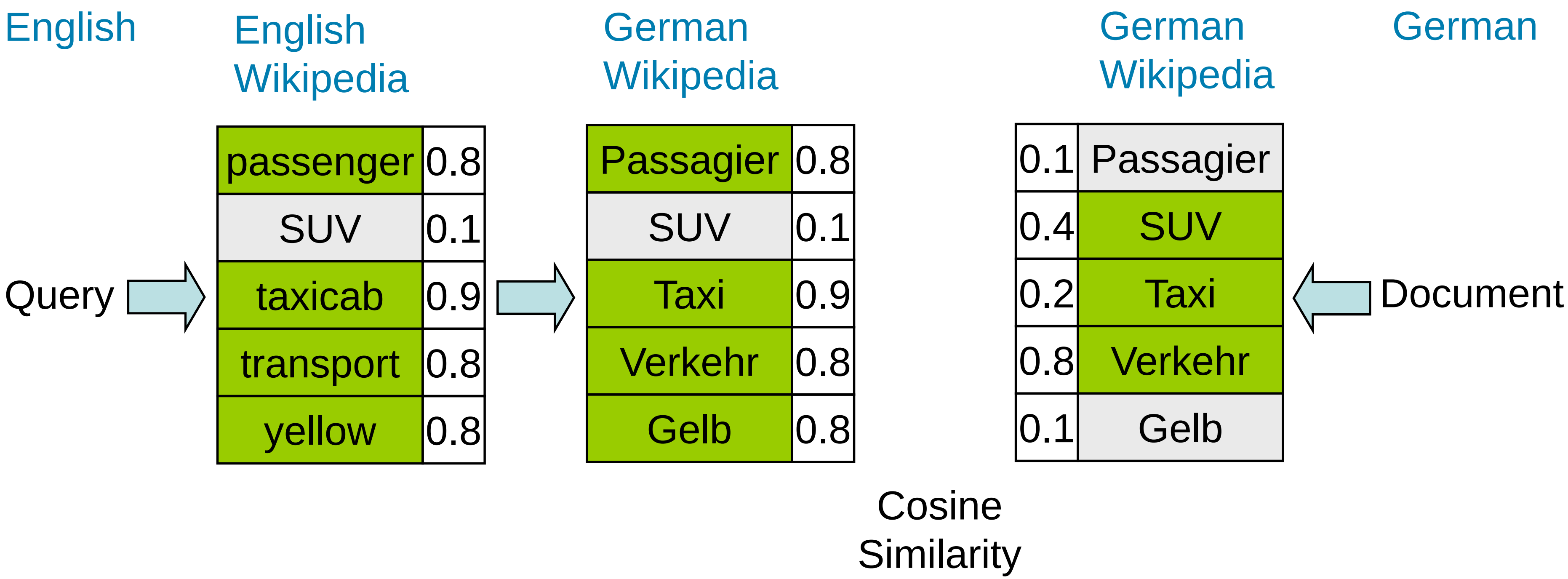
Outline

- IR models based on SR
- Preprocessing of queries and documents
- Results for monolingual runs: English, German, Russian
- **Approach for bilingual IR**
- Results for bilingual runs: English – German
- Additional experiments concerning efficiency
- Summary and future work

Crosslingual Approach

- Machine translation using **Systran**

- SR-Text: **Cross-language links** in Wikipedia



English

Query

English Wikipedia

passenger	0.8
SUV	0.1
taxicab	0.9
transport	0.8
yellow	0.8

German Wikipedia

Passagier	0.8
SUV	0.1
Taxi	0.9
Verkehr	0.8
Gelb	0.8

Cosine Similarity

German Wikipedia

0.1	Passagier
0.4	SUV
0.2	Taxi
0.8	Verkehr
0.1	Gelb

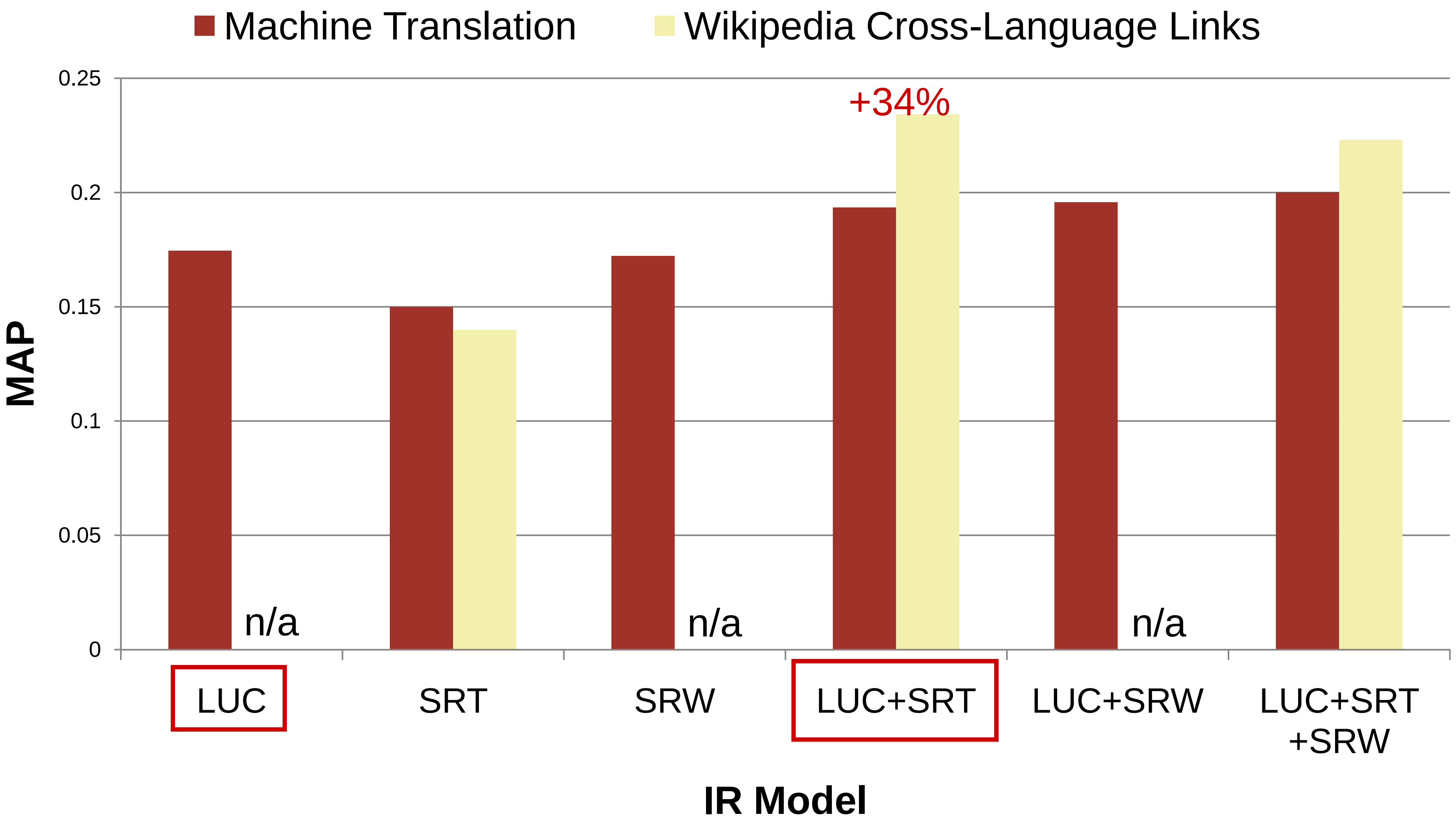
German

Document

Outline

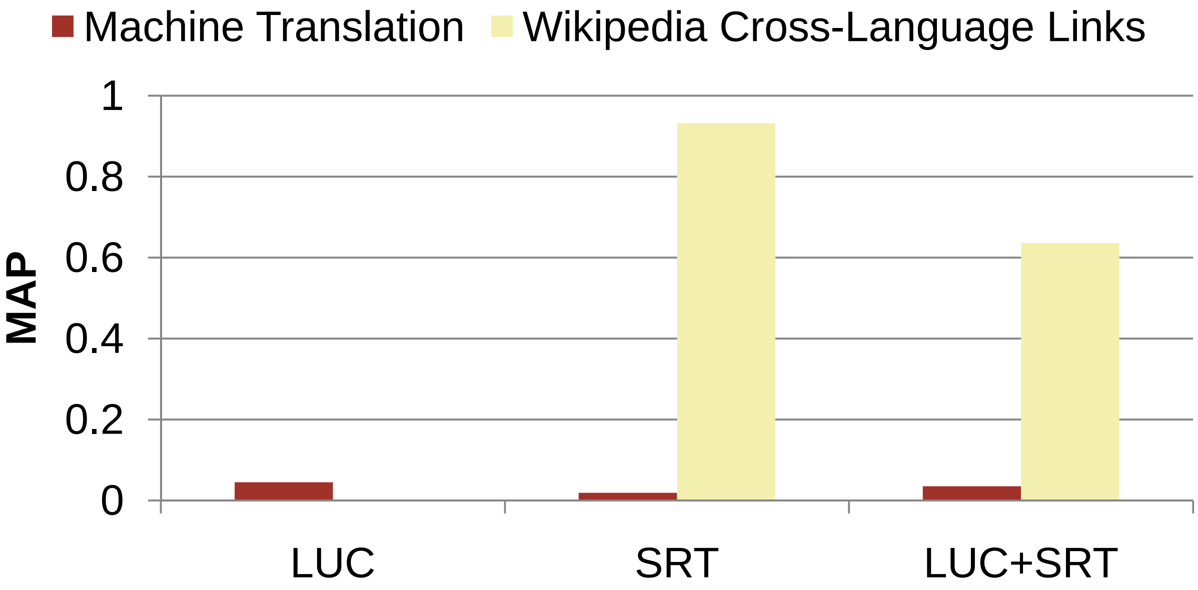
- IR models based on SR
- Preprocessing of queries and documents
- Results for monolingual runs: English, German, Russian
- Approach for bilingual IR
- **Results for bilingual runs: English – German**
- Additional experiments concerning efficiency
- Summary and future work

Bilingual Runs: English – German
IR Models and Translation Approach



Example Topic

- Topic No. 209
- English title
  - *Doping and sports*
- *German title translated by Systran*
  - *Lackieren und Sport* (engl.: *painting/lacquer and sports)*



Outline

- IR models based on SR
- Preprocessing of queries and documents
- Results for monolingual runs: English, German, Russian
- Approach for bilingual IR
- Results for bilingual runs: English – German
- **Additional experiments concerning efficiency**
- Summary and future work

Additional Experiments

- SR based models are computationally expensive

- Reducing dimensions of concept vectors in SR-Text
  - Experiments on CLEF data from previous years
  - Using only the highest 20,000 – 25,000 concepts for the concept vectors of queries and documents yields same MAP values

- Only ranking the documents retrieved by Lucene
  - Experiments on monolingual English data from this year
  - Only a fraction of query – document comparisons necessary
  - No decrease of MAP
    - SR based models might not retrieve additional relevant documents
    - But they help to rank relevant documents higher

Summary

- Introduction of **Wiktionary as new resource** for IR
- **Combination of Wikipedia and Wiktionary** as knowledge base for two SR based IR models: SR-Text and SR-Word
- **Combination of SR based models with Lucene** using CombSUM
- Monolingual:
  - Improvement of combination of models compared to baseline:
    - **English: 9%, German: 14%, Russian: 16%**
  - **SR-Word outperforms SR-Text**
- Bilingual:
  - Machine translation
  - SR-Text: **cross-language links** in Wikipedia
    - Improvement by combining Lucene and SR-Text:
      - **English – German: 34%**
- **Efficiency** of SR based models **can be improved**

Future Work

- Combination of SR based models with other IR models (Okapi BM25, …)
- Different combination methods for models and knowledge bases
- Using cross-language links in Wiktionary for cross-lingual IR

Acknowledgments

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Ubiquitous Knowledge Processing Lab

http://www.ukp.tu-darmstadt.de

KTS
KLAUS TSCHIRA STIFTUNG
GEMEINNÜTZIGE GMBH

e-learning
center of
research
excellence

Bundesministerium
für Wirtschaft
und Technologie

Photo by Jens Steingässer

VolkswagenStiftung

Deutsche
Forschungsgemeinschaft
DFG

Emmy
Noether-
Programm
Deutsche
Forschungsgemeinschaft
DFG

Wikipedia & Wiktionary API

http://www.ukp.tu-darmstadt.de/software