Application of Axiomatic Approaches to Crosslanguage Retrieval Overview of the Know-Center System for Robust WSD @ CLEF2009

Roman Kern
rkern@know-center.at

http://www.know-center.at

COMET
Competence Centers for
Excellent Technologies

Introduction

Overview

- System Overview
  - Index Types
  - Index Fields
  - Query Construction
  - Ranking Functions
- System Performance
  - Baseline Performance
  - Impact of WSD Information
  - Impact of Translation

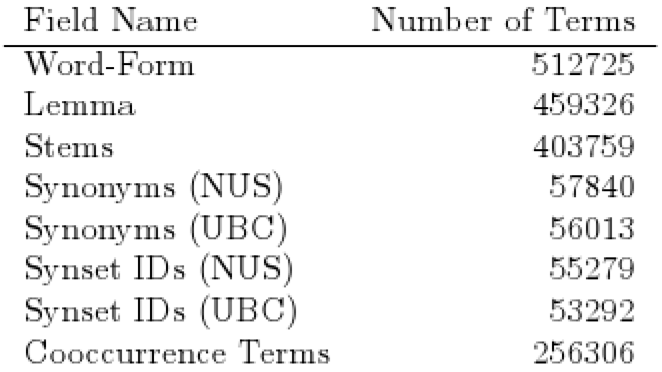
Index Types

- Document Index
  - Created using the ~170k documents
  - Contains **WSD information** (Synonyms & Synset-IDs)
- Multilingual Index
  - **Aligned** documents
  - Used for **translation** of (query) terms
  - One multilingual index per corpus

System Overview / Index Fields

Document Index

- Build using article body
  - Headline not used
- Token Fields
  - Word-Form
  - Lemma
  - Stems *(Snowball Stemmer)*
- **WSD Fields**
  - Synonyms of *top ranked* synset
  - ID of *top ranked* synset
- Co-Occurrence Field
  - Build using the stemmed terms
  - CondPMI for term-term weights



Field Name	Number of Terms
Word-Form	512725
Lemma	459326
Stems	403759
Synonyms (NUS)	57840
Synonyms (UBC)	56013
Synset IDs (NUS)	55279
Synset IDs (UBC)	53292
Cooccurrence Terms	256306

$$S_{CondPMI}(w_i, w_j) = \frac{log_2 \frac{P(w_j|w_i)}{P(w_j)}}{log_2(\frac{1}{P(w_j)})}$$

System Overview / Index Fields

Multilingual Index

- Build using multilingual corpora
- Document aligned: **Wikipedia**

  Exploit cross-lingual links between articles

- Sentence aligned: **Europarl**

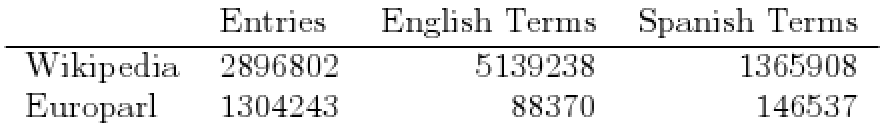
  Proceedings of the European Parliament

- Translation

  Search in *source language*

  Collect top-n results in *target language* *(n = 50)*

  Extract terms and select top-m as translation *(m = 2)*



	Entries	English Terms	Spanish Terms
Wikipedia	2896802	5139238	1365908
Europarl	1304243	88370	146537

Multilingual Query

- Pluggable weighting scheme for term translation
- Keyword Extraction

  Use the term with the highest **TFIDF weight**

$$w_i^{TFIDF} = log(\frac{N}{docFreq_i + 1} + 1) * \sum_j^D score_j$$

- Query Reconstruction

  Aggregation of **differences** between expected and observed score

$$w_i^{reconstruction} = \frac{1}{\sum_j^D |tf_{i,j} * log(\frac{N}{docFreq_i + 1} + 1) - score_j| + 1}$$

Query Construction

- Using the **Title** and **Description** part of the topics
  - Description terms did get lower weight *(0.25)*
- No blind relevance feedback
  - Only global QE methods
- Incorporate WSD information via **Query Expansion**
  - The synonyms of the top scores sense are used
  - The synset-id of the top sense
- Co-occurrence terms were also added via QE
  - Add co-occurring terms to query *(2 size of query)*
  - Co-occurrence reflects all semantic relatedness (hypernyms, meronyms, ...)

System Overview

Ranking Functions

- Pluggable retrieval function for scoring
- Default Lucene **TFIDF** boolean query
- Lucene **Disjunction Max** query
- Variant of the **BM25** weighting function

$$S_{BM25}(Q,D) = \sum_{t \in Q \cap D} \frac{tf_{t,D}}{k_1((1-b)+b*\frac{docLength_D}{averageDocLength})+tf_{t,D}} * log\frac{N-docFreq_t+0.5}{docFreq_t+0.5}$$

- **Axiomatic** retrieval function

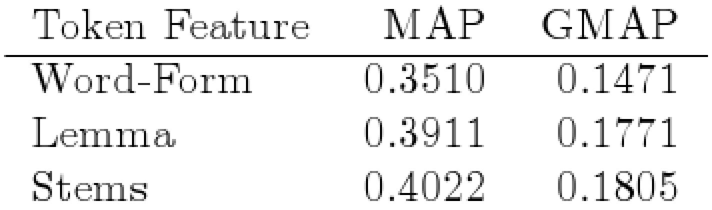
Famlily of weighting function derived using an axiomatic approach

$$S_{Axiomatic}(Q,D) = \sum_{t \in Q \cap D} (\frac{N}{docFreq_t})^{\alpha} * \frac{tf_{t,D}}{tf_{t,D}+0.5+\beta\frac{docLength_D}{averageDocLength}}$$

Baseline Performance

- Comparison of the **token features**

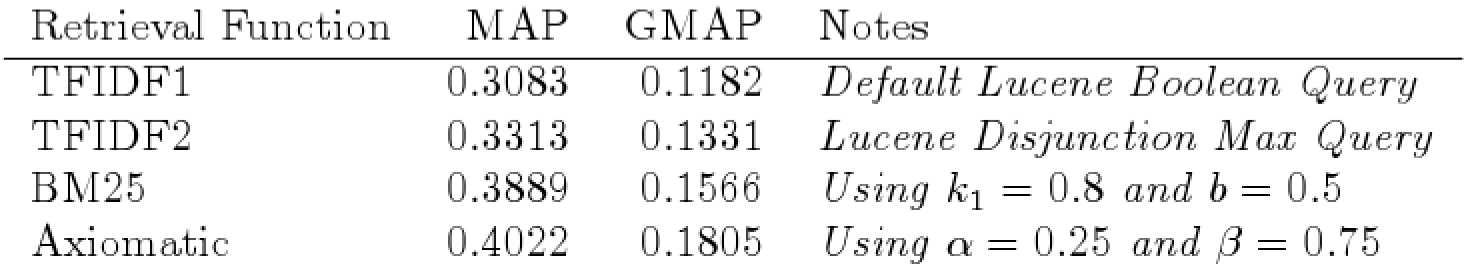
  Best performance using *stems*



Token Feature	MAP	GMAP
Word-Form	0.3510	0.1471
Lemma	0.3911	0.1771
Stems	0.4022	0.1805

- Comparison of the **retrieval functions**

  Best performance using *axiomatic approach*



Retrieval Function	MAP	GMAP	Notes
TFIDF1	0.3083	0.1182	*Default Lucene Boolean Query*
TFIDF2	0.3313	0.1331	*Lucene Disjunction Max Query*
BM25	0.3889	0.1566	*Using* $k_1 = 0.8$ *and* $b = 0.5$
Axiomatic	0.4022	0.1805	*Using* $\alpha = 0.25$ *and* $\beta = 0.75$

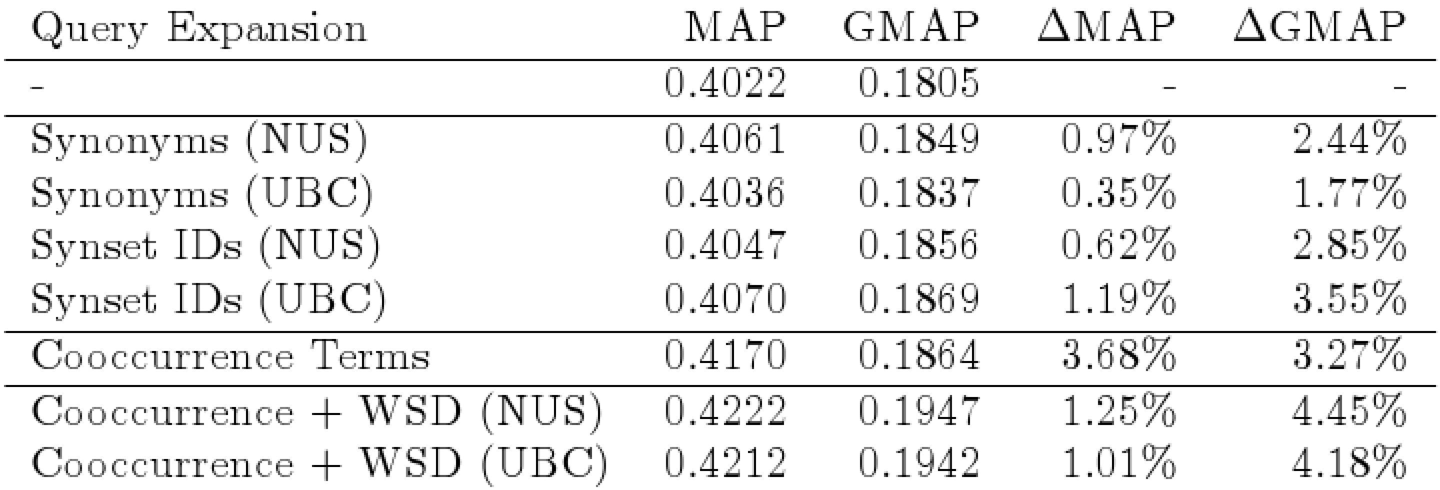
System Performance / Monolingual

Performance Impact of WSD

- Comparison of the **query expansion strategy**

  WSD information does improve the monolingual retrieval

  Query expansion using co-occurrence does out-perform pure synonym approach

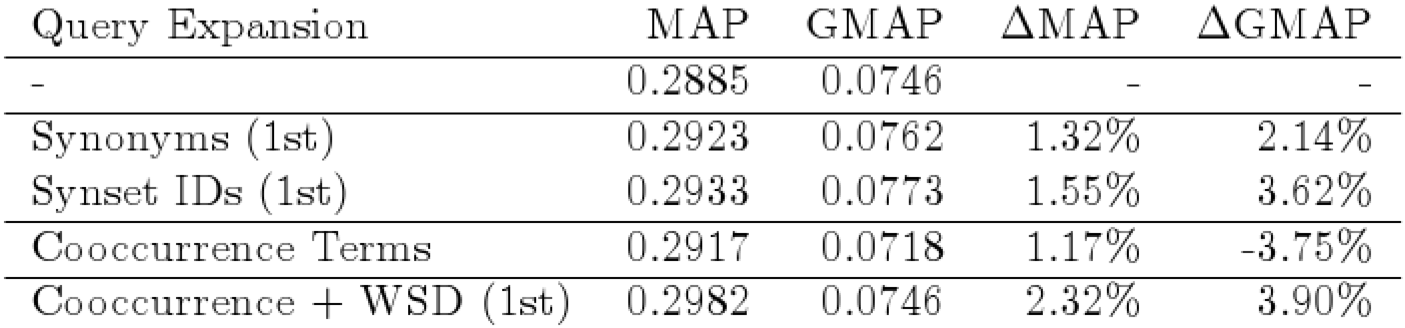


Query Expansion	MAP	GMAP	ΔMAP	ΔGMAP
-	0.4022	0.1805	-	-
Synonyms (NUS)	0.4061	0.1849	0.97%	2.44%
Synonyms (UBC)	0.4036	0.1837	0.35%	1.77%
Synset IDs (NUS)	0.4047	0.1856	0.62%	2.85%
Synset IDs (UBC)	0.4070	0.1869	1.19%	3.55%
Cooccurrence Terms	0.4170	0.1864	3.68%	3.27%
Cooccurrence + WSD (NUS)	0.4222	0.1947	1.25%	4.45%
Cooccurrence + WSD (UBC)	0.4212	0.1942	1.01%	4.18%

Performance Bilingual

- Comparison of the system with **query translation**

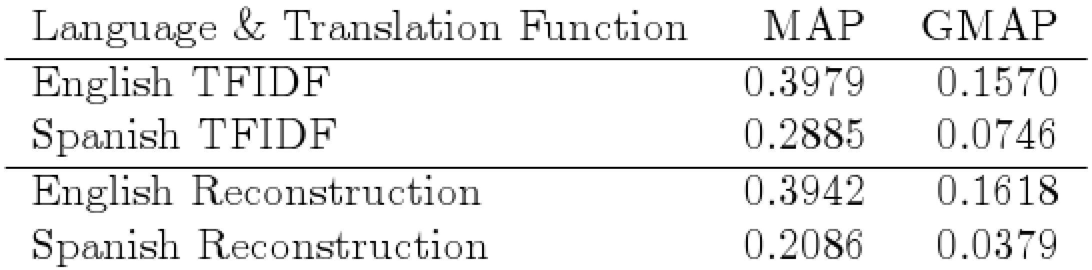
Improvements of WSD information smaller than for monolingual



Query Expansion	MAP	GMAP	ΔMAP	ΔGMAP
-	0.2885	0.0746	-	-
Synonyms (1st)	0.2923	0.0762	1.32%	2.14%
Synset IDs (1st)	0.2933	0.0773	1.55%	3.62%
Cooccurrence Terms	0.2917	0.0718	1.17%	-3.75%
Cooccurrence + WSD (1st)	0.2982	0.0746	2.32%	3.90%

- Influence of the query translation

Pronounced difference between keyword extraction for the spanish topics



Language & Translation Function	MAP	GMAP
English TFIDF	0.3979	0.1570
Spanish TFIDF	0.2885	0.0746
English Reconstruction	0.3942	0.1618
Spanish Reconstruction	0.2086	0.0379

Conclusion

Summary

- Axiomatic based retrieval model does provide robust performance
  - Even better performance than BM25
- WSD information does show improvements in the monolingual task
  - Improvement of up to 3.5% for GMAP
- WSD information does improve performance even if applied additionally to an exisisting QE technique
  - Improvements of more than 3% for MAP and GMAP
- WSD information does also increase the performance in the bilingual task
  - Improvements of WSD information smaller than for monolingual

Thank You!

Questions?

Roman Kern
Knowledge Relationship Discovery
Know-Center Graz
Inffeldgasse 21a
8020 Graz

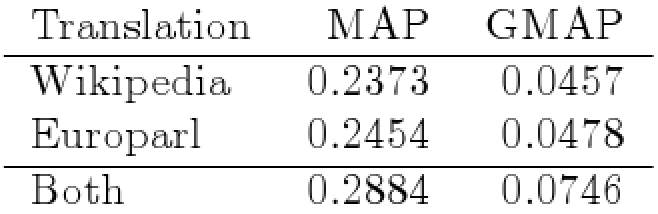
+43 316 873 66
rkern@know-center.at
www.know-center.at

System Performance / Bilingual

Impact of Translation Corpus

- Comparison of the system with a combination of the **corpus** used for translation

  Performance of Wikipedia and Europarl about the same, but combination works best



Translation	MAP	GMAP
Wikipedia	0.2373	0.0457
Europarl	0.2454	0.0478
Both	0.2884	0.0746