Exploiting Keyword Structure for Domain-Specific Retrieval

Christof Monz Jaap Kamps Maarten de Rijke

Language & Inference Technology Group
Institute for Logic, Language, and Computation
University of Amsterdam

CLEF 2002 Workshop
September 20, 2002

Motivation

Many domain-specific collections, such as the scientific collections of GIRT and Amaryllis, contain meta-information such as keywords.

One of the main goals of our participation in the GIRT and Amaryllis tasks was to experiment with the keywords used in the collections.

Our strategy for CLEF 2002 was to compute the similarity of keywords based on their occurrence in the collection, and explore whether the resulting keyword space can be used to improve retrieval effectiveness.

Keyword usage

GIRT collection

6745 keywords
755333 occurrences

5 most frequent

29561 *Bundesrepublik Deutschland*
9246 *Frau*
6133 *historische Entwicklung*
4736 *Entwicklung*
4451 *neue Bundesländer*

Repeated keywords

704 (4 in GIRT-19955040)

Amaryllis collection

1255360 keywords
1599653 occurrences
10274 occur $\geq$ 25 times

5 most frequent

20514 *Homme*
17283 *France*
7888 *Traitement*
6619 *Etude expérimentale*
5987 *Etude cas*

Repeated keywords

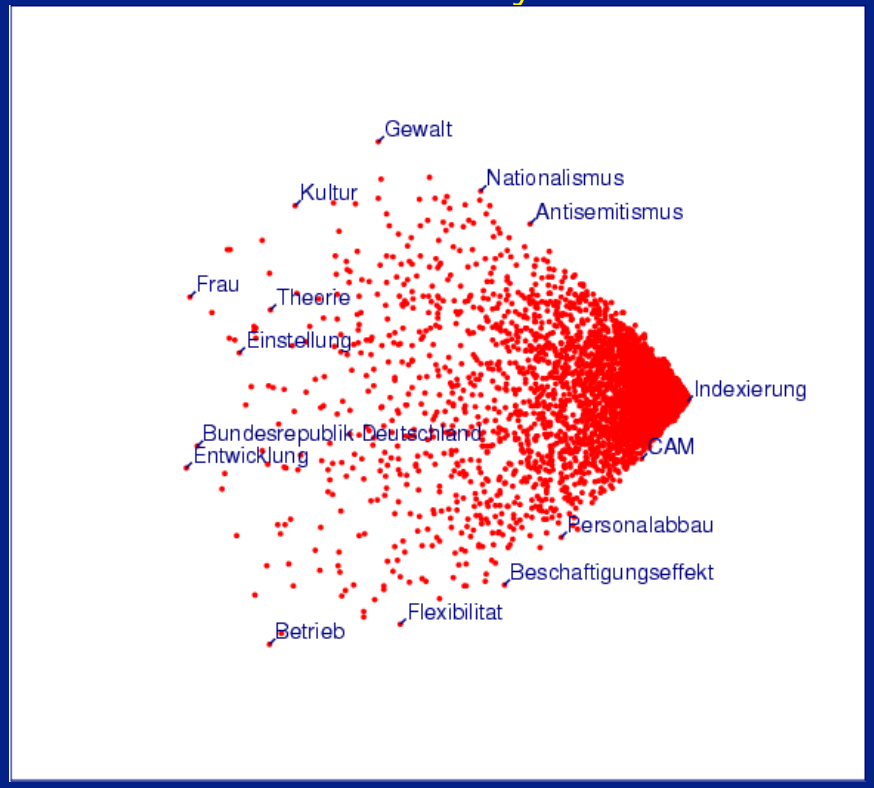
562 (9 times *France* in AM-079660)

- Many collections contain meta-information such as keywords
- Good quality dictionary/thesaurus is not always available
- We experiment with extracting the 'meaning' of keywords from their usage

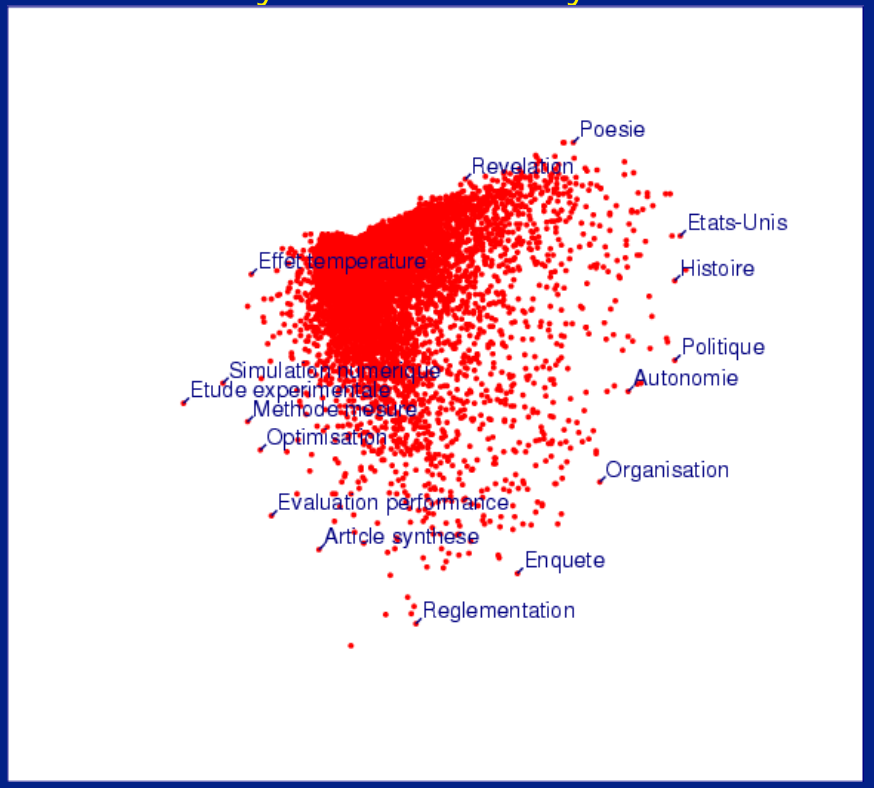
10-Dimensional Keyword Space

- Relative co-occurrence frequency keywords indicates semantic similarity
- We define a metric, and reduce the $N \times N$ matrix to the 10 principal dimensions

GIRT 6745 keywords



Amaryllis 10274 keywords

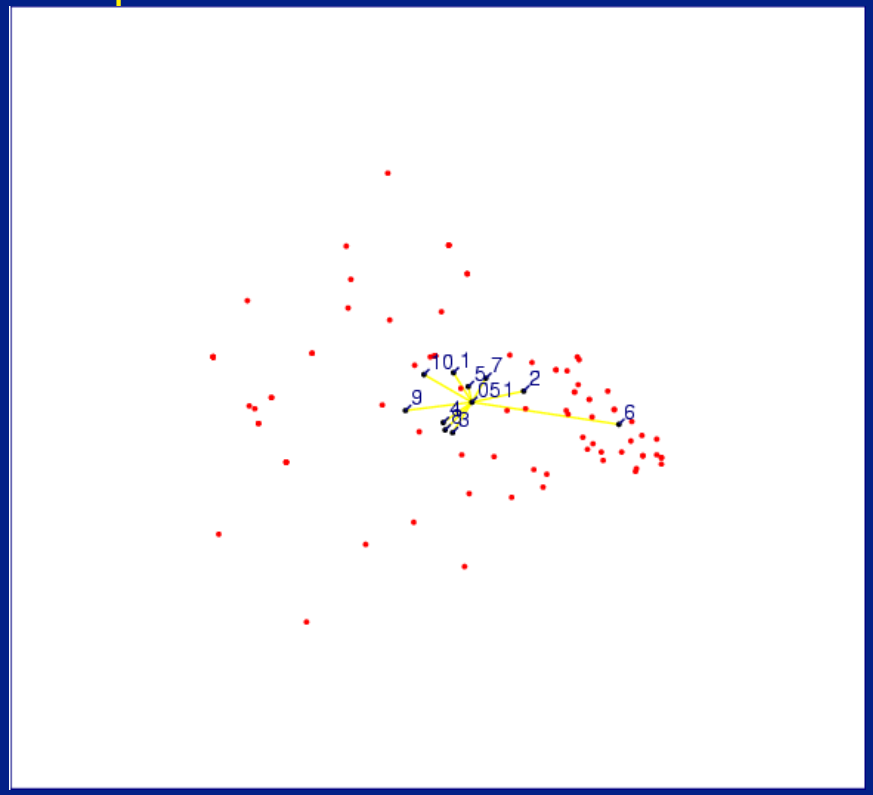


- Note that this also reveals the main research dimensions in the domains!

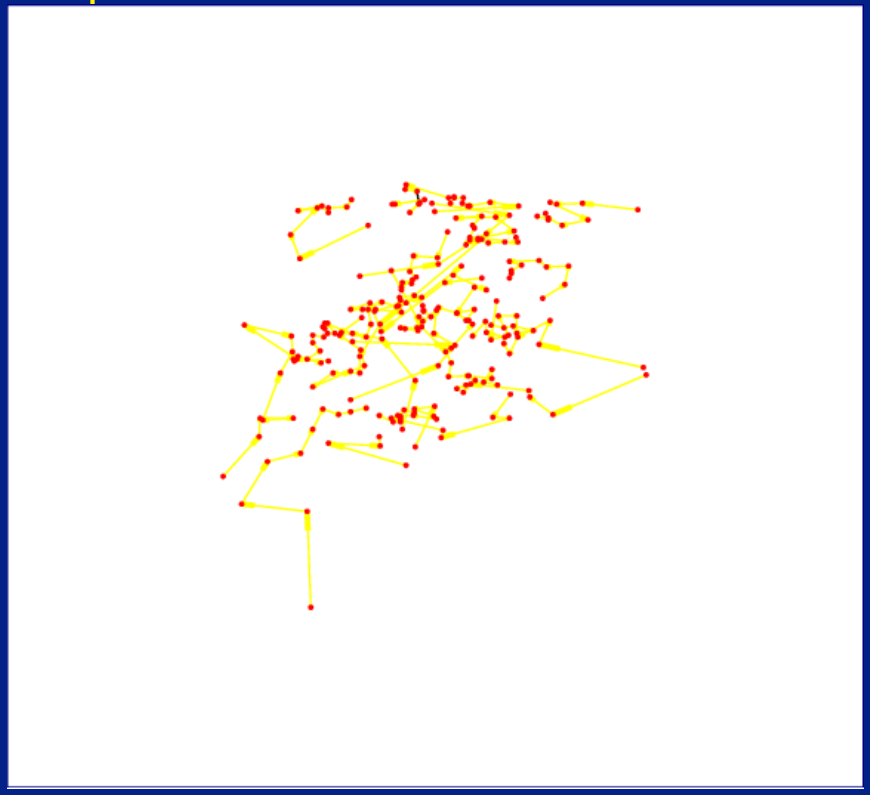
Visualizations using Keyword Space

- Document vector is the mean of keyword vectors
- Topic vector is the weighted mean of top n document vectors

Top 10 documents for GIRT-051



Topic drift for GIRT-051–GIRT075



- We use the topic vector to reranks documents and recover keywords

Keyword Recovery

GIRT topic 051

⟨title⟩ *Selbstbewusstsein von Mädchen*
⟨desc⟩ *Finde Dokumente, die über den Verlust des Selbstbewusstseins junger Mädchen während der Pubertät berichten.*

⟨title⟩ Self-confidence of girls
⟨desc⟩ Find documents which report on the loss of self-confidence of young girls during the puberty.

- Topic vector based on top 10 documents of base run

Ten closest used keywords	Ten closest global keywords	Ten densest used keywords
Selbstbewußtsein	*Erwartung*	*Erziehungsstil*
familiale Sozialisation	*Selbstbewußtsein*	*Pubertät*
Junge	*familiale Sozialisation*	*Menstruation*
Adoleszenz	*Identitätsbildung*	*körperliche Entwicklung*
Subkultur	*Identifikation*	*Selbstzerstörung*
Erziehungsstil	*Sozialisationsbedingung*	*Selbstbewußtsein*
soziale Isolation	*Junge*	*Heimerziehung*
Marginalität	*Adoleszenz*	*geistige Behinderung*
Bewußtseinsbildung	*Freundschaft*	*Grieche*
Pubertät	*Verhaltensmuster*	*Bewußtseinsbildung*

Provided versus Recovered Keywords

Amaryllis topic 1

⟨title⟩*Impact sur l'environnement des moteurs diesel* ⟨desc⟩*Pollution de l'air par des gaz d'échappement des moteurs diesel et méthodes de lutte antipollution. Emissions polluantes (NOX, SO2, CO, CO2, imbrûlés, ...) et méthodes de lutte antipollution*

⟨title⟩The impact of diesel engine on environment ⟨desc⟩Air pollution by the exhaust of gas from diesel engines and methods of controlling air pollution. Pollutant emissions (NOX, SO2, CO, CO2, unburned product, ...) and air pollution control

Monolingual, provided

Concentration et toxicité des polluants
Mécanisme de formation des polluants
Réduction de la pollution
Choix du carburant
Réglage de la combustion
Traitement des gaz d'échappement
Législation et réglementation

Monolingual, recovered

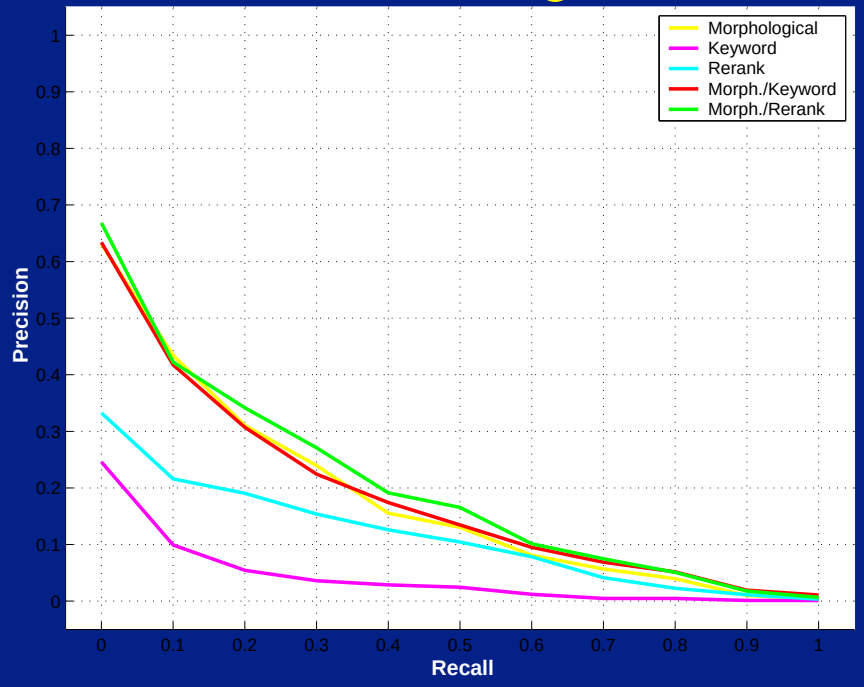
Moteur diesel
Qualité air
Azote oxyde
Norme ISO
Produit pétrolier
Lutte antipollution air
Véhicule à moteur
Gas oil
Consommation carburant
Carburant

Bilingual, recovered

Qualité air
Moteur diesel
Trafic routier urbain
Autobus
Azote oxyde
Exposition professionnelle
Véhicule à moteur
Carburant diesel
Inventaire source pollution
Carburant remplacement

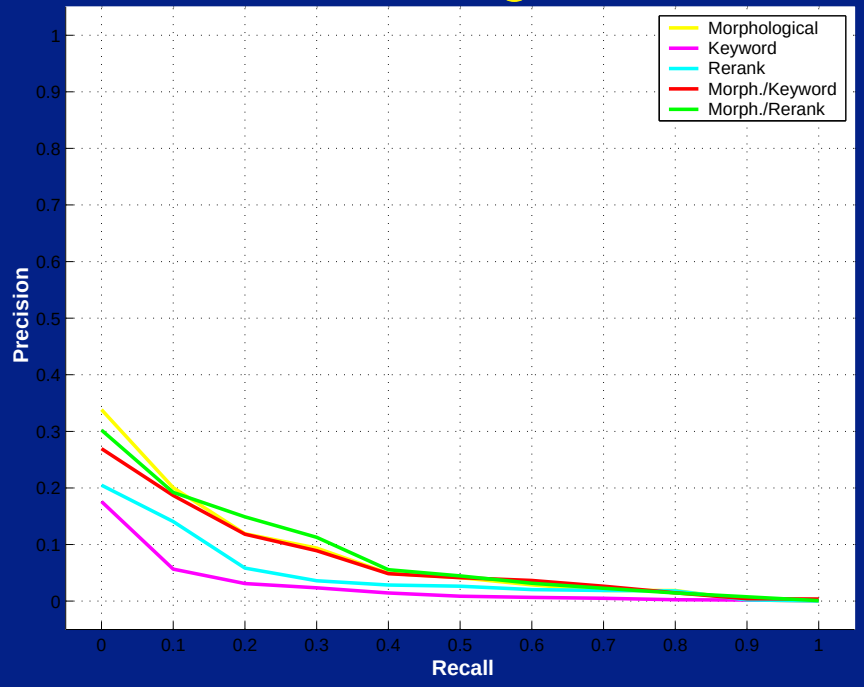
Results for GIRT

GIRT Monolingual



Morphological: 0.1639
Morph/Rerank: 0.1906 (+16.3%)

GIRT Bilingual



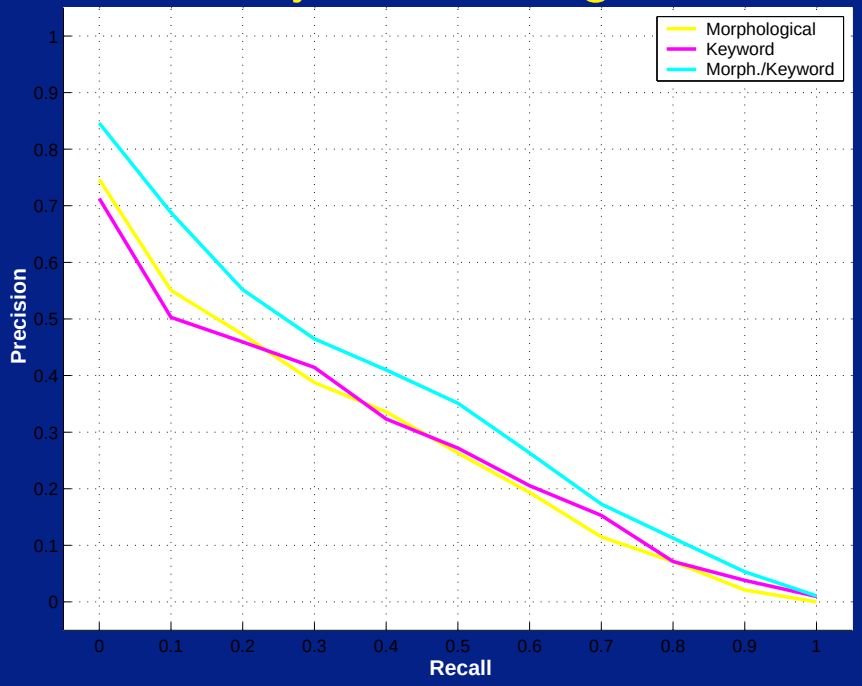
Morphological: 0.0666
Morph/Rerank: 0.0704 (+5.7%)

- Results for GIRT disappointing:

  German Base run: 0.4476 GIRT01 Base run: 0.3083 GIRT00 Base run: 0.3145
- Gain by combination consistent with GIRT01 and GIRT00

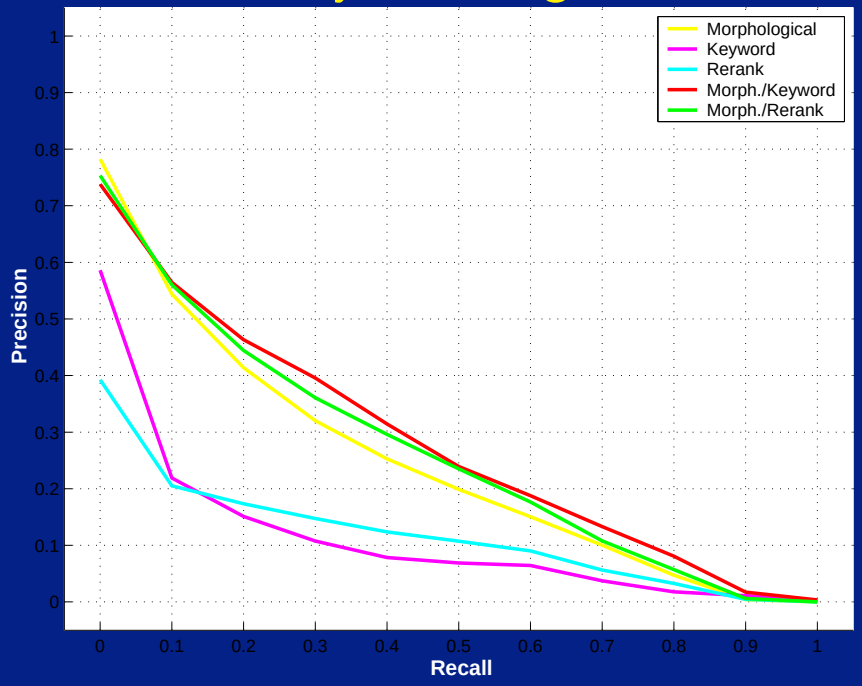
Results for Amaryllis

Amaryllis Monolingual



Morphological: 0.2681
Morph/Keyword: 0.3401 (+26.7%)

Amaryllis Bilingual



Morphological: 0.2325
Morph/Keyword: 0.2660 (+14.4%)

Recovered versus Provided Keywords

Amaryllis monolingual keyword-only runs

Provided Keywords $\longrightarrow$ 0.2684
Recovered Keywords $\longrightarrow$ 0.1120

Provided keywords score remarkably well

Amaryllis monolingual combined runs

Combination with text-only run	Both combinations improve
Recovered Keywords $\longrightarrow$ 0.2923	+9.0%
Provided Keywords $\longrightarrow$ 0.3401	+26.7%

- Used combination factor proved far from optimal
  - 0.7 Morphological / 0.3 Keyword $\longrightarrow$ 0.3401 (+26.7%)
  - 0.4 Morphological / 0.6 Keyword $\longrightarrow$ 0.4175 (+55.6%)

Effectiveness of Combining Runs

Standard explanation [Lee, SIGIR'97]:

- a high overlap between relevant documents (R_{overlap})
- low overlap between non-relevant documents (N_{overlap})

Task	Runs	R_{overlap}	N_{overlap}	Improvement	$\overline{\text{RSV}}(R_{\text{overlap}})$	$\overline{\text{RSV}}(N_{\text{overlap}})$
GIRT monolingual	Morph/Keyword	0.4493	0.1031	+2.9%	0.2506	0.1775
GIRT bilingual	Morph/Keyword	0.2984	0.0756	-6.9%	0.2643	0.1719
Amaryllis monolingual	Morph/Keyword	0.6586	0.1236	+26.7%	0.2919	0.2068
Amaryllis bilingual	Morph/Keyword	0.6506	0.1301	+14.4%	0.2803	0.2092
GIRT monolingual	Morph/Rerank	1	1	+16.3%	0.5448	0.4098
GIRT bilingual	Morph/Rerank	1	1	+5.7%	0.5330	0.3982
Amaryllis bilingual	Morph/Rerank	1	1	+9.1%	0.5641	0.4546

Lee's rationale fails for rerank runs!
There are other reasons for effective combination of runs
A candidate is average (normalized) RSV

Conclusions

We experimented with exploiting meta-information such as keywords:

- Derive 'meaning' from usage in collection
- Reduce matrix to 10 principal dimensions
- This allows for visualizing documents, topics, topic drift, ...

The resulting keyword space was used for keyword recovery

- Informal evaluation of recovered keywords suggests viability
- This can be used to assign keywords to new documents

Experience on GIRT and Amaryllis

- Recovered keywords (and rerankings) score worse than base runs
- The combined runs do improve retrieval effectiveness