Evaluation of IR Strategies for Polish

Akasereh Mitra, Malak Piotr and Pawłowski Adam

Research presentation

The collection

- 1,093,705 documents (meta-data of CH objects)
- average size of a document: 35 terms
- description schema:
  - Dublin Core
  - Qualified Dublin Core
  - Europeana Semantic Elements

Topics

- 50 topics, 2.82 token per topic av.
- Polish topics with English translation
- chronological topics
  - set time frames,
  - particular period
- named entities topics
  - personal names,
  - geographical names,
  - historical names
- general entities topics
  - religion or beliefs,
  - social groups or functions.

Indexing strategies

- automatic indexing and manually enriched topics,
- light stemming, rules based – mostly nouns, the most efficient if applied for tokens longer than 6 characters,
- weighting scheme – OKAPI (BM25), Z-scored data fusion, tf.idf, DFR-I(ne)B2

Evaluation

- MAP, P@5, P@10, p-value, GMAP, MFRS
- main – MAP for first 1000 matches

Automatic and enriched runs

- two manual enriched topics sets, emulating:
  - educated users,
  - experts in the field.
- both with light stemmer, and without ls
- one baseline run
- ALL enriched runs gave worse results than baseline one:
  - more keywords (2.82 in baseline to 6.1 educated, and 9.8 expert)
  - narrower queries by experts,

Evaluation of automatic runs

N	Run	Parameter Setting	MAP	P@10
1	Torun_Auto	tf.idf, stopwords rem., light stem., Boolean	0.348	
2	UniNE_Fusion	data fusion (Okapi: light stem., trunc-5)	0.343	0.614
3	UniNE_DFR	DFR-I(ne)B2, light stem., stopwords rem.	0.331	0.568
4	UniNE_PRF	data fusion, PRF (Rocchio, 5 docs, 10 terms)	0.258 †	0.494
5	UniNE_Baseline	*tf.idf* (cosine), no stemming, stopwords rem.	0.257 †	0.492
6	UniNEGramPRF	data fusion, 5-gram, PRF	0.220 †	0.472
	Baseline run	Okapi, no stemming, stopwords removing	0.314	0.520

Name	Parameter Setting	MAP	*% of change in MAP*	P@10
PLTO1EduLS	Educated, light stemmer	0.2774	−11.66%	0.454
PLTO1EduNO	Educated, no stemmer	0.2724	−13.25%	0.460
PLTO2HighLS	High, light stemmer	0.2709	−14.33%	0.528
PLTO2HighNO	High, no stemmer	0.2690	−13.73%	0.528
PLWR2Exp	Experts (no stemming)	0.1795 †	−42.83%	0.378
PLWR1Edu	Educated (no stemming)	0.1529 †	−51.31%	0.350
PLWR3Stu	Students (no stemming)	0.1279 †	−59.27%	0.268
Base Line	Basic (no stemming)	0.3140	n/a	0.552

Conclusions

- conjunction of keywords for indexing,
- light stemming increases matching performance,
- recognition of personal names (e.g. Jaroslaw – a masculine name, and a city in Poland),
- CH objects even with old spelling, are indexable because of contemporary terms in meta fields
- no additional dictionaries old – modern language are necessary.

Contact: Piotr Malak, Nicolaus Copernicus University, piomk@uni.torun.pl

This research was supported in part by the Sciex-NMS ,under Grant **POL 11.219** *Information Retrieval and Text Categorization for Polish*.