A comparison of off-the-shelf IR engines for question answering

Jörg Tiedemann
Alfa-Informatica, University of Groningen, The Netherlands
tiedeman@let.rug.nl

ABSTRACT

Common question answering (QA) systems are based on the extraction of answers from large document collections. The task of the IR component in QA is to retrieve relevant segments in order to reduce the search space. The performance (especially in terms of recall) of this component is crucial for such QA systems. We compared seven off-the-shelf IR engines using the test set from the CLEF 2003 competition on Dutch question answering.

Open source IR engines

Amberfish: http://www.etymon.com/tr.html
GPL, C/C++, plain text, semi-structured/XML (with nested fields), wild-card search, phrase search, boolean queries, relevance ranking

Lucene: http://jakarta.apache.org/lucene/docs/index.html
Apache License, Java, plain/semi-structured documents, snowball stemmers, phrase search, boolean queries, relevance ranking

Managing Gigabytes (MG): http://www.cs.mu.oz.au/mg/
GPL, C, plain text, images, boolean or ranked queries

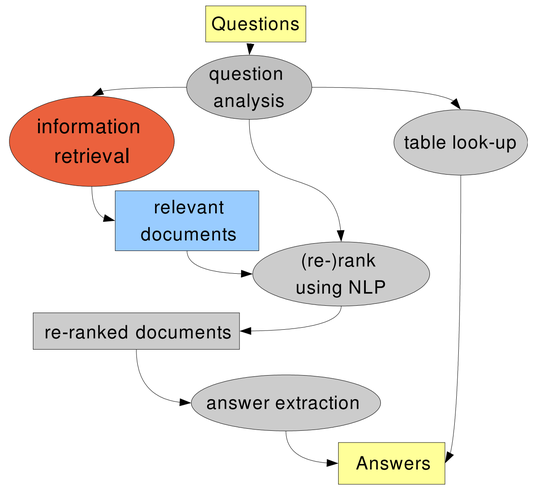
Swish-e: http://swish-e.org/
GPL, C, plain/semi-structured documents, snowball stemmers, wild card search, phrase search, fuzzy search (soundex, metaphone), flexible configuration (input/output, tokenisation etc), boolean queries, relevance ranking, Perl bindings

Xapian: http://www.xapian.org/
GPL, C++, plain text, snowball stemmers, phrase search, proximity search, relevance feedback, wide range of boolean operators, relevance ranking, Perl/SWIG bindings

Zebra: http://www.indexdata.dk/zebra/
GPL, C, structured (XML), phrase search, boolean queries, relevance ranking, wild-card search, Z39.50 protocoll, client-server implementaton

Zettair: http://www.seg.rmit.edu.au/zettair/
BSD-style license, C, plain, semi-structured (TREC), phrase search, boolean queries, relevance ranking, summary function

Joost - QA with dependency relations



IR results (CLEF 2003 data, 200 retrieved documents)

	documents		paragraphs		sentences	
MRR (in %)	doc	answer	doc	answer	doc	answer
Swish-e	26.02	54.01	28.62	43.52	23.85	32.87
Zettair	32.10	52.69	29.90	42.09	28.32	31.04
Xapian	28.25	50.49	30.11	41.41	25.14	28.90
Zebra	26.50	45.06	27.79	37.53	25.47	30.67
Lucene	29.74	47.87	30.14	36.48	27.82	29.61
Amberfish	21.05	44.31	20.67	28.05	21.15	23.06
MG	20.86	39.98	20.98	22.53	21.18	15.44

... number of paragraphs required to obtain >= x% recall

	30	40	50	60	65	70	75	80	85	90
Zettair:	1	2	4	8	13	20	34	87	-	-
Swish-e:	1	2	4	7	12	19	35	112	-	-
Lucene:	2	3	5	10	15	24	43	114	-	-
Xapian:	1	2	4	11	17	37	71	-	-	-
MG:	5	8	14	25	36	57	99	-	-	-
Amberfish:	3	6	11	23	32	57	128	-	-	-
Zebra:	2	3	6	18	36	79	191	-	-	-

Evaluation methodology

Performance is measured in terms of mean reciprocal ranks (MRR).

$$MRR = \frac{1}{x}\sum_{x}\frac{1}{rank(first\_answer)}$$

Two types of scores are distinguished: document MRR and answer MRR.

doc MRR: mean reciprocal rank of relevant documents retrieved; i.e. documents listed in the gold standard

answer MRR: mean reciprocal rank of relevant answers retrieved, i.e. documents which include the answer string

Information retrieval and Joost

The CLEF corpus:
188,651 documents
1,101,790 paragraphs
4,039,614 sentences
76,692,515 words

- paragraph/sentence level index
- retrieval of 200 paragraphs/ sentences per question
- evaluation using MRR for top 5 answers (sentences)

CLEF 2003 (Dutch):

- 450 questions
- 370 with answers

Results

MRR (in %)	paragraphs	sentences
Zettair:	54.4	51.9
Lucene:	53.9	50.6
MG:	45.3	40.4
Swish-e:	37.9	44.9

Conclusions

- QA may gain a lot from appropriate IR
- there is large performance differences between open-source IR engines
- IR performance is not (always) correlated to QA performance

Future Work

- NLP in IR (compound analysis, dependency relations, multi-word-units/phrases)
- IR voting
  - different IR engines
  - different index types
- parameter optimisation