Experiments with Semantic-flavored Query Reformulation of Geo-Temporal Queries

Nuno Cardoso[1] and Mário J. Silva[2]
[1]University of Lisbon Faculty of Sciences, LaSIGE, Lisbon, Portugal
and SINTEF Language Technologies Group, SINTEF CT, Oslo, Norway
[2]University of Lisbon Faculty of Sciences, LaSIGE, Lisbon. Portugal
ncardoso@xldb.di.fc.ul.pt mjs@di.fc.ul.pt

Motivation

Simple queries work well with simple IR systems (term-match based document retrieval).

Query expansion (QE) helps...
More terms → matching odds increased → better retrieval results
… but sometimes not.
Bad selection of terms → drift from initial topic → noisy results

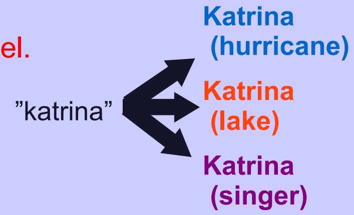
Why don't we understand what the user want, instead of retrieving what the user said?

Why don't we reason answers instead of guess terms? Is there a better approach for elaborated queries with geographic and temporal scopes?

Queries have ***entities***, and entities have semantic information.

Statistics-based QE works at term level.

Reasoning-based QE requires working at entity level, where its semantic role is ***grounded***.


"katrina"
Katrina (hurricane)
Katrina (lake)
Katrina (singer)

Objectives

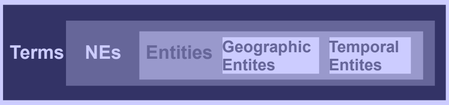
- ***Build*** a semantically-flavored query reformulation (SQR) approach, using external knowledge resources and reasoning approaches to reformulate queries at entity level.
- ***Evaluate*** how suitable is a SQR approach on retrieving documents for geographically-challenging queries.

System overview

1. Detect and ground entities in user queries and in the ***whole*** document collection

requires a named entity recognition (NER) software.

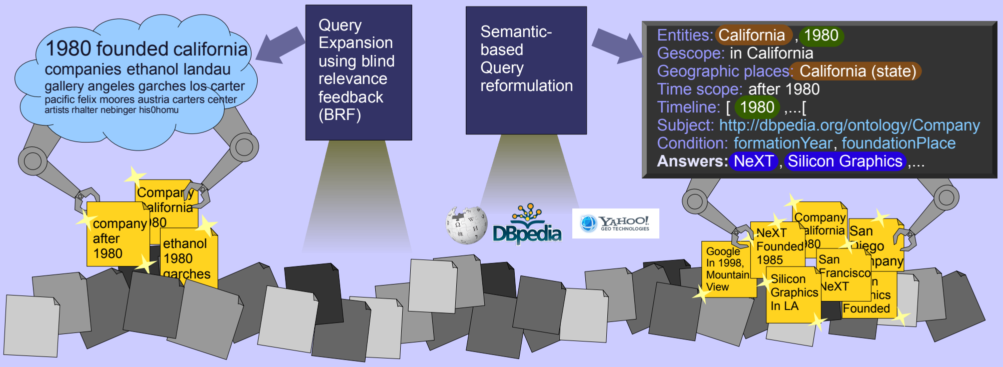
2. Use external knowledge bases (Wikipedia, DBpedia, geographic ontologies) to access more information about entities.



3. Index terms and semantic information (NEs, entities, places and time expressions)

4. Extend a retrieval engine to cope with term / semantic indexes, reformulate queries to use against those indexes

"Companies founded in California after 1980"



Entities: California , 1980
Gescope: in California
Geographic places: California (state)
Time scope: after 1980
Timeline: [1980 ,...[
Subject: http://dbpedia.org/ontology/Company
Condition: formationYear, foundationPlace
Answers: NeXT , Silicon Graphics ,...

Query Parsing

Initial query

"Where and when did Astrid Lindgren die ?"

Question type:
Where ,When
Expected answer types: PLACE, TIME

NE:Person
Entity:
http://dbpedia.org/resource/Astrid_Lindgren

Property:
http://dbpedia.org/ontology/deathPlace
http://dbpedia.org /ontology/deathDate

About: Astrid Lindgren

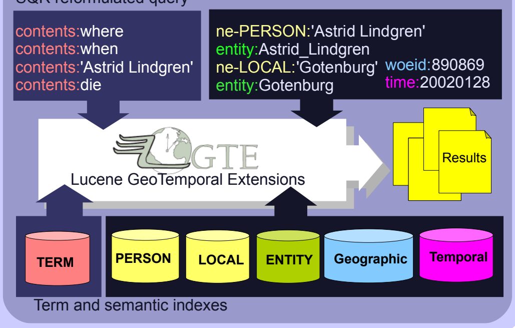
SPARQL query to DBpedia:

```
SELECT ?place, ?date where {
  dbpedia:Astrid_Lindgren
     dbpedia-owl:deathPlace ?place.
  dbpedia:Astrid_Lindgren
     dbpedia-owl:deathDate ?date.
}
```

Place: http://dbpedia.org/resource/Stockholm
Date: 2002-01-28

Document retrieval

SQR reformulated query



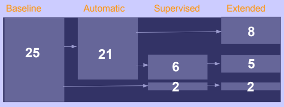
contents:where
contents:when
contents:'Astrid Lindgren'
contents:die

ne-PERSON:'Astrid Lindgren'
entity:Astrid_Lindgren
ne-LOCAL:'Gotenburg'
entity:Gotenburg
woeid:890869
time:20020128

Term and semantic indexes

Experiments and results

1. Baseline run, plain terms with no expansion
2. Automatic run, with DBpedia ontology lookup
3. Supervised run, with DBpedia ontology lookup
4. Extended run, with DBpedia abstract entities



NYT 2002-2005 Collection

Nr of **documents**	315.371
Nr of **NEs**	17.952.142
Nr of **classifications** assigned for **NEs**	18.364.572
Nr of classifications grounded to **entities**	3.344.235
Nr of classifications grounded do a **place**	588.621
Nr of docs with non-empty **GeoSignature**	202.624 (64%)
Nr of docs with non-empty **TimeSignature**	70.403 (22%)

Run	mean AP	mean Q	nDCG
2. Automatic	**0.3354**	0.3584	0.5705
1. Baseline	0.3301	**0.5701**	0.5701
3. Supervised	0.3255	0.3482	0.5593
4. Extended	0.2978	0.3205	0.5325

Baselines performed well, subjects were much more Important than geoscopes or timescopes

references to Astrid Lindgren only about her death...

No control over term:semantic index weights → recipe for disaster

more semantic information → more indexes on retrieval

summing multiple indexes from BM25 unbalances retrieval (best term:semantic ratios around 5:1)

Conclusions

1. Semantic query reformulation can achieve good retrieval performances for geographic-flavoured queries
2. Reasoning answers to add entities is hard, but grounding entities and detecting their roles is easier and very important for document ranking
3. Mixing term and semantic indexes must be done carefully: untuned index weight rations bias retrieval

We would like to thank Jorge Machado, for the support and development of LGTE, Diana Santos for comments in the paper, and SINTEF ICT. This work is supported by FCT for its LASIGE Multi-annual support, GREASE-II project (grant PTDC/EIA/73614/2006) and a PhD scholarship grant SFRH/BD/45480/2008, and by the Portuguese Government, the European Union (FEDER and FSE) through the Linguateca project, under contract ref. POSC/339/1.3/C/NAC, UMIC and FCCN.