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

Nuno Cardoso¹ and Mário J. Silva²

¹ Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal and SINTEF Natural Language Technologies Group, SINTEF ICT, Oslo, Norway
² Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal
ncardoso@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt

NTCIR-8 GeoTemporal task – 15-18th June, 2010 - Tokyo, Japan
Overview

1. PhD motivation
2. Objectives
3. System overview
4. Experiments and results
5. Lessons learned and conclusions
PhD motivation

- Simple queries work well with simple IR systems (term-match based document retrieval)
- What about complex queries?
- Current query expansion (QE) methods help... More terms → matching odds increased → better retrieval results
- ... but sometimes not. Bad selection of terms → drift from initial topic → noisy results
Most 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.

PhD motivation (cont.)

- Katrina (hurricane)
- Katrina (lake)
- Katrina (singer)

"katrina"
PhD motivation (cont.)

- Why don't we try to understand what the user wants, instead of retrieving what the user said?
- Why don't we reason to get answers instead of guessing terms?
- Is there a better approach for elaborated queries, such as queries with concrete geographic and temporal scopes?
Statistics-based Query Expansion

“Companies founded in California after 1980”

Query Expansion using blind relevance feedback (BRF)

1980 founded california companies ethanol landau gallery angeles garches los carter pacific felix moores austria carters center artists rhalter nebinger his0homu

terms in cloud obtained with LucQE using the NYT collection
Semantic-based Query Reformulation

“Companies founded in California after 1980”

Semantic-based Query reformulation

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, ...
PhD 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.

That's where NTCIR GeoTemporal task comes in...
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.
System overview

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

Query Parsing example

“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
rdf:label – “Astrid Lindgren”@pt
“アストリッド・リンドグレーン”@jp

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

About: Astrid Lindgren
An Entity of Type: person, from Named Graph:
http://dbpedia.org, within Data Space: dbpedia.org

SELECT ?place, ?date where {
  dbpedia:Astrid_Lindgren dbpedia-owl:deathDate?date .
}
Question type: none (→ Which)
Expected answer type: “japanese animators”

SPARQL query to DBpedia:

```sparql
SELECT ?s WHERE {
        ?category skos:broader dbpedia:Category:Japanese_animators
    }
    ?s dbpedia-owl:birthPlace dbpedia:Tokyo .
}
```
Document retrieval example

SQR reformulated query

Terms

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

Semantic information

ne-PERSON:'Astrid Lindgren'
entity:Astrid_Lindgren

ne-LOCAL:'Gotenburg'
entity:Gotenburg

woeid:890869
time:20020128

Lucene GeoTemporal Extensions

Term index

TERM

+ PERSON

LOCAL

ENTITY

Geographic

Temporal

Semantic indexes

Results

Results

Results
GeoTime experiments (EN only)

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

Queries accumulate more semantic information from 1 to 4.
Query reformulation example


Added in Baseline run
Added in Automatic run
Added in Supervised run
Added in Extended run
### NYT 2002-2005 collectionn (EN)

<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr of <strong>documents</strong></td>
<td>315,371</td>
</tr>
<tr>
<td>Nr of <strong>NEs</strong></td>
<td>17,952,142</td>
</tr>
<tr>
<td>Nr of <strong>classifications</strong> assigned for <strong>NEs</strong></td>
<td>18,364,572</td>
</tr>
<tr>
<td>Nr of classifications grounded to <strong>entities</strong></td>
<td>3,344,235</td>
</tr>
<tr>
<td>Nr of classifications grounded do a <strong>place</strong></td>
<td>588,621</td>
</tr>
<tr>
<td>Nr of docs with <strong>geographic places</strong></td>
<td>202,624 (64%)</td>
</tr>
<tr>
<td>Nr of docs with <strong>temporal expressions</strong></td>
<td>70,403 (22%)</td>
</tr>
</tbody>
</table>
## Official results

<table>
<thead>
<tr>
<th>Run</th>
<th>mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>0.3301</td>
</tr>
<tr>
<td>2. Automatic</td>
<td>0.3354</td>
</tr>
<tr>
<td>3. Supervised</td>
<td>0.3255</td>
</tr>
<tr>
<td>4. Extended</td>
<td>0.2978</td>
</tr>
</tbody>
</table>

GeoTime best: 0.4158

- Only topic title
- No entity index at the time
- No stemming, 1:1 term:semantic index weight
Post-hoc experiments

- Prefer entity index to NE index
- Stemming, different term:semantic index weights
- Compare/combine BRF and SQR

1. **Baseline** run, term index, no expansion
2. **BRF** run, term index, BRF expansion
3. **SQR** runs, term + semantic, SQR expansion
4. **BRF+SQR** runs, term + semantic, BRF expanded terms + SQR expanded semantic content
## Post-hoc results

### MAP values (trec_eval)

<table>
<thead>
<tr>
<th>SQR</th>
<th>no BRF</th>
<th>With BRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>no SQR</td>
<td>0.3418</td>
<td>0.3246</td>
</tr>
<tr>
<td>1:1</td>
<td>0.2869</td>
<td>0.2631</td>
</tr>
<tr>
<td>2:1</td>
<td>0.3289</td>
<td>0.2958</td>
</tr>
<tr>
<td>5:1</td>
<td><strong>0.3441</strong></td>
<td><strong>0.3157</strong></td>
</tr>
<tr>
<td>10:1</td>
<td>0.3439</td>
<td>0.3269</td>
</tr>
<tr>
<td>100:1</td>
<td>0.3415</td>
<td>0.3204</td>
</tr>
<tr>
<td>1000:1</td>
<td>0.3379</td>
<td>0.3183</td>
</tr>
</tbody>
</table>

**XLDB official best:** 0.3354

**GeoTime best:** 0.4158
Lessons learned

● 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 means more indexes used on retrieval
  - summing partial scores from multiple indexes with BM25 unbalances retrieval focus
  - Best term:semantic ratios around 5:1
Conclusions

- **Semantic query reformulation** can achieve good retrieval performances for geographic and temporal-flavoured queries.

- **Reasoning answers** to add entities is hard, but grounding entities and detecting their roles is easier and very important for document ranking.

- **Mixing term and semantic indexes** must be done carefully: untuned index weights may bias retrieval.
Experiments with Semantic-flavored Query Reformulation of Geo-Temporal Queries

Nuno Cardoso\textsuperscript{1} and Mário J. Silva\textsuperscript{2}

\textsuperscript{1} Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal and SINTEF Natural Language Technologies Group, SINTEF ICT, Oslo, Norway
\textsuperscript{2} Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE, Lisbon, Portugal
ncardoso@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt
Prototype snapshots

**NEW FOREST MANAGEMENT RULES DRAW PRAISE AND CRITICISM**

The nation’s 155 national forests and making it easier for regional forest managers to decide whether (forest and grasslands). They also cut back on requirements for public participation in forest planning decisions. Forest Service officials said the rules are designed to give local foresters more (forest) foresters see as a legal and regulatory gridlock that has delayed forest plans for years because (forest) want the forest to look and be in the future,” said Rick D. Cables, the Forest Service 3KB – 2004-12-22 | Show | Details | More info

**FOREST DISERVICE (FOR USE)**

FOREST DISERVICE UNTIL NOW, both national forests in New England have held the line against the all-terrain vehicles that chew up the forest floor, speed erosion, spread invasive species. Forest in New Hampshire and Maine maintains the ban on ATVs. But the recently released draft plan for the Green Mountain National Forest in Vermont foresees letting ATV owners cross national forest land on “corridors” connecting with a larger trail system outside the national forest. National 5KB – 2005-04-24 | Show | Details | More info

**AFTER THE STORMS CLEARING OF FOREST READY TO RESUME**

New plant growth and renewal. A lack of fires in the forest created a dense canopy that allowed relatively little sunlight to reach the forest floor, discouraging many plants from growing. Now, those plans are back on track. A salvage logging operation to remove large trees from the forest floor of the 26,488-acre state forest in eastern Polk County(...) so forest visitors will know what to expect. Signs will alert visitors to what’s happening. “We...
Prototype snapshots
Prototype snapshots

[Image of a webpage with search results for Forest Fires in Portugal]