GIRSA-WP at GikiCLEF: Integration of Structured Information and Decomposition of Questions

Sven Hartrumpf\textsuperscript{1} and Johannes Leveling\textsuperscript{2}

\textsuperscript{1}: Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen),
Hagen, Germany
\textsuperscript{2}: Centre for Next Generation Localisation (CNGL)
Dublin City University
Dublin, Ireland

CLEF 2009, 30 Sept. – 2 Oct., Corfu, Greece
Outline

Motivation

System Description

Recursive Question Decomposition

Experiments

Results

Conclusions

Future Work
Prior Work: QA, GIR, and their Combination

InSicht question answering system
(participated at QA@CLEF 2004–2008)
+
GIRSA geographic information retrieval system
(participated at GeoCLEF 2006–2008)
=
GIRSA-WP combination of methods
(participated at GikiP 2008)
GIRSA-WP

- **InSicht**: Wikipedia categories, infobox information → natural language description; Wikipedia articles and NL descriptions → syntactic-semantic parser WOCADI → semantic networks

- **GIRSA**: standard IR on geographically annotated documents (normalized location names etc.); index abstracts of Wikipedia articles, but as full-text (not on a per-sentence basis)

- **GIRSA-WP**: semantic filtering / EAT; merge results (combMAX); add multilingual results using links to other languages; add support snippets
What capitals of Dutch provinces received their town privileges before the fourteenth century?

→ Name capitals of Dutch provinces.
   → Name Dutch provinces.
   = Zeeland (support from article 1530: Besonders betroffen ist die an der Scheldemündung liegende niederländische Provinz Zeeland.)
   → Name capitals of Zeeland.
   = Middelburg (support from article Miniatuur Walcheren: ... in Middelburg, der Hauptstadt von Seeland (Niederlande).)
   = Middelburg (answer to revised question can be taken without change)
   → Did Middelburg receive its town privileges before the fourteenth century?
   = Ja./Yes. (support from article Middelburg: 1217 wurden Middelburg durch Graf Willem I. ... die Stadtrechte verliehen.)
   = Middelburg (support: three sentences, from three articles, see above)

: :
Experiments

Three runs:

- **Run 1**: only results from InSicht.
- **Run 2**: results from InSicht and GIRSA, using a standard query formulation and a standard IR model (tf-idf) in GIRSA.
- **Run 3**: results from InSicht and GIRSA, using a Boolean conjunction of the standard query formulation employed for GIRSA and (at most two) keywords extracted from the topic.
## GIRSA-WP Results for GikiCLEF 2009

<table>
<thead>
<tr>
<th>Run</th>
<th>Answers</th>
<th>Correct answers</th>
<th>Precision</th>
<th>GikiCLEF score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>30</td>
<td>0.7895</td>
<td>24.7583</td>
</tr>
<tr>
<td>2</td>
<td>994</td>
<td>107</td>
<td>0.1076</td>
<td>14.5190</td>
</tr>
<tr>
<td>3</td>
<td>985</td>
<td>142</td>
<td>0.1442</td>
<td>23.3919</td>
</tr>
</tbody>
</table>
Conclusions

- GikiCLEF questions seem to be harder than QA@CLEF questions
- Temporal and geographical constraints pose additional problems for QA techniques
- Correct answers can often not be found in one step; instead, subproblems must be solved
- Indexing shorter (abstracted) Wikipedia articles returned a higher number of correct results
- The annotation of geographic entities in the documents ensured a relatively high recall
- GIRSA-WP’s multilingual approach is too simple: it relies on the Wikipedia of one language (German) and adds linked articles in other languages
Future Work

▶ Improve the use of additional information on Wikipedia pages (image captions, lists, categories, infoboxes, other tables)
▶ Improve multilingual support