A fully-automatic approach to answer geographic queries: GIRSA-WP at GikiP

Johannes Leveling    Sven Hartrumpf

Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de
Main idea

InSicht (Hartrumpf, 2005)
- open-domain QA system
- based on matching semantic network representations of question and documents
- supports question decomposition e.g. temporal or geographical constraints

+ GIRSA (Leveling and Hartrumpf, 2008)
- textual GIR system
- supports methods to boost recall e.g. normalizing location indicators
- supports methods to boost precision e.g. metonymy recognition

= GIRSA-WP (GIRSA for Wikipedia)
- automatic combination of InSicht and GIRSA
• applies semantic filter on answer candidates
• merges results from InSicht and GIRSA by using the maximum score of documents
• returns list of Wikipedia article names
• simple multilingual approach: follow German Wikipedia links to articles in English and Portuguese
Semantic filter (1/2)

- in QA: check expected answer type of answer candidates
- for GIRSA-WP: check semantic answer types (semantic sort and features, see Helbig (2006))
  - extract word representing the answer type from topic title and description (the first noun not a proper noun)
  - parse these words with WOCADI, a syntactico-semantic parser (includes a disambiguation of words) and find semantic features corresponding to the extracted words
  - parse the answer candidates (titles of Wikipedia articles) and determine their semantic features
  - test if unification of semantic features succeeds; discard answer candidate, otherwise
Semantic filter (2/2)

- *Which Swiss cantons border Germany?*
  → extracted word: *cantons*
- parse result: corresponding concept is *canton*
  - artificial geographical entity *or* regional institution
  - legal-person: +, movable: −, etc.
- answer candidate *Cross-Border-Leasing:*
  - prototypical-theoretical-concept
  - legal-person: −, movable: −
  → semantic features not unifiable
- answer candidate *Aargau:*
  → unifiable semantic features
Experiments and results

- six runs submitted: three with threshold score of 0.01 and varied settings for stemming, location name normalization, and noun decompounding; additional three experiments with threshold score of 0.03
- 798 (372) answers found
- 79 correct answers in best run
Conclusions (1/2)

GikiP topics

- are at least as difficult as QA or GeoCLEF topics
- aim at a wider range of expected answer types
- include complex geographic relations
  (GP2: *outside*, GP4: *on the border*),
  restrictions on measurable properties
  (GP3: *more than*, GP13: *longer than*), and
  temporal constraints

⇒ new challenge for QA and GIR community
Conclusions (2/2)

- GIRSA:
  - indexing single sentences was meant to ensure a high precision (but did not work);
  - geographic entities have not been annotated at all in the Wikipedia documents

- InSicht:
  - important information is given in tables (like inhabitant numbers), but WOCADI ignores these
  - the semantic matching approach is still too strict for the IR oriented parts of GikiP queries (similarly for GeoCLEF)

⇒ tasks for future work
Selected References

