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

Sven Hartrumpf[1]
Johannes Leveling[2]
[1]Intelligent Information and Communication Systems (IICS),
University of Hagen (FernUniversität in Hagen), Germany
[2]Centre for Next Generation Localisation (CNGL),
Dublin City University, Dublin 9, Ireland
Sven.Hartrumpf@fernuni-hagen.de
Johannes.Leveling@computing.dcu.ie

Abstract

This paper describes the current GIRSA-WP system and the experiments performed for GikiCLEF 2009. GIRSA-WP (GIRSA for Wikipedia) is a fully-automatic, hybrid system combining methods from question answering (QA) and geographic information retrieval (GIR). It merges results from InSicht, a deep (text-semantic) open-domain QA system, and GIRSA, a system for textual GIR.

For the second participation (the first participation was for the pilot task GikiP 2008), the GIR methods were adjusted by switching from a sentence-based retrieval to an abstract-based retrieval. Furthermore, geographic names and location indicators in Wikipedia articles were annotated before indexing. The QA methods were extended by allowing more general recursion with question decomposition. In this way, complex questions, which are frequent in GikiCLEF, can be answered by first answering several depending questions and exploiting their answers. Two new resources of structured information from Wikipedia were integrated, namely the categories assigned to articles and the infobox file from DBpedia, which is an automatic information extraction approach for Wikipedia data. Both resources were exploited by reformulating them in a restricted natural language form. In this way, they can be used as any other text corpus. A semantic filter in GIRSA-WP compares the expected answer type derived from the question parse to the semantics of candidate answers.

Three runs were submitted. The first one contained only results from the QA system; as expected it showed high precision, but low recall. The combination with results from the GIR system increased recall considerably, but reduced precision. The second run used a standard IR query, while the third run combined such queries with a Boolean query with selected keywords. The evaluation showed that the third run was significantly better than the second run. In both cases, the combination of the GIR methods and the QA methods was successful in combining their strengths (high precision of deep QA, high recall of GIR), but the overall performance leaves much room for improvements. For example, the multilingual approach is too simple. All processing is done in only one Wikipedia (the German one); results for the nine other languages are collected only by following the translation links in Wikipedia.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods; Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation; Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Experimentation, Measurement, Performance

Keywords

Geographic Information Retrieval, Question Answering, Questions Beyond Factoids, Temporal and Local Constraints, Cross-language Information Retrieval, Wikipedia

1 Introduction

GIRSA-WP (GIRSA for Wikipedia) is a fully-automatic, hybrid system combining methods from question answering (QA) and geographic information retrieval (GIR). It merges results from InSicht, an open-domain QA system [3], and GIRSA, a system for textual GIR [6]. GIRSA-WP has already participated at the preceding pilot task, GikiP 2008 [7, 8], and was improved based on this and other evaluations.

2 System Description

The GIRSA-WP system used for GikiCLEF 2009 integrates two basic systems: a deep (text-semantic) QA system (InSicht) and a GIR system (GIRSA, GIR with semantic annotation). Each question is processed by both basic systems; GIRSA-WP filters their results semantically to improve precision and combines both result streams yielding a final result of Wikipedia article names, additional supporting article names (if needed), and supporting text snippets (the latter is not required by the GikiCLEF guidelines, but helpful for users).

The semantic filter checks whether the expected answer type (EAT) of the question and the title of a Wikipedia article are semantically compatible. This technique is widely known from QA for typical answer types such as PERSON, ORGANIZATION, or LOCATION. In our system, a concept (a disambiguated word) corresponding to the EAT is extracted from the question. This concept and the title of a candidate article are parsed by WOCADI [2], a syntactico-semantic parser for German. The semantic representations (comprising the sort and the semantic features, see [5] for details on the semantic representation formalism MultiNet) of the semantic heads are unified. If this unification succeeds, the candidate article is kept; otherwise it is discarded. For example, from topic GC-2009-06 (*Which Dutch violinists held the post of concertmaster at the Royal Concertgebouw Orchestra in the twentieth century?*), the concept extracted as EAT is violinist.1.1, whose semantic representation belongs to the class human (*human-object* in MultiNet). There are 87 such semantic classes, which can also be disjunctively connected for underspecification or for so-called semantic molecules (or semantic families).

The retrieval in the GIR system works on the first few (two or three) sentences of the Wikipedia articles. Geographic names and location indicators in the articles were automatically annotated (see [6] for a discussion of this approach). As a result of our participation in GikiCLEF last year, we found that the full Wikipedia articles may be too long and indexing on a per-sentence basis does not provide enough context for matching. Therefore, we focused on the most important parts of the Wikipedia articles (to increase precision for GIRSA), and changed to full-document indexing.

For the GikiCLEF 2009 experiments, the questions were analyzed by the parser and sent to GIRSA and InSicht. In GIRSA, the top 1000 results were retrieved, with scores normalized to the interval $[0, 1]$; on average, GIRSA returned 153 and 395 documents per question for run 2 and 3, respectively (see Sect. 3).

For results returned by both GIRSA and InSicht, the maximum score was chosen (combMAX, [1]). Results whose score was below a given threshold were discarded and the semantic filter was applied to the remaining results. To obtain multilingual results, the German article names were 'translated' to the nine other languages using the Wikipedia linking between languages.

Besides the inter-wiki links, GIRSA-WP uses one further information type from Wikipedia: the categories assigned to articles. Note that other Wikipedia information types like intra-wiki (i.e. inter-article) links and Internet links are still ignored.

For the first time, two resources that contain structured information and are derived directly (categories) or indirectly (DBpedia) from Wikipedia were integrated into GIRSA-WP. The direct source of categories assigned to articles was exploited by extracting categories from the Wikipedia XML file. The resulting relations of the form *in_category(article_title, category)* were reformulated in the following form: *article_title ist ein/ist eine/ ... category*/'*article_title is a ... category*'. Some automatic corrections for frequent cases where the text would be syntactically and/or semantically incorrect were implemented. The remaining errors were largely unproblematic because the processing by InSicht's parser detects them and avoids incorrect semantic networks. In this way, 1.1 million semantic networks were generated for 1.5 million sentences derived from around 2 million *in_category* relations.

The DBpedia data (more specifically: version 3.2 of the file infobox_de.nt, the infobox information from the German Wikipedia encoded in N-Triples, a serialization of RDF; see `http://wiki.dbpedia.org/` for details) is integrated similarly to the category data by rephrasing it in natural language. As there are many different relations in DBpedia only some frequent and relevant relations are covered currently. Each selected relation (currently 19) is linked to an abstract relation (currently 16) and a natural language pattern. For example, the triple

```
<http://dbpedia.org/resource/Andrea_Palladio>
<http://dbpedia.org/property/geburtsdatum>
"1508-11-08"^^<http://www.w3.org/2001/XMLSchema#date>
```

is translated to *Andrea Palladio wurde geboren am 08.11.1508.*/'*Andrea Palladio was born on 08.11.1508.*' This generation process led to around 460,000 sentences derived from around 4,400,000 triples in the DBpedia file.

The detour via natural language for structured information resources is slower and can introduce some errors. But the advantage is that all resources are treated in the same way (and hence can be used in the same way to provide answer support etc.). In addition, the parser is able to deal with ambiguities (for example, names referring to different kinds of entities) that had to be resolved explicitly on the structured level otherwise.

The QA system (InSicht) compares the semantic representation of the question and the semantic representations of document sentences. To go beyond exact matching, InSicht applies many techniques, e.g. coreference resolution, query expansion by inference rules and lexicosemantic relations, and splitting the query semantic network at certain semantic relations. In the context of GikiCLEF, InSicht results (which are generated answers in natural language) must be mapped to Wikipedia article names; if this is not straightforward, the article name of the most important support is taken.

InSicht employed a new special technique called *question decomposition* (or *query decomposition*, see [4] for details) for GeoCLEF 2007, GeoCLEF 2008, and GikiP 2008. An error analysis showed that sometimes it is not enough to decompose a question once. For example, question GC-2009-07 (*What capitals of Dutch provinces received their town privileges before the fourteenth century?*) is decomposed into the subquestion *Name capitals of Dutch provinces.* and revised question *Did ⟨subanswer-1⟩ receive its town privileges before the fourteenth century?* Unfortunately, the subquestion is still too complex and unlikely to deliver many (if any) answers. This situation changes if one decomposes the subquestion further into a subquestion (second level) *Name Dutch provinces.* and revised question (second level) *Name capitals of ⟨subanswer-2⟩*. InSicht's processing of question GC-2009-07 is illustrated in Fig. 1. Note that for readability the supporting texts are shortened and not translated. All subquestions and revised questions are shown in natural language, while the system operates mostly on the semantic (network) level.

Question decomposition, especially in its recursive form, is a very powerful technique that can provide answers and justifications for complex questions. However, the success rates at each decomposition combine in a multiplicative way. For example, if the QA system has an average success rate of 0.5, a double

question:
Welchen Hauptstädten niederländischer Provinzen wurde vor dem vierzehnten Jahrhundert das Stadtrecht gewährt?
'*What capitals of Dutch provinces received their town privileges before the fourteenth century?*'
subquestion level 1:
Nenne Hauptstädte niederländischer Provinzen.
'*Name capitals of Dutch provinces.*'
subquestion level 2:
Nenne niederländische Provinzen.
'*Name Dutch provinces.*'
1st subanswer level 2:
Zeeland (support from article *1530*:
Besonders betroffen ist die an der Scheldemündung liegende niederländische Provinz Zeeland.)
2nd subanswer level 2: ...
⋮
1st revised question level 2:
Nenne Hauptstädte von Zeeland.
'*Name capitals of Zeeland.*'
2nd revised question level 2: ...
⋮
1st answer to 1st revised question level 2:
Middelburg (support from article *Miniatuur Walcheren*:
... in Middelburg, der Hauptstadt von Seeland (Niederlande).)
1st answer to 2nd revised question level 2: ...
⋮
1st subanswer level 1:
Middelburg (note: answer to 1st revised question level 2 can be taken without change)
2nd subanswer level 1: ...
⋮
1st revised question level 1:
Wurde Middelburg vor dem vierzehnten Jahrhundert das Stadtrecht gewährt?
'*Did Middelburg receive its town privileges before the fourteenth century?*'
2nd revised question level 1: ...
⋮
answer to 1st revised question level 1:
Ja./'*Yes.*' (support from article *Middelburg*:
1217 wurden Middelburg durch Graf Willem I. ... die Stadtrechte verliehen.)
answer to 2nd revised question level 1: ...
⋮
1st answer:
Middelburg (support: three sentences, here from different articles, see supports listed in previous steps)
2nd answer: ...
⋮

Figure 1: Illustration of successful recursive question decomposition for GC-2009-07.

Table 1: Evaluation results for all GIRSA-WP runs.

Run	Answers	Correct answers	Precision	GikiCLEF score
1	38	30	0.7895	24.7583
2	994	107	0.1076	14.5190
3	985	142	0.1442	23.3919

decomposition as described above (leading to questions on three levels) will have an average success rate of 0.125 ($= 0.5 \cdot 0.5 \cdot 0.5$).

3 Experiments

We produced three runs with the following experiment settings:

- 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.

4 Evaluation and Discussion

InSicht achieved a higher precision than GIRSA-WP as a whole: 0.7895 compared to 0.1076 and 0.1442 for run 2 and 3, respectively (see Table 1; the definition of the GikiCLEF score and other task details can be found in the GikiCLEF overview paper), but InSicht's low recall (30 correct answers compared to 107 and 142 for run 2 and 3, respectively) is still problematic as already seen in similar evaluations, e.g. GikiP 2008. As intended, InSicht aims for precision, GIRSA for recall, and GIRSA-WP tries to combine both in an advantageous way.

The overall performance of GIRSA-WP is not satisfying, yet. We made the following general observations:

- On average, GikiCLEF questions seem to be harder than QA@CLEF questions from the years 2003 till 2008.
- Especially the presence of temporal and spatial (geographical) constraints in GikiCLEF questions poses challenges for QA techniques.
- As our question decomposition experiments indicate, correct answers can often not be found in one step; instead, subproblems must be solved or subquestions must be answered in the right order.
- Indexing shorter (abstracted) Wikipedia articles returned a higher number of correct results (which was tested on some manually annotated data before submission). Similarly, the annotation of geographic entities in the documents (i.e. conflating different name forms etc.) ensured a relatively high recall.
- The use of the query formulation which combines keywords extracted from the query with a standard IR query (run 3) increases precision (+34%) and recall (+33%) compared to the standard IR query formulation (run 2).
- The system's multilingual approach is too simple because it relies only on the Wikipedia of one language (German) and adds results by following title translation links to other languages. Therefore for questions that have no or few articles in German, relevant articles in other languages cannot be found.

5 Future Work

Some resources are not yet exploited to their full potential. For example, almost half of the category assignments are ignored (see Sect. 2). Similarly, many attribute-value pairs from infoboxes in DBpedia are not covered by GIRSA-WP currently. The cross-language aspect should be improved by processing at least one more Wikipedia version, preferably the largest one: the English Wikipedia.

Acknowledgments

This research was in part supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL).

References

[1] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, NIST Special Publication 500-215, pages 243–252. National Institute for Standards and Technology, 1994.

[2] Sven Hartrumpf. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany, 2003.

[3] Sven Hartrumpf. Question answering using sentence parsing and semantic network matching. In Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck, and Bernardo Magnini, editors, *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, volume 3491 of *Lecture Notes in Computer Science*, pages 512–521. Springer, Berlin, 2005.

[4] Sven Hartrumpf. Semantic decomposition for question answering. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikos Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI)*, pages 313–317, Patras, Greece, July 2008.

[5] Hermann Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, 2006.

[6] Johannes Leveling and Sven Hartrumpf. Inferring location names for geographic information retrieval. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science (LNCS)*, pages 773–780, Berlin, 2008. Springer.

[7] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. Getting geographical answers from Wikipedia: the GikiP pilot at CLEF. In *Results of the CLEF 2008 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark, September 2008.

[8] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling, and Yvonne Skalban. GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In *Evaluating Systems for Multilingual and Multimodal Information Access, CLEF 2008*, volume 5706 of *Lecture Notes in Computer Science (LNCS)*, pages 894–905, Berlin, 2009. Springer.