Recursive Question Decomposition for Answering Complex Geographic Questions

Sven Hartrumpf[1] and Johannes Leveling[2]

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

Abstract. This paper describes the GIRSA-WP system and the experiments performed for GikiCLEF 2009, the geographic information retrieval task in the question answering track at CLEF 2009. Three runs were submitted. The first one contained only results from the InSicht QA system; it showed high precision, but low recall. The combination with results from the GIR system GIRSA 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 achieved significantly higher mean average precision (MAP) than the second run. In both cases, integrating GIR methods and QA methods was successful in combining their strengths (high precision of deep QA, high recall of GIR), resulting in the third-best performance of automatic runs in GikiCLEF. The overall performance still leaves 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 by following the translation links in 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 [1], and GIRSA, a system for textual GIR [2]. GIRSA-WP has already participated in the preceding pilot task, GikiP 2008 [3, 4], and was improved based on this and other evaluations.

2 System Description

2.1 GIRSA-WP Subsystems

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 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. The title of each candidate article is parsed by a syntactico-semantic parser for German [5]. The resulting semantic representations (comprising the sort and the semantic features, see [6] for details on the semantic representation formalism MultiNet) of the representations from the question and from the article title 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 combined to form disjunctive expressions 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 (e.g. name variants and adjectives corresponding to toponyms) in the articles were automatically annotated and normalized (see [2] 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 InSicht's 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 run 3, respectively (see Sect. 3). For results returned by both GIRSA and InSicht, the maximum score was chosen (combMAX, [7]). Results whose score was below a given threshold were discarded and the semantic filter was applied to the remaining results.

2.2 External Knowledge

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 is integrated in a similar way into GIRSA-WP by rephrasing it in natural language. 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 was processed (see `http://wiki.dbpedia.org/` for details). As there are many different relations in DBpedia, only some frequent and relevant relations are covered currently. Each selected relation (currently 19) is associated with 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 of translating structured information resources to natural language is performed with the goal to treat all resources in the same way, i.e. parsing them to obtain their representation as semantic networks. Hence, the results can be used in the same way, e.g. for reasoning and to provide answer support. In addition, the parser is able to resolve ambiguities; for example, names referring to different kinds of entities that had to be disambiguated 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 lexico-semantic 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.

2.3 Recursive Question Decomposition

InSicht employed a new special technique called *question decomposition* (or *query decomposition*, see [8] 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* $\langle$*SubA(nswer)*$^1\rangle$ *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 a revised question (second level) *Name capitals of* $\langle$*SubA(nswer)*$^2\rangle$. InSicht's processing of question GC-2009-07 is illustrated in more detail in Fig. 1. For brevity and better readability, additional question reformulation phases and intermediate stages have been omitted and 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 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: 0.7895 compared to 0.1076 and 0.1442 for run 2 and run 3, respectively (see Table 2). The definition of the GikiCLEF score and other task details can be found in [9]. But InSicht's low recall (only 30 correct answers compared to 107 and 142 correct answers for run 2 and run 3, respectively) is still problematic as has already been 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.

Table 1. Illustration of successful recursive question decomposition for topic GC-2009-07. The superscript designates the level of recursion, the subscript distinguishes alternatives on the same level of recursion.

Q^0	*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?'*
$\mathrm{SubQ}^1 \leftarrow \mathrm{Q}^0$	*Nenne Hauptstädte niederländischer Provinzen.* *'Name capitals of Dutch provinces.'*
$\mathrm{SubQ}^2 \leftarrow \mathrm{SubQ}^1$	*Nenne niederländische Provinzen.* *'Name Dutch provinces.'*
$\mathrm{SubA}^2_1 \leftarrow \mathrm{SubQ}^2$	*Zeeland* (support from article *1530*: *Besonders betroffen ist die an der Scheldemündung liegende niederländische Provinz Zeeland.*)
$\mathrm{SubA}^2_2 \leftarrow \mathrm{SubQ}^2$	*Overijssel* ...
$\vdots$	
$\mathrm{RevQ}^1_1 \leftarrow \mathrm{SubA}^2_1 + \mathrm{SubQ}^1$	*Nenne Hauptstädte von Zeeland.* *'Name capitals of Zeeland.'*
$\mathrm{RevA}^1_1 \leftarrow \mathrm{RevQ}^1_1$	*Middelburg* (support from article *Miniatuur Walcheren*: *...in Middelburg, der Hauptstadt von Seeland (Niederlande).*; note that the orthographic variants *Zeeland/Seeland* are identified correctly)
$\mathrm{SubA}^1_1 \leftarrow \mathrm{RevA}^1_1$	*Middelburg* (note: answer to revised question can be taken without change)
$\mathrm{RevQ}^0_1 \leftarrow \mathrm{Q}^0 + \mathrm{SubA}^1_1$	*Wurde Middelburg vor dem vierzehnten Jahrhundert das Stadtrecht gewährt?* *'Did Middelburg receive its town privileges before the fourteenth century?'*
$\mathrm{RevA}^0_1 \leftarrow \mathrm{RevQ}^0_1$	*Ja.*/*'Yes.'* (support from article *Middelburg*: *1217 wurden Middelburg durch Graf Willem I. ... die Stadtrechte verliehen.*)
$\mathrm{A}^0_1 \leftarrow \mathrm{RevA}^0_1 + \mathrm{SubA}^1_1$	*Middelburg* (support: three sentences, here from different articles, see supports listed in previous steps)

In order to investigate the complementarity of GIRSA and InSicht, two experimental runs were performed after the campaign. In run 4 and 5, only results from GIRSA are included; the settings correspond to the ones from run 2 and run 3, respectively. The number of correct answers (compare run 2 and sum of run 1 and 4; compare run 3 and sum of run 1 and 5) shows that the overlap of GIRSA and InSicht is minimal: only 1 correct answer is shared. Hence, the combination of both systems is very effective. The results indicate also that the combination of the two systems profits from keeping most of InSicht's correct results and discarding some incorrect results from GIRSA.

Table 2. Evaluation results for the three official GIRSA-WP runs and two experimental runs.

Run	System	Answers	Correct answers	Precision	GikiCLEF score
1	InSicht	38	30	**0.7895**	**24.7583**
2	InSicht+GIRSA	994	107	0.1076	14.5190
3	InSicht+GIRSA	985	**142**	0.1442	23.3919
4	GIRSA	964	78	0.0809	7.8259
5	GIRSA	961	113	0.1176	15.0473

We made the following general observations:

Complexity of Questions GikiCLEF topics are open-list questions and do not include factoid or definition questions. 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 and GIR techniques, which cannot be met successfully by shallow (i.e. syntactically-oriented) natural language processing or traditional IR techniques alone.

Combination of standard and Boolean IR As the GikiCLEF topics resemble open list questions, the aim of the GIR approach was to retrieve results with a high initial precision. 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).

Question decomposition 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. For some topics, a promising subquestion leads to many answers (for example, the subquestion *Nenne Städte in Deutschland./'Name cities in Germany.'* for topic GC-2009-47), which cannot be efficiently handled for the revised questions so that correct answers are missed.

Abstract indexing 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.

Multilingual Results The system's multilingual approach is too simple because it relies only on the Wikipedia in one language (German) and adds results by following title translation links to other languages. For eleven GikiCLEF topics (5, 10, 15, 16, 18, 24, 26, 27, 28, 36, and 39) no articles in German were assessed as relevant. Therefore for questions that have no or few articles in German, relevant articles in other languages cannot be found. Processing the Wikipedia articles in parallel for another language in the same way also will allow to find subanswers supported by articles in other languages, i.e. the supporting texts may not only be distributed among different articles of only one languages, but also among articles in different languages.

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 material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142.

References

1. Hartrumpf, S.: Question answering using sentence parsing and semantic network matching. In Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B., eds.: Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004. Volume 3491 of LNCS. Springer, Berlin (2005) 512–521
2. Leveling, J., Hartrumpf, S.: Inferring location names for geographic information retrieval. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A., Petras, V., Santos, D., eds.: Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007. Volume 5152 of LNCS. Springer, Berlin (2008) 773–780
3. Santos, D., Cardoso, N., Carvalho, P., Dornescu, I., Hartrumpf, S., Leveling, J., Skalban, Y.: 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)

4. Santos, D., Cardoso, N., Carvalho, P., Dornescu, I., Hartrumpf, S., Leveling, J., Skalban, Y.: GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia. In: Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17–19, Revised Selected Papers. Volume 5706 of LNCS. Springer, Berlin (2009) 894–905
5. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. Der Andere Verlag, Osnabrück, Germany (2003)
6. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Berlin (2006)
7. Fox, E.A., Shaw, J.A.: Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2). NIST Special Publication 500-215, National Institute for Standards and Technology (1994) 243–252
8. Hartrumpf, S.: Semantic decomposition for question answering. In Ghallab, M., Spyropoulos, C.D., Fakotakis, N., Avouris, N., eds.: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI), Patras, Greece (July 2008) 313–317
9. Santos, D., Cabral, L.M.: GikiCLEF: Expectations and lessons learned. This volume.