University of Hagen at QA@CLEF 2005: Extending Knowledge and Deepening Linguistic Processing for Question Answering

Sven Hartrumpf
Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
Sven.Hartrumpf@fernuni-hagen.de

Abstract

The German question answering (QA) system InSicht participated in QA@CLEF for the second time. It relies on complete sentence parsing, inferences, and semantic representation matching. This year, the system was improved in two main directions. First, the background knowledge was extended by large semantic networks and large rule sets. InSicht's query expansion step can produce more alternatives using these resources. A second direction for improvement was to deepen linguistic processing by treating a phenomenon that appears prominently on the level of text semantics: coreference resolution.

A new source of lexico-semantic relations and equivalence rules has been established based on compound analyses. WOCADI's compound analysis module determined the structure and semantics of compounds when parsing the German QA@CLEF corpus and the German GIRT (German Indexing and Retrieval Test database) corpus. The compound analyses were used in three ways: to project lexico-semantic relations from compound parts to compounds, to establish a subordination hierarchy between compounds, and to derive equivalence rules between nominal compounds and their analytic counterparts, e.g. between *Reisimport* (‘*rice import*’) and *Import von Reis* (‘*import of rice*’). Another source of new rules were verb glosses from GermaNet, a German WordNet variant. The glosses were parsed and automatically formalized.

The lack of coreference resolution in InSicht was one major source of missing answers in QA@CLEF 2004. Therefore the coreference resolution module CORUDIS was integrated into the parsing during document processing. The resulting coreference partition of mentions (or markables) from a document is used to derive additional networks where mentions are replaced by mentions from the corresponding coreference chain in that partition.

The central step in the QA system InSicht, matching (one by one) semantic networks derived from the question parse to document sentence networks, was generalized. Now, a question network can be split at certain semantic relations (e.g. relations for local or temporal specifications); the resulting semantic networks are conjunctively connected.

To evaluate the different extensions, the QA system was run on all 400 German questions from QA@CLEF 2004 and 2005 with varying setups. Some of these extensions showed positive effects, but currently they are minor and not yet statistically significant. At least three explanations play a role. First, the differences in the semantic representation of questions and document sentences are often minimal and do not require much background knowledge to be related. Second, there are some questions that need a lot of inferential steps. For many such inference chains, formalized inferential knowledge like axioms and meaning postulates for concepts are missing. Third, the low recall values of some natural language processing modules, e.g. the parser and the coreference resolution module, can cause a missing inferential link and thereby a wrong empty answer. Work on the robustness of these modules will help to answer more questions correctly.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding, Language generation*

General Terms

Experimentation, Measurement, Performance

Keywords

Question answering, Questions beyond factoids, Deep semantic processing of questions and documents

1 Introduction

The German question answering (QA) system InSicht participated in QA@CLEF for the second time. This year, the system was improved in two main directions. First, the background knowledge was extended by large semantic networks and rule sets. InSicht's query expansion step produces more alternative representations using these resources. A second direction for improvement was to deepen linguistic processing by treating a phenomenon that appears prominently on the level of text semantics: coreference resolution.

The paper starts with a summary of the basic InSicht system (Section 2). Then, the most important improvements since QA@CLEF 2004 are described (Section 3). In Section 4, the resulting system is evaluated on the 400 German questions from QA@CLEF 2004 and 2005. The contribution of different modifications is investigated by running the system with different setups. Some conclusions appear in the final Section 5.

2 Overview of InSicht

The deep semantic question answering system InSicht (Hartrumpf, 2005) relies on complete sentence parsing, inferences, and semantic representation matching. It comprises six main steps.

In the *document processing* step, all documents from a given collection are preprocessed by transforming them into a standard XML format (CES, corpus encoding standard, Ide et al. (1996)) with word, sentence, and paragraph borders marked up by XML elements. Then, all preprocessed documents are parsed by the WOCADI parser (Hartrumpf, 2003), yielding a syntactic dependency structure and more importantly a semantic network representation of the MultiNet formalism (Helbig, 2005) for each document sentence.

In the second step (*query processing*), the user's question is parsed by the same parser that processed the documents. Determining the sentence type (here, often a subtype of *question*) is especially important because it controls some parts of two later steps: query expansion and answer generation. The system does not deal with (expected) answer types or similar concepts; whatever semantic network for a document sentence matches the question and can be reformulated as a natural language expression is a candidate answer.

Next comes *query expansion*: Equivalent and similar semantic networks are derived by means of lexico-semantic relations from a computer lexicon (HaGenLex, see Hartrumpf et al. (2003)) and a lexical database (GermaNet), equivalence rules, and inferential rules like entailments for situations (applied in backward chaining). The result is a set of disjunctively connected semantic networks that try to capture variations of explicit and implicit representations of sentences possibly containing an answer for the user's question.

In the fourth step (*semantic network matching*), all document sentences matching at least one of the semantic networks from query expansion are collected. A two-level approach is chosen for efficiency reasons. First, an index of concepts (disambiguated words with IDs from HaGenLex) is consulted with the relevant concepts from the query networks. Second, the retrieved documents are compared sentence network by sentence network to find a match with a query network.

Answer generation is next: natural language generation rules are applied to matching semantic networks and try to generate a natural language answer from the deep semantic representations. The sentence type and the semantic network itself control the selection of answer rules. The rules also act as a filter for uninformative or bad answers. The result of this step are tuples of generated answer string, answer score, supporting document ID, and supporting sentence ID.

To deal with typically many candidate answers resulting from answer generation, an *answer selection* step is required at the end. It implements a strategy that combines a preference for more frequent answers and a preference for more elaborate answers. The best answers (by default only the best answer) and the supporting sentences (and/or the IDs of supporting sentences or documents) are presented to the user that posed the question.

3 Improvements over the System for QA@CLEF 2004

There are several areas where InSicht has been improved since QA@CLEF 2004. The most notably changed phases are document processing, query expansion, and semantic network matching.

(sub ”riesenpython.1.1” ”riesenschlange.1.1”) (‘*giant python*’ ‘*giant snake*’)
(sub ”rollhockeynationalmannschaft.1.1” ”hockeymannschaft.1.1”) (‘*roller hockey national team*’ ‘*hockey team*’)
(sub ”weizenexport.1.1” ”getreideexport.1.1”) (‘*wheat export*’ ‘*crop export*’)
(syno ”metrosuizid.1.1” ”u-bahnselbstmord.1.1”) (‘*metro suicide*’ ‘*underground train self-murder*’)
(syno ”rehabilitationskrankenhaus.1.1” ”rehaklinik .1.1”) (‘*rehabilitation hospital*’ ‘*rehab hospital (clinic)*’)
(syno ”wirtschaftmodell.1.3” ”ökonomiemodell.1.3”) (‘*(economy) model*’ ‘*economy model*’)

Figure 1: Examples of inferred lexico-semantic relations for compounds.

3.1 Document Processing

The coverage of the WOCADI parser has been increased so that for 51% of all QA corpus sentences a full semantic network is produced (compared to 49% for QA@CLEF 2004, see Hartrumpf (2005)). This was achieved by extending the lexicon HaGenLex and by refining the parser itself. The concept index (a mapping from concept IDs to document IDs), which is used by the matcher for reducing run time, provides more efficient creation and lookup operations than last year because we switched from an external binary tree to a freely available system, qdbm (Quick Database Manager, http://qdbm.sourceforge.net).

3.2 More Query Expansions by Larger Lexico-Semantic Networks and Larger Rule Sets

A new source of lexico-semantic relations and equivalence rules has been established based on compound analyses. WOCADI’s compound analysis module determines structure and semantics of compounds when parsing a text corpus. The 470,000 compound analyses from parsing the German QA@CLEF corpus and the German GIRT corpus were collected. Only compounds that the compound analysis module assigned a semantics where the right (base) compound part is a hyperonym of the compound were considered. Given a compound, synonyms and hyponyms[1] of each compound part are collected by following corresponding relations in the lexicon. Then, each element from the Cartesian product of these alternatives is looked up in the compound analyses mentioned above. If it exists with a given minimal frequency (currently: 1), a relation is inferred based upon the relations between corresponding parts. In case of a contradiction (e.g. the first parts are in a hyponymy relation while the second parts are in a hyperonymy relation), no relation is inferred. This algorithm delivered 16,526 relations: 5,688 SUB (subordination) edges and 10,838 SYNO (synonymy) edges. All of them are lexico-semantic relations between compounds. Some examples are shown in Figure 1.

A more direct use of compound analyses is the extraction of subordination edges (MultiNet uses SUB, SUBR, and SUBS edges for nominal compounds depending upon the nouns ontological sort) representing a hyponymy relation between a compound and its base noun (or adjective). This process led to 387,326 new edges.

A third use of automatic compound analyses is the production of equivalence rules for complement-filling compounds. One can generate for such compounds an equivalence to an analytical form, e.g. between *Reisimport* (‘*rice import*’) and *Import von Reis* (‘*import of rice*’).[2] Currently, only compounds where the base noun has exactly one complement in the lexicon that can (semantically) be realized by the determining noun are treated in this way, so that for 360,000 analyzed nominal compounds in the QA corpus

[1] Hyperonyms can be ignored because hyperonymy is the inverse relation of hyponymy and all inferable relations will also be produced when treating the compound analyses containing the corresponding hyperonym.

[2] By way of a lexical change relation, the representations of both formulations are linked to the representation of a formulation with a verb: *Reis importieren* (‘*to import rice*’).

```
((pre ((member ?r1 (preds subs))))
 (rule ((?r1 ?n1 "drogenkonsum.1.1") ; drug consumption
        <->
        (?r1 ?n1 "konsum.1.1")
        (aff ?n1 ?n2)
        (sub ?n2 "droge.1.1")))
 (name "compound_analysis.sg.drogenkonsum.1.1"))

((pre ((member ?r1 (preds subs))))
 (rule ((?r1 ?n1 "gebäudesanierung.1.1") ; building sanitation
        <->
        (?r1 ?n1 "sanierung.1.1")
        (aff ?n1 ?n2)
        (sub ?n2 "gebäude.1.1")))
 (name "compound_analysis.sg.gebäudesanierung.1.1"))

((pre ((member ?r1 (preds subs))))
 (rule ((?r1 ?n1 "holzerzeugung.1.1") ; wood production
        <->
        (?r1 ?n1 "erzeugung.1.1")
        (rslt ?n1 ?n2)
        (sub ?n2 "holz.1.1")))
 (name "compound_analysis.sg.holzerzeugung.1.1"))
```

Figure 2: Three automatically generated rules for compounds involving a complement of the base noun.

around 13,000 rules were generated. Three simplified MultiNet rules are shown in Figure 2. Variables are preceded by a question mark. The attribute *pre* contains preconditions for variables occurring on both sides of an equivalence rule. The MultiNet relation PREDS corresponds to SUBS and instantiates (or subordinates) not just a single concept but a set of concepts. The relations AFF (affected object) and RSLT (result of a situation) stem from the HaGenLex characterization of the direct object of the base nouns *Konsum* (‘*consumption*’), *Sanierung* (‘*sanitation*’), and *Erzeugung* (‘*production*’). Such an equivalence rule fired only for question qa05_023 (*Welcher frühere Fußballspieler wurde wegen Drogenkonsum verurteilt?*, ‘*Which former soccer player was convicted of taking drugs?*’) because most questions from QA@CLEF 2004 and 2005 are not related to such compounds or their analytical forms.

Another set of rules available to this year’s InSicht stems from parsing verb glosses from GermaNet (a German WordNet variant) and further automatic formalization (see Glöckner et al. (2005) for details). Each rule relates one verb reading with one or more readings of other verbs. None of these rules fired during query expansion of QA@CLEF questions. This was not too much of a surprise because the rule set is quite small (around 200 rules). Nevertheless, this path seems promising as soon as more German glosses become available (from GermaNet or other sources like Wikipedia).

3.3 Coreference Resolution for Documents

Looking at last year’s questions that turned out to be hard to answer for most systems (see Hartrumpf (2005) for error classes and frequencies) and looking at some of our own test questions, the lack of coreference resolution was identified as one major source of errors. (This lack caused 6% of InSicht’s wrong empty answers for questions from QA@CLEF 2004.) Therefore the coreference resolution module

CORUDIS (COreference RUles with DIsambiguation Statistics, see Hartrumpf (2003, 2001) for details) was integrated into the parsing during document processing. If a coreference partition of mentions (or markables) from a document is found the simplified and normalized document networks are extended by networks where mentions are replaced by mentions from the corresponding coreference chain in that partition. For example, if document network d contains mention m_i and m_i is in a nontrivial (i.e. $n > 1$) coreference chain $\langle m_1, \ldots, m_i, \ldots, m_n \rangle$ the following networks are added to the document representation: $d_{m_i|m_1}, \ldots, d_{m_i|m_{i-1}}, d_{m_i|m_{i+1}}, \ldots, d_{m_i|m_n}$ where $d_{m_1|m_2}$ denotes network d with the semantics of mention m_1 substituted by the semantics of mention m_2. Some mention substitutions are avoided if no performance improvement is possible or likely (e.g. if the semantic representations of m_1 and m_2 are identical), but there is still room for beneficial refinements.

The CORUDIS module is not yet efficient enough (which is not surprising because finding the best coreference partition of mentions is NP-hard) so that the search had to be limited by some parameter settings in order to reduce run time. On the down side, this caused that only for 40% of all texts a partition of mentions was found. Furthermore, only 70% of all texts had been analyzed for coreferences when the evaluation from Section 4 was run. Therefore the improvements achievable by coreference resolution will increase in the near future.

Coreference resolution has been rarely described for natural language processing (NLP) applications. For example in information extraction, Zelenko et al. (2004) presented and extrinsically evaluated several coreference resolution modules for such an application but it restricts coreferences to mentions that could be relevant for the given information extraction task. In contrast, WOCADI's coreference resolution module treats all mentions that meet the MUC definition of a markable (Hirschman and Chinchor, 1997).

3.4 More Flexible Matching by Splitting Question Networks

Last year's InSicht matched semantic networks derived from a question parse to document sentence networks one by one. A more flexible approach turned out to be beneficial for IR and geographical IR; so it was tested in InSicht's QA mode, too. The flexibility comes from the fact that a question network is split if certain graph topologies exist: a network is split in two networks at CIRC, CTXT (nonrestrictive and restrictive context, respectively), LOC (location of objects or situations), and TEMP (temporal specification) edges. The resulting parts are conjunctively connected. Section 4 shows which splitting configurations turned out to be most profitable.

4 Evaluation

The current InSicht QA system has been evaluated on the QA@CLEF questions from 2004 and 2005. To investigate the impact of different improvements described in Section 3 the setup was varied in different ways, as can be seen in the second and third column of Table 1. The evaluation metrics reported are the number of right, inexact, and wrong answers and the K1-measure (see Herrera et al. (2004) for a definition). Note that the number of unsupported answers was always zero, so it was omitted for brevity.

For better comparison, the results for InSicht's official run at QA@CLEF 2004 are shown, too. The K1-measure was much lower than this year because the confidence score of last year's system was tuned for confidence-weighted score (CWS). Now InSicht tries to optimize the K1-measure because the K1-measure seems to be a more adequate metric for evaluating QA systems (Herrera et al., 2004).

Table 1: Results for German question sets from QA@CLEF 2004 and 2005. *lexsem* stands for lexico-semantics relations projected to compounds and *hypo* for hyponymy relations for compounds (see Section 3.2); S is the set of relations where query networks can be split (see Section 3.4).

Question Set	Setup		Results			
	Query Expansion	Matching	# Right	# Inexact	# Wrong	K1
2004	*run from QA@CLEF 2004*		67	7	0	-0.327
2004	lexsem	no coreference, $S = \{\textsc{loc}\}$	83	6	0	0.285
2004	lexsem	coreference, $S = \{\textsc{loc}\}$	83	6	0	0.285
2005	lexsem	no coreference, $S = \{\}$	80	7	1	0.260
2005	lexsem	no coreference, $S = \{\textsc{loc}\}$	84	7	1	0.280
2005	lexsem	coreference, $S = \{\textsc{loc}\}$	84	8	0	0.280
2005	lexsem, hypo	coreference, $S = \{\textsc{loc}\}$	86	8	0	0.290

To experiment with cross-language QA, a machine translation of questions was employed. After the machine translation system Systran (as provided on the web) had translated the 200 English questions (from QA@CLEF 2005) into German, they were put into the standard InSicht system. The number of right answers dropped by around 50%, which was mainly due to incorrect or ungrammatical translations. Some translation problems seemed to be systematic, so that a simple postprocessing component could correct some wrong translations, e.g. the temporal question word *when* was translated as *als* instead of *wenn*.

5 Conclusion

The QA system InSicht was extended by new sources of knowledge. From automatic compound analyses, large semantic networks and numerous equivalence rules were derived. The linguistic processing was deepened by integrating the coreference resolution module CORUDIS into document processing.

When evaluated on all 400 questions from QA@CLEF 2004 and 2005 some of these extensions showed positive effects. But unfortunately the effects are minor and not yet statistically significant. The reasons need further investigation but here are three preliminary observations.

First, the differences in the semantic representation of questions and document sentences are often minimal and do not require the kind of knowledge that was generated. We hope that larger test sets will show significant positive effects.

On the other extreme, there are some questions that need much more inferential steps than currently produced by query expansion. The matching approach is quite strict (precision-oriented, while for example the QA system described by Jijkoun et al. (2004) is recall-oriented) and can require long inference chains in order to find answers. The main hindrance to building such chains are missing pieces of formalized inferential knowledge, like axioms for MultiNet and meaning postulates for concepts. Some parts of this knowledge can be automatically generated, see for example Section 3.2.

A third explanation regards the quality of some NLP modules. The still limited recall values of the parser (see Section 3.1), the coreference resolution module (see Section 3.3), and other modules can cause that an inferential link (e.g. a coreference between two nominal phrases) is missing so that a question remains unanswered and a wrong empty (NIL) answer is produced. Such negative effects are typical for applications building on deep syntactico-semantic processing. Therefore the robustness of some NLP modules will be increased in order to answer more questions.

References

Glöckner, Ingo; Sven Hartrumpf; and Rainer Osswald (2005). From GermaNet glosses to formal meaning postulates. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen – Beiträge zur GLDV-Tagung 2005 in Bonn* (edited by Fisseni, Bernhard; Hans-Christian Schmitz; Bernhard Schröder; and Petra Wagner), pp. 394–407. Frankfurt am Main: Peter Lang.

Hartrumpf, Sven (2001). Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proceedings of the Fifth Computational Natural Language Learning Workshop (CoNLL-2001)*, pp. 137–144. Toulouse, France. http://www.aclweb.org/anthology/W01-0717.

Hartrumpf, Sven (2003). *Hybrid Disambiguation in Natural Language Analysis*. Osnabrück, Germany: Der Andere Verlag. ISBN 3-89959-080-5.

Hartrumpf, Sven (2005). Question answering using sentence parsing and semantic network matching. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign* (edited by Peters, C.; P. D. Clough; G. J. F. Jones; J. Gonzalo; M. Kluck; and B. Magnini), volume 3491 of *Lecture Notes in Computer Science (LNCS)*, pp. 512–521. Berlin: Springer.

Hartrumpf, Sven; Hermann Helbig; and Rainer Osswald (2003). The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

Helbig, Hermann (2005). *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer.

Herrera, Jesús; Anselmo Peñas; and Felisa Verdejo (2004). Question answering pilot task at CLEF 2004. In *Results of the CLEF 2004 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2004 Workshop* (edited by Peters, Carol and Francesca Borri), pp. 445–452. Bath, England.

Hirschman, Lynette and Nancy Chinchor (1997). MUC-7 coreference task definition (version 3.0). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/.

Ide, Nancy; Greg Priest-Dorman; and Jean Véronis (1996). *Corpus Encoding Standard*. http://www.cs.vassar.edu/CES/.

Jijkoun, Valentin; Gilad Mishne; Maarten de Rijke; Stefan Schlobach; David Ahn; and Karin Müller (2004). The University of Amsterdam at QA@CLEF 2004. In *Results of the CLEF 2004 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2004 Workshop* (edited by Peters, Carol and Francesca Borri), pp. 321–324. Bath, England.

Zelenko, Dmitry; Chinatsu Aone; and Jason Tibbetts (2004). Coreference resolution for information extraction. In *ACL-2004: Workshop on Reference Resolution and its Applications* (edited by Harabagiu, Sanda and David Farwell), pp. 24–31. Barcelona, Spain: Association for Computational Linguistics.