University of Hagen at QA@CLEF 2007: Coreference Resolution for Questions and Answer Merging

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

Abstract

The German question answering (QA) system InSicht participated in QA@CLEF for the fourth time. InSicht realizes a deep QA approach: it builds on full sentence parses, rule-based inferences on semantic representations, and matching semantic representations derived from questions and document sentences. InSicht was improved for QA@CLEF 2007 in the following main areas: questions containing pronominal or nominal anaphors are treated by a coreference resolver; the shallow QA methods are improved; and finally, our system for the CLEF Answer Validation Exercise is employed for answer merging. Results showed a performance drop compared to last year mainly due to unstable and incomplete handling of the newly added Wikipedia corpus. However, dialog treatment by coreference resolution delivered very accurate results so that follow-up questions can be handled similar to isolated questions.

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*
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*
I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*

General Terms

Experimentation, Measurement, Performance

Keywords

Question answering, Deep semantic processing of questions and documents, Follow-up questions, Coreference resolution, Answer merging

1 Overview

Research described in this paper is part of IRSAW[1], a question answering (QA) framework integrating modules for different tasks such as natural language analysis (by the WOCADI parser), combining dif-

[1]The IRSAW project (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web; LIS 4 – 554975(2) Hagen, BIB 48 HGfu 02-01) is funded by the DFG (Deutsche Forschungsgemeinschaft).

ferent data streams with answer candidates (answer streams), logical answer validation (MAVE), and natural language generation. Three different approaches to create answer candidates are employed. These approaches are applied to the two corpora for QA@CLEF 2007, namely CLEF-NEWS and Wikipedia, actually resulting in six different streams of answer candidates, which are merged in the answer validation phase.

The first answer producer is InSicht (Hartrumpf and Leveling, 2006), a precision-oriented QA system using a semantic network representation of questions and documents (see Sect. 2). It realizes a deep (semantic) QA approach because it tries to employ deep methods in many natural language processing (NLP) areas: it builds on full sentence parses for documents and questions, rule-based inferences on semantic representations, matching semantic representations derived from questions and documents, and natural language generation of answers from semantic representations of documents. Specialized NLP modules resolve temporal deixis and coreferences.

The other two answer producers are QAP (Question Answering by Pattern matching) and MIRA (Modified Information Retrieval (IR) Approach for question answering), see Sect. 3. They employ shallow NLP methods and aim at a high recall to provide a fallback strategy for InSicht. These shallow approaches rely on preprocessed document corpora with sentence boundaries detected. Text segments are transformed into XML and indexed in a database management system supporting the *tf-idf* IR model. Both the QAP and the MIRA module rely on the WOCADI parser for resolving ellipsis and anaphoric references in questions.

2 Changes of InSicht for QA@CLEF 2007

The QA@CLEF task was considerably changed in 2007: The document collection was increased in size and diversity by adding the Wikipedia documents from 2006-11-30 (for German, the number of sentences grew from 5.0 million sentences to 16.5 million sentences); and follow-up questions were included in the test set. These two issues are discussed in the following subsections.

2.1 Wikipedia

For document processing in InSicht, all documents are parsed by WOCADI and intratextual coreferences are resolved by CORUDIS (COreference resolution by RUles and DIsambigation Statistics, Hartrumpf (2001, 2003)). Because the time between guideline release and test set release was too short to process all Wikipedia documents in the normal way, we had to rely on an older parse of the Wikipedia (from 2006-09-25). Although this Wikipedia version is only 2 months older than the recommended snapshot, it turned out to be considerably different in many articles (e.g. 6,813 articles disappeared by renaming, merging, or complete removal, while 38,478 articles were added); this is just another witness of the very dynamic nature of Wikipedia. To save time, coreferences were not resolved.

The German Wikipedia is 2.3 times larger than the traditional German QA@CLEF collection, CLEF-NEWS. In general, InSicht was able to deal with the size increase, but unfortunately the quite unrestricted form of article names (and in InSicht's context, file names) led to an inconsistent concept index that rendered many Wikipedia articles inaccessible to InSicht. So, InSicht's answers came too rarely from Wikipedia, which seems to be the main reason for the performance drop. Aggravating this situation, most test set questions seem to target the Wikipedia subcollection only. Fortunately, the performance drop in InSicht was in part compensated by the improved shallow QA subsystems (see Sect. 3) and the newly integrated answer validator MAVE (see Sect. 4).

2.2 Dialog Treatment

In the years before 2007, all questions could be answered in isolation without any reference to the context, like previous questions or answers. The guidelines for QA@CLEF 2007 removed this restriction by allowing coreferences *to the topic expressed in the first question/answer pair*. In the test set, one question (165) is even more unrestricted: it contains a pronominal anaphor that corefers with an antecedent from the second question of its topic context.

To treat such context-dependent questions, the basic idea was to keep a dialog history containing semantic representations of questions and answers. The dialog history is initialized (i.e. deleted), if the start of a new topic is encountered in the test set. (We have not tried to detect topic boundaries automatically, yet.) On these semantic representations in the dialog history, coreferences are resolved by the general coreference resolver CORUDIS. This module has already been used successfully for coreference resolution on documents since QA@CLEF-2005 (Hartrumpf, 2006).

CORUDIS is a hybrid coreference resolver because it contains both symbolic, linguistically motivated *coreference rules* that license possible coreferences and a statistical *multi-dimensional back-off model* derived from a manually annotated corpus for selecting among licensed alternatives. In addition, CORUDIS employs a whole range of bonus factors like syntactic parallelism, semantic parallelism, and maximality of noun phrases. The back-off model for CORUDIS, which was derived from annotated newspaper articles (from the *Süddeutsche Zeitung*), was taken without any modifications. First, we considered retraining on a dialog corpus with anaphors in questions, which would have been more similar to the application type in QA@CLEF 2007. But to save time, we kept the old model and only added some positive (negative) scores for specific coreference rules that are more (less) likely to be correct in question sequences.

Table 1 contains all 29 questions from the 200 German questions (of QA@CLEF 2007) where coreference resolution is required to find an answer. So, only 34.5% of all 84 follow-up questions for the 47 topics with more than one question require coreference resolution. The second column in the table shows which dependency types occur in the test set; q_1 (a_1) stands for the first question (answer) of a topic. Only two questions (046, 107) contain an anaphor that corefers with an answer; therefore, an answer should only be an alternative to an antecedent candidate from a preceding question. As no answer could be found for the corresponding topic-first questions (045, 105) it did not matter whether one includes or excludes the semantic representation of the first answer from subsequent coreference resolutions for the same topic.

The pronominal anaphor *sie* in question 165 corefers with *Hanamachi* from the *second* question of the topic 163–165). To handle such references to non-first questions, we adapted the dialog processing as follows: a subsequent question is deleted from the dialog history only if it contains an anaphor which was successfully resolved.

The answer producers used only representations where coreferences had been resolved. For the deep producer InSicht, it suffices to integrate the antecedents (and remove the anaphors) on the semantic network level. For the shallow producers to profit from coreference resolution most easily, the integration is performed on the surface level. This leads to question formulations that can be answered without any further context. The original questions and the automatically revised questions can be seen in the first column of Table 1. The average size of the resulting semantic networks (measured by the number of relations after coreference resolution) was 11.39; this shows that QA@CLEF's questions stayed astonishingly stable in terms of semantic size (and approximately specificity) over the last 5 years: 11.30 (2006), 11.30 (2005), 9,73 (2004), 10.98 (2003). But the question difficulty increased because of added phenomena like temporal restrictions, temporal deixis, and anaphors in questions.

Related work for coreference resolution on the question side and on the document side is presented in an overview by Vicedo and Ferrández (2006). Some systems for English that employ coreference resolution on questions in the Context Task of TREC-10 are described by Harabagiu et al. (2001); Lin and Chen (2001); Oh et al. (2001). All three approaches handle coreferences in order to add the keywords from the question (or answer) containing the antecedent for an anaphor in the current question. In our approach, we try to construct a question that can be answered as a question without context.

Elliptical questions, although not occurring in this year's test set, were implemented because they are frequent and central for QA systems with dialog handling. A simple heuristic was applied to detect an elliptical question: if no verb is contained in the parse of the question and the question is short, the question will be treated as elliptical. Ellipsis resolution simply replaces the *question focus* of the previous question by the question focus of the current question, e.g. the elliptical question *Wo?* (*Where?*) after the question *Seit wann X?* (*Since when X?*) becomes *Wo X?* (*Where X?*).

Table 1: Analysis of coreferences in the German questions of QA@CLEF 2007. The question in the first column contains in parentheses the correct antecedent for an anaphor. The English translations are from the English-German test set. The parenthesized item in the second column (dependency type) indicates that this resolution alternative is less likely to lead to an answer. The resolution result of CORUDIS in the third column can be right (R; 26 cases), missing (M; 3 cases), or wrong (W; 0 cases).

Question	Dep. type	Res.
Gegen wen ist sie (Steffi Graf) im Halbfinale der French Open im Jahr 1994 ausgeschieden? (038) Against whom did she drop out in the semifinal of the French Open in 1994?	$q_2 \rightarrow q_1$	R
In welchem Jahr hat er (Goethe) die Krönung von Kaiser Joseph dem Zweiten beobachtet? (040) In which year did he watch the coronation of the Emperor Joseph the second?	$q_2 \rightarrow q_1$	R
In welchem Jahr wurde er (Goethe) geboren? (041) When was he born?	$q_3 \rightarrow q_1$	R
In welcher Stadt ist er (Goethe) gestorben? (042) In which city did Goethe die?	$q_4 \rightarrow q_1$	R
In welchem Jahr erhielt er (answer 045: Victor Fleming; question 045: der Regisseur von "Vom Winde verweht") einen Oscar für "Vom Winde verweht"? (046) In which year did he receive the Oscar for "Gone with the Wind"?	$q_2 \rightarrow \mathbf{a_1}$ ($q_2 \rightarrow q_1$)	M
In welchem Jahr ist er (Freddy Mercury) gestorben? (051) In which year did he die?	$q_2 \rightarrow q_1$	R
Welches Musical von ihm (Andrew Lloyd Webber) führt der ökumenische Jugendchor Friedrichsdorf 1994 auf? (061) Which musical by him is performed by the ecumenical youth choir Friedrichsdorf in 1994?	$q_2 \rightarrow q_1$	R
Für wie viel Millionen Dollar hat er (Andrew Lloyd Webber) ein Picasso-Gemälde ersteigert? (062) For how much million dollars did he purchase a painting by Picasso at auction?	$q_3 \rightarrow q_1$	R
Wer war er (Al Capone)? (069) Who was he?	$q_2 \rightarrow q_1$	R
Wann wurde er (Al Capone) erschossen? (070) When was he shot?	$q_3 \rightarrow q_1$	R
Wie alt war er (Al Capone), als er seine Familie nach Chicago brachte? (071) How old was he when he brought his family to Chicago?	$q_4 \rightarrow q_1$	R
Wie viele Mitgliedsstaaten hatte die Organisation (die UNESCO) 1995? (074) How many member states did the organization have in 1995?	$q_3 \rightarrow q_1$	R
Wie viele Soldaten hatte die USA während des Krieges (Vietnamkriegs) in den 60er Jahren in Vietnam stationiert? (088) How many soldiers did the USA base in the sixties in Vietnam during the war?	$q_2 \rightarrow q_1$	R
Wer war US-Präsident, als der Krieg (der Vietnamkrieg) zu Ende ging? (090) Who was President of the USA when the war came to an end?	$q_4 \rightarrow q_1$	R
Nenne drei Alben dieser Rockband (Red Hot Chili Peppers). (107) Name three albums of the rock band.	$q_3 \rightarrow \mathbf{a_1}$	M
Von welcher Organisation wurde sie (Audrey Hepburn) als Sonderbotschafterin ernannt? (111) Which organization appointed she as a special ambassador?	$q_3 \rightarrow q_1$	R
An welcher Universität studierte er (Martin Scorsese) 1960 Filmkunst? (114) At which university did Martin Scorsese study cinematography in 1960?	$q_2 \rightarrow q_1$	R
Für wie viele Oscars wurde er (Martin Scorsese) bereits nominiert? (115) For how many Academy Awards he has been already nominated?	$q_3 \rightarrow q_1$	R
Wie heißt der Kriminalkommissar in seinem (Henning Mankells) Roman "Mörder ohne Gesicht"? (134) What is the name of the police inspector in his novel "Faceless Killers"?	$q_2 \rightarrow q_1$	R
Nenne drei Staaten, die das Protokoll (Kyoto-Protokoll) unterzeichnet haben. (153) Name three countries which signed the Protocol?	$q_2 \rightarrow q_1$	R
Wieviele Mitglieder hat die Organisation (die Organisation Pro Familia)? (156) How many members does the organization have?	$q_2 \rightarrow q_1$	R
Wie heißt ihr (Angela Merkels) Bruder? (161) What is the name of her brother?	$q_2 \rightarrow q_1$	R
In welchen japanischen Städten existieren sie (Hanamachi) noch? (165) In which japanese cities do Hanamachi still exist?	$q_3 \rightarrow \mathbf{q_2}$	R
Auf welchem Treffen ließ er (Gerhard Schröder) sich von Chirac vertreten? (174) At which event was he represented by Chirac?	$q_2 \rightarrow q_1$	R
Wie heißt die vierte Ehefrau von ihm (Gerhard Schröder)? (175) What is the name of his fourth wife?	$q_3 \rightarrow q_1$	R
Wie heißen die drei großen Wasserfälle im Canyon (Grand Canyon)? (177) What are the names of the three waterfalls of the Canyon?	$q_2 \rightarrow q_1$	M
Wie hieß ihr (Cate Blanchetts) erster Kinofilm? (185) What is the name of her first cinema film?	$q_2 \rightarrow q_1$	R
Wohin wanderte er (Alfred Hitchcock) 1939 aus? (197) Where did he emigrate to in 1939?	$q_2 \rightarrow q_1$	R
Wie heißt sein (Alfred Hitchcocks) erster amerikanischer Film? (198) What is the name of his first American film?	$q_3 \rightarrow q_1$	R

3 Shallow QA Subsystems

QAP (Question Answering by Pattern matching, see Leveling (2006)) employs pattern matching on a per-sentence basis. For this approach, about 30 classes of questions were defined, based on an analysis of the QA@CLEF questions from 2003 to 2006. These classes correspond to relational triples of the form *<relname>(<keyword>, <answer>)*, where *relname* is the name of a relation, *keyword* is a term describing the question topic, and *answer* is a string representing the answer. QAP returns answers exactly as they occur in the document. Several large resources were utilized to create question-answer pairs for training this method, including the PND data as used in the German Wikipedia (PND – Personennamendatei, see Hengel and Pfeifer (2005)). An entry in the PND data contains information about a famous person such as his/her place of birth (relation *born_in*), date of birth (*born_on*), place of death (*died_in*), date of death (*died_on*), aliases (*has_pseudonym*), and profession (*has_role*). This data can be transformed to represent question-answer pairs. In addition to explicit information, additional question-answer pairs can be derived, e.g. the age at death (*died_at_age*) can be computed from the date of birth and the date of death.

QA in QAP consists of determining the question type (*relname*), extracting the main keywords from the question, and retrieving a set of document sentences containing answer candidates. The patterns corresponding to the relation identified are applied to retrieved sentences. Pattern matching instantiates the answer variables in the patterns. Their values are returned as answer candidates and sent to the MAVE module for validation. QAP was introduced in QA@CLEF 2006, while the following approach (MIRA) is new in QA@CLEF 2007.

MIRA (Modified Information Retrieval Approach for QA, see Leveling (2007)) is a recall-oriented approach to QA based on information retrieval combined with the selection of the most frequent word sequence of the expected answer type. Question processing in MIRA consists of the following steps: The natural language question is tokenized and stopwords are eliminated to identify the keywords. A naïve Bayesian classifier is applied on shallow features of the question such as the first four word forms. The classifier returns a ranking of expected answer types, of which the most probable type is selected. The classes for expected answer type include DEFINITION, LOCATION, MEASURE, ORGANIZATION, PERSON, SUBSTANCE, and TIME. An IR query is created utilizing all morphologic variants of keywords and submitted to the database system. The top 250 documents (sentences) are retrieved and tokenized. The tokens are categorized according to the expected answer types, i.e. named entities are tagged with LOCATION, ORGANIZATION, or PERSON; temporal expressions (dates) are annotated with TIME, and numeric expressions followed by a unit are associated with MEASURE. Answer candidates are then selected by choosing the most frequent word sequences tagged with the expected answer type.

4 Merging Answer Streams by Validation

The answer producers delivered six answer streams that had to be merged by a new component, the answer validator MAVE (Multinet-based Answer VErification). The first MAVE prototype (Glöckner, 2006) originated from the Answer Validation Exercise (AVE) at CLEF 2006 and was later extended to handle not only the basic validation task (i.e. checking the correctness of an answer with respect to a supporting text passage), but answer selection as well (Glöckner et al., 2007). The system accepts streams of validation items composed of the question string, the answer string, and a supporting witness text extracted from the document collection. It then uses deep linguistic processing and logical reasoning for validating the correctness of answers, i.e. by checking if they are verified by the witness texts. This is usually the case if the *hypothesis* expressed by the answer given the question (i.e. by the answer in affirmative form) can be proved from the representation of the witness and from the assumed background knowledge.

In order to gain more robustness, the theorem prover of MAVE is embedded in a feedback loop which subsequently skips literals until a proof of the reduced set of query literals succeeds. The number of skipped literals then serves as a robust indicator for (non)entailment. The system is backed with additional tests for false positives which reject trivial or circular answers. Answer selection needs some more effort because it involves a re-ranking of answers which permits a selection of the *best* one, rather than a clear-cut validation decision. MAVE exploits the aggregated evidence of all witness texts supporting a given answer in order to assign a useful validation score.

Table 2: Results for the German question set from QA@CLEF 2007.

Run	Results				
	# Right	# Unsupported	# Inexact	# Wrong	K1
fuha071DEDE	48	2	4	146	-0.1789
fuha072DEDE	30	2	4	164	-0.3180

The version of MAVE used for filtering the QA@CLEF 2007 results was mostly identical to the system described by Glöckner et al. (2007). The system even reused the error model extracted in these experiments, which determine the probability estimates needed for aggregation. However, there were two main changes compared to the published method: First, extraction of a threshold which makes it possible to reject rather than select the best answer candidate in the case that the evidence for selection of the best answer is still too weak. Second, integration of large lexical-semantic resources (like GermaNet and OpenThesaurus) which allow more flexible inferences.

The current state of the system is detailed by Glöckner (2007b). In particular, the system now avoids several errors which still deteriorated results in QA@CLEF 2007. The improvements achieved in the meantime include an answer type filter which compares the expected answer type of the question and the found answer type; an improved informativeness test for definition questions; combination of the skipped literal count (as a special kind of an *edge overlap* metric) with a simple lexical overlap matcher, etc. These changes resulted in very satisfying performance in the Answer Validation Exercise 2007, though the AVE task of answer selection from a few end results of QA systems is by no means comparable to a realistic problem of answer selection from hundreds of answer candidates, as produced by our six streams for QA@CLEF 2007.

5 Evaluation and Discussion

We submitted two runs for the German monolingual task in QA@CLEF 2007 (see Table 2). The first run was generated from all six answer streams by applying MAVE for answer selection. Noticing the heterogenity of the involved QA systems (i.e. the precision-oriented InSicht system and the shallow QAP and MIRA systems), the second run was compiled from the QAP and MIRA streams only in order to obtain a baseline for the shallow QA subsystems. As noted above, the results dropped in comparison to previous years, mainly because of the addition of Wikipedia and several problems in adapting system components (see Sect. 2.1). In part, the shallow QA subsystems managed to back-up the performance of the deep QA system: 18 correct answers that InSicht did not find came from the shallow streams.

Compared to previous years with positive K1 values, our system somewhat lost the ability to judge its own answers by assigning accurate scores and hence the ability to identify questions with no answers in the collection (NIL questions). This effect was partly due to remaining bugs in the answer validator – an error which disabled applicability of important axioms and an error in the processing of numerals which spoiled results for COUNT and MEASURE questions. Moreover the system used improper parameter settings based on the QA@CLEF tasks of previous years and the characteristics of earlier versions of InSicht, QAP, and MIRA, which were no longer valid due to the extension of the text collection by the German Wikipedia and due to changes in the underlying QA systems to which the error model was not yet adapted.

The dialog treatment was very successful as can be seen in the third column of Table 1. 89.6% of the questions with anaphors were correctly treated by the coreference resolver CORUDIS.

Due to time constraints, patterns for the Wikipedia data were not produced in time by the shallow QA methods. Instead, the patterns created from CLEF-NEWS were utilized for Wikipedia documents as well. As articles sometimes follow a template-like style in the Wikipedia (e.g. for biographical information like the place and date of birth), QAP missed important patterns applicable to reliably find answers to factoid questions in Wikipedia. The shallow QA systems profited from resolving anaphoric references (by the coreference resolver CORUDIS in the WOCADI parser). But for three follow-up questions, an anaphor stayed unresolved. This led to missing bits of information for precise results from the IR phase.

6 Conclusion

The system for QA@CLEF 2007 showed a performance drop compared to 2004, 2005, and 2006. Error analysis hinted at the massive change in the size and type of the document collection caused by the addition of Wikipedia. On the positive side, the system architecture (under the umbrella of IRSAW) matured by the solid integration of two shallow QA subsystems beside the main, deep QA system, InSicht. Finally, the addition of MAVE as an answer validator completed the system by an important component. Some cutting edge QA systems participating in CLEF or TREC employ components fulfilling a similar function.

In the future, the document processing (especially preprocessing and parsing) of the Wikipedia subcollection should be improved by adjusting to some frequent peculiarities of Wikipedia. The successful dialog handling should be tested on more diverse discourse dependency types and structures linking questions and answers.

References

Glöckner, Ingo (2006). University of Hagen at QA@CLEF 2006: Answer validation exercise. In *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop* (edited by Nardi, Alessandro; Carol Peters; and José Luis Vicedo). Alicante, Spain.

Glöckner, Ingo (2007b). University of Hagen at QA@CLEF 2007: Answer validation exercise. In *Results of the CLEF 2007 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2007 Workshop*. Budapest, Hungary.

Glöckner, Ingo; Sven Hartrumpf; and Johannes Leveling (2007). Logical validation, answer merging and witness selection – a case study in multi-stream question answering. In *Proceedings of RIAO 2007 (Recherche d'Information Assistée par Ordinateur – Computer assisted information retrieval), Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*. Pittsburgh, USA: Le Centre de Hautes Etudes Internationales d'informatique Documentaire – C.I.D.

Harabagiu, Sanda; Dan Moldovan; Marius Paşca; Mihai Surdeanu; Rada Mihalcea; Roxana Gîrju; Vasile Rus; Finley Lăcătuşu; Paul Morărescu; and Răzvan Bunescu (2001). Answering complex, list and context questions with LCC's question-answering server. In *Proceedings of TREC-10*, pp. 355–361.

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.

Hartrumpf, Sven (2003). *Hybrid Disambiguation in Natural Language Analysis*. Osnabrück, Germany: Der Andere Verlag.

Hartrumpf, Sven (2006). Extending knowledge and deepening linguistic processing for the question answering system InSicht. In *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria* (edited by Peters, Carol; Fredric C. Gey; Julio Gonzalo; Gareth J. F. Jones; Michael Kluck; Bernardo Magnini; Henning Müller; and Maarten de Rijke), volume 4022 of *Lecture Notes in Computer Science*, pp. 361–369. Berlin: Springer.

Hartrumpf, Sven and Johannes Leveling (2006). University of Hagen at QA@CLEF 2006: Interpretation and normalization of temporal expressions. In *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop* (edited by Nardi, Alessandro; Carol Peters; and José Luis Vicedo). Alicante, Spain.

Hengel, Christel and Barbara Pfeifer (2005). Kooperation der Personennamendatei (PND) mit Wikipedia. *Dialog mit Bibliotheken*, 17(3):18–24.

Leveling, Johannes (2006). On the role of information retrieval in the question answering system IRSAW. In *Proceedings of the LWA 2006, Workshop Information Retrieval*, pp. 119–125. Hildesheim, Germany: Universität Hildesheim.

Leveling, Johannes (2007). A modified information retrieval approach to produce answer candidates for question answering. In *Proceedings of the LWA 2007, Workshop FGIR*. Halle/Saale, Germany: Gesellschaft für Informatik.

Lin, Chuan-Jie and Hsin-Hsi Chen (2001). Description of NTU system at TREC-10 QA task. In *Proceedings of TREC-10*, pp. 406–411.

Oh, Jong-Hoon; Kyung-Soon Lee; Du-Seong Chang; Chung Won Seo; and Key-Sun Choi (2001). TREC-10 experiments at KAIST: Batch filtering and question answering. In *Proceedings of TREC-10*, pp. 347–354.

Vicedo, Jose L. and Antonio Ferrández (2006). Coreference in Q & A. In *Advances in Open Domain Question Answering* (edited by Strzalkowski, Tomek and Sanda Harabagiu), volume 32 of *Text, Speech and Language Technology*, pp. 71–96. Dordrecht: Springer.