University of Hagen at QA@CLEF 2006: Interpretation and Normalization of Temporal Expressions

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

Abstract

The German question answering (QA) system InSicht participated in QA@CLEF for the third time. InSicht realizes a deep QA approach: it builds on full sentence parses, inferences on semantic representations, and matching between semantic representations derived from questions and document sentences. InSicht was improved for QA@CLEF 2006 in the following main areas: temporal expressions are better normalized and temporal deictic expressions are resolved to explicit date representations; the coreference module was extended by a fallback strategy for increased robustness; equivalence rules can introduce negative information that should not occur in document sentences used for answering questions; answer candidates are clustered in order to avoid multiple occurrences of one real-world entity in the answers to a list question; and finally a shallow QA subsystem that produces a second answer stream was integrated into InSicht. The current system is evaluated on the German questions from QA@CLEF 2006. An ablation study indicates which changes had the most positive effects.

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*

General Terms

Experimentation, Measurement, Performance

Keywords

Question answering, Deep semantic processing of questions and documents, Temporal deixis

1 Introduction

The German question answering (QA) system InSicht participated in QA@CLEF for the third time. InSicht realizes a deep (or semantic) approach because it builds on full sentence parses, inferences on semantic representations, and matching between semantic representations derived from questions and document sentences. For QA@CLEF 2005, a coreference resolution module was integrated and several large knowledge bases containing relations between lexical concepts were added (Hartrumpf, 2006). These knowledge bases were derived mainly from nominal compound analyses. InSicht was improved in the following directions for QA@CLEF 2006: temporal expressions are better normalized and temporal deictic expressions are resolved to explicit date representations (Sect. 2.1); the coreference module was extended by a fallback strategy for increased robustness (Sect. 2.2); equivalence rules can now introduce negative information that should not occur in document sentences used for answering questions (Sect. 2.3); answer candidates are clustered in order to avoid multiple occurrences of one real-world entity in the answers to a list question (Sect. 2.4); and finally a shallow QA subsystem that produces a second answer stream was integrated into InSicht (Sect. 2.5).

2 Improvements over the System for QA@CLEF 2005

2.1 Temporal Expressions

As many questions are related to temporal expressions, InSicht was extended to use not just explicit temporal expressions like *before September 18, 2005* and *in 2001* but also deictic expressions like *today* or *2 years ago*. To this end, around 40 MultiNet rules (Helbig (2006) describes the MultiNet semantic network formalism in detail) were developed which resolve temporal deictic expressions into explicit forms. For a wider perspective on *indexicals* (or deictic expressions) in general with their classificatory, relational, and deictic components, see (Nunberg, 1993).

Before the rules can be applied one must determine for each document the date it was published. The special variable *?now*, which is used in many deixis resolution rules, contains this value. The date encoded in the document identifier was taken as a starting point, which was refined (if possible) by looking for a date given at the start of the article text. In 61.4% of the articles, the date could be determined from the text; in the remaining 38.6% (which all came from the FR subcorpus and not the other subcorpora SDA and SPIEGEL), the document identifier was used as the fallback strategy. Currently, the SGML attribute *DATE* (*DT*, *WEEK*) is not used because it is often inaccurate: the FR subcorpus contains only a *WEEK* attribute which corresponds to the Sunday or Monday starting the week. For the SDA subcorpus, the identification method from article text yields the same results as the *DT* attribute.

Figure 1 shows two examples of deixis resolution rules that InSicht applies in the document processing step. The rules are transformation rules that can introduce new relations and delete existing relations (operation: *$delete*). Variables start with a question mark; by convention, node (relation) variables are of the form *?n* (*?r*) followed by an integer, e.g. *?n1* and *?r1* in the first premise term of rule *deixis.übermorgen*. In addition to testing the presence or absence of a relation and testing the value of a node feature (e.g. the MultiNet layer feature CARD), the premise can also contain some predefined declarative predicates (e.g. the *member* predicate in the example rules). On the conclusion side, the value of node variables can be stated by constants or by functional expressions (e.g. functions for calculating with dates like *plus-days*, functions for accessing components of a date like *date-day*, and week-aware rounding functions like *wday-round-down*).

In addition to rules that directly introduce canonical date representations (details below the resolution of days are currently ignored because they seem to be less prominent in questions over newspaper and newswire articles), some preparatory normalization rules were needed. Examples are the adjectives that correspond to deictic adverbs, e.g. the adjective *morgig* corresponding to the adverb *morgen*/'*tomorrow*'.

```
((rule (
   (?r1 ?n1 "übermorgen.1.1") ; temporal adverb referring to the day after tomorrow
   (member ?r1 (ante dur fin strt temp))
→
   ($delete (?r1 ?n1 "übermorgen.1.1"))
   (?r1 ?n1 ?n2)
   (attr ?n2 ?n3) (sub ?n3 "jahr.1.1") (val ?n3 ?n4)
   (card ?n4 (date-year (plus-days ?now 2)))
   (attr ?n2 ?n5) (sub ?n5 "monat.1.1") (val ?n5 ?n6)
   (card ?n6 (date-month (plus-days ?now 2)))
   (attr ?n2 ?n7) (sub ?n7 "tag.1.1") (val ?n7 ?n8)
   (card ?n8 (date-day (plus-days ?now 2)))))
(name "deixis.übermorgen"))

((rule (
   (?r1 ?n1 ?n2)
   (member ?r1 (ante dur fin strt temp))
   (temp ?n1 "past.0")
   (sub ?n2 ?n3)
   (*pmod ?n3 ?n4 ?n5)
   (member ?n5 ("montag.1.1" ...)) ; 'Monday', week days and compounds involving week days
   (refer ?n2 det)
   (member ?n4 ("letzt.1.1" ...)) ; 'last' ...
→
   ($delete (?r1 ?n1 ?n2))
   ($delete (sub ?n2 ?n3))
   (?r1 ?n1 ?n6)
   (attr ?n6 ?n7) (sub ?n7 "jahr.1.1") (val ?n7 ?n8)
   (card ?n8 (date-year (wday-round-down ?now ?n5 ?n4)))
   (attr ?n6 ?n9) (sub ?n9 "monat.1.1") (val ?n9 ?n10)
   (card ?n10 (date-month (wday-round-down ?now ?n5 ?n4)))
   (attr ?n6 ?n11) (sub ?n11 "tag.1.1") (val ?n11 ?n12)
   (card ?n12 (date-day (wday-round-down ?now ?n5 ?n4)))))
(name "deixis.weekday+pmod.past"))
```

Figure 1: Two simplified rules for resolving temporal deixis. The concepts jahr.1.1 ('*year*'), monat.1.1 ('*month*'), and tag.1.1 ('*day*') are attributes used in normalized date representations with MultiNet.

The semantic representations of these adjectives are currently flat in the sense that they are equal to the ones of other operational qualities. Therefore, preparatory normalization rules introduce the temporal interpretation of these adjectives, e.g. for *das morgige Treffen*/'*tomorrow's meeting*' a temporal relation TEMP from treffen.1.1 to morgen.1.1. The newly introduced temporal concepts like morgen.1.1 are then in turn normalized by rules like the rule *deixis.übermorgen.1.1* in Figure 1. Note that the rules can be easily transferred to other languages because they work on semantic representations of the MultiNet formalism. Only some lexical concepts (like week days) must be translated; if one already has a mapping between the involved lexicons, rules can be *translated* automatically for the use in another language.

For some sentences, the deictic expression has been introduced by the prepositional phrase interpretation. For example, the *vor*-PP in the noun phrase *das Konzert vor 20 Tagen* ('*the concert 20 days ago*') introduces an artificial concept now.0 (besides a relation to the semantic representation of *20 days*) that needs to be resolved by using the notion of publication date mentioned above.

The ten most successful deixis resolution rules (i.e. the ten rules that fired most often for semantic networks of document sentences) are listed in Table 1. The percentages of rule application for successful (or

Table 1: Description of rules for resolving temporal deixis

Rule	Natural language example
deixis.weekday.past	*the debate on last Monday*
deixis.today	*The book is published today.*
deixis.weekday.nonpast	*The group will meet on Friday.*
deixis.monthname.past	*The law was passed in August.*
deixis.year+opminus	*A storm came 2 years ago.*
deixis.tomorrow	*The president will arrive tomorrow.*
deixis.monthname.nonpast	*The team will win in December.*
deixis.today's	*today's news paper*
deixis.year+prop.past	*the revenues in the past year*
deixis.yesterday	*The peace treaty was signed yesterday.*

Table 2: Statistics on rule application for temporal deixis

Rule	Application (% of complete semantic networks)			
	FR	SDA	SPIEGEL	all
deixis.weekday.past	1.4067	6.4929	0.0658	3.5824
deixis.today	0.7950	0.5791	0.8363	0.7013
deixis.weekday.nonpast	0.4939	1.0044	0.0515	0.6867
deixis.monthname.past	0.1807	0.4948	0.1622	0.3204
deixis.year+opminus	0.2301	0.2291	0.2759	0.2335
deixis.tomorrow	0.1179	0.3692	0.0357	0.2241
deixis.monthname.nonpast	0.1748	0.2489	0.0984	0.2018
deixis.today's	0.1978	0.1479	0.0714	0.1649
deixis.year+prop.past	0.1264	0.1512	0.0964	0.1351
deixis.yesterday	0.2471	0.0060	0.0178	0.1196
one or more rules	3.7937	9.6008	1.8837	6.2463

complete) parses from the subcorpora in Table 2 reflect different text types. The SDA subcorpus (newswire text) contains short texts with a large percentage of temporal deixis (almost 1 out of 10 sentences). The reference to past week days is extremely frequent because the current day of newswire text is very prominent to the reader or hearer. In the daily newspaper subcorpus (FR), the distribution of firing rules is more uniform than in the SDA subcorpus. And finally, the weekly newspaper subcorpus (SPIEGEL) shows the smallest rule activation rates because a weekly newspaper has longer articles and cannot clearly refer to yesterday, certain week days, or similar concepts because it is unknown on what day the typical reader of a weekly newspaper will read the article. Table 3 lists some deixis resolution examples that helped InSicht to find an answer.

For related work, see for example Pan and Hobbs (2005) who briefly describe a rule-based translation from the output of an English semantic parser to an OWL-Time representation for temporal expressions, especially temporal aggregates. Mani and Wilson (2000) cover with their system similar cases like the deixis resolution rules in InSicht but their system looks for lexical and syntactic clues based on part-of-speech tags and does not build on semantic representations derived from full syntactico-semantic parsing like InSicht does. Schilder and Habel (2001) describe a system for German financial news that extracts temporal information based on a part-of-speech tagger and specialized finite state transducers. In contrast to Mani and Wilson (2000), Schilder and Habel (2001) and InSicht's approach also take into account the semantics of prepositional phrases (see (Hartrumpf et al., 2006) for details on the underlying prepositional phrase interpretation in InSicht's parser).

Table 3: Successful deixis resolution. Note that the official question qa06_172 contains a small grammatical error, which InSicht's parser ignored.

Question	Document sentence	Answer
In welchem Jahr starb Charles de Gaulle? (qa06_079)	*Frankreichs Staatschef Jacques Chirac hat die Verdienste des* vor 25 Jahren *gestorbenen Generals und Staatsmannes Charles de Gaulle gewürdigt.* (SDA.951109.0236)	1970
'*In which year did Charles de Gaulle die?*'	'*France's chief of state Jacques Chirac acknowledged the merits of general and statesman Charles de Gaulle, who died* 25 years ago.'	1970
An welchen[!] *Tag haben Jordanien und Israel den Friedensvertrag unterzeichnet?* (qa06_172)	*Israel und Jordanien haben* am Mittwoch *im Grenzgebiet zwischen beiden Ländern einen Friedensvertrag unterzeichnet.* (SDA.941026.0140)	*am 26.10.1994*
'*On which day did Jordan and Israel sign the peace treaty?*'	'*Israel and Jordan signed a peace treaty in the border area between both countries* on Wednesday.'	'*on 26th October 1994*'

2.2 Robust Coreference Resolution

As only 45% of the texts received a partition from the coreference resolution module CORUDIS (Hartrumpf, 2006) (mainly because of parameter settings that allow efficient coreference resolution for the 277,000 texts (without duplicates) in the QA@CLEF corpus), a fallback strategy was added to InSicht. If no partition of mentions (markables) can be found by CORUDIS, a *fallback partition* of mentions is calculated as follows. All pronouns are resolved to their most likely antecedents (as estimated by CORUDIS); all other mentions are ignored, i.e. each of these mentions ends up in a singleton set, which is an element of the fallback partition.

2.3 Negative Information for Network Matching

The open world assumption (OWA) in semantic network representations can be problematic if the *query expansion* step produces semantic networks that are similar to the original question network but a little bit less specific. For example, InSicht employs a rule named *drop.first_name* that drops the first name of a person if the last name is also in the question (see Figure 2). Without additional measures, the resulting additional query network will also match a document network containing a first name, even though this first name might differ from the one in the original question.

Therefore rules can introduce negative information. In the given example of rule *drop.first_name*, the conclusion specifies the term (*no-rel ?n1* "vorname.1.1"), which means that the node variable of the person (where the first name has been deleted by this rule, *?n1*) should not contain any (direct or indirect) relation to a first name (vorname.1.1) in a document network.

2.4 Answer Clustering

The QA system InSicht worked without investigating the relationship between answer candidates in previous years; it only combined identical answer candidates into one answer candidate with an increased frequency score that influenced answer selection. But as InSicht was extended to answer list questions, clustering became unavoidable because otherwise the system could give many answers (meant as different

```
((rule (
   (sub ?n1 ?)
   (attr ?n1 ?n2) (sub ?n2 "nachname.1.1") (val ?n2 ?)
   (attr ?n1 ?n3) (sub ?n3 "vorname.1.1") (val ?n3 ?n4)
→
   ($delete (attr ?n1 ?n3))
   ($delete (sub ?n3 "vorname.1.1"))
   ($delete (val ?n3 ?n4))
   (no-rel ?n1 "vorname.1.1")))
(name "drop_first_name")
(quality 0.6))
```

Figure 2: A transformation rule that introduces negative information. The variable *?* is the anonymous variable.

entities on the answer list) which all refer to the same entity. For example, if one asks for the president of a country, one could receive one person in many forms: with all first names, with some first names, with initials, only last name, etc.

Non-identical answer candidates *a* and *b* are clustered together if *a* is a substring of *b*, possibly with additional interleaving characters. For example, the answer strings *Peter Schmidt* and *P. Schmidt* will be in the same answer cluster. This surface definition for a clustering condition will be replaced by a semantics based definition: if *a* is more specific than *b*, cluster *a* and *b* with representative *a*; if *b* is more specific than *a*, cluster *a* and *b* with representative *b*; if *a* and *b* are unifiable, cluster *a* and *b* with the unification of *a* and *b* as the representative.

2.5 Shallow QA Subsystem

WOCADI produces a partial (and not a complete) semantic network for about 50% of all sentences in the German QA@CLEF corpus. For example, WOCADI can rarely analyze sentences containing grammatical errors, spelling errors, or conflated sentence parts originating from erroneous document preprocessing. Thus, InSicht will not be able to find many of those answers appearing in malformed sentences only. Therefore, we experimented with an additional shallower approach which can be seen as producing an additional answer *stream* (see (Ahn et al., 2005) for a QA system with many answer streams).

We applied a sentence boundary detector and a tokenizer to the corpus and split all documents into single sentences. Then, question-answer pairs for the QA task in previous years were extracted from the MultiEight corpus and augmented manually.

These pairs were fed into a standard IR system to determine candidate sentences containing an answer. The answer candidates were then processed as follows: Words from closed word categories were annotated with part-of-speech information, keywords in the query were replaced with symbols representing their type (name, upper-case word, lower-case word, number) and the answer string was replaced with an answer variable representing one or more words. We then formed several patterns from this substitution, including a maximum of 5 tokens left and right of the answer variable.

In this shallow approach, question answering consists of 5 steps: 1. Perform relevance-ranked IR on the corpus and determine the top 250 ranked sentences. 2. Apply pattern matching with the answer patterns extracted. 3. If a match is found, the answer variable contains the string representing the answer. 4. Answer candidates are validated using a simple heuristic: all answers starting with punctuation marks or containing words from a closed word category only are eliminated. 5. Answers are ranked by cumulative frequency and the most frequent answer is chosen.

Table 4: Results for the German question set from QA@CLEF 2006. W_n (W_{nn}) is the number of wrong NIL (non-NIL) answers.

Setup and modification	Results					
	# Right	# Unsupported	# Inexact	# W_{nn}	# W_n	K1
InSicht (= FUHA061dede)	62	4	0	3	129	0.1799
− deixis resolution (Sect. 2.1)	62	2	1	3	130	
− robust coreference resolution (Sect. 2.2)	61	4	0	3	130	
− negative information (Sect. 2.3)	62	4	0	5	127	
− answer clustering (Sect. 2.4)	60	3	3	3	129	
+ shallow QA (Sect. 2.5, = FUHA062dede)	65	4	1	3	125	0.1895

3 Run Description and Evaluation

The current QA system has been evaluated on the German questions from QA@CLEF 2006. The shallow approach found 17 answers in total for the 198 questions assessed for QA@CLEF 2006. 13 of them were also found by the deep approach (run FUHA061dede in Table 4), but 3 of them were new and correct and 1 was new and inexact (because it omitted a phrase-final acronym). This was our first attempt to combine the deep processing with a shallower method and it leaves many chances for further improvements. However, it already shows that a combination of processing methods will further improve performance for the IRSAW system, a more general retrieval system which will encompass the QA system InSicht. The combination of the deep approach and the shallow approach corresponds to run FUHA062dede in Table 4.

The result rows that are preceded by a minus sign (−) in Table 4 are from an ablation study: each extension from Sect. 2 was omitted in an evaluation run in order to see the impact on system performance. All extensions had positive effects, but their statistical significance will probably appear only when evaluating on much more questions.

4 Conclusion

The extensions of the QA system InSicht after QA@CLEF 2005 (deixis resolution, increased robustness of coreference resolution, use of negative information during network matching, clustering of answers) improved system results. To investigate the statistical significance of these improvements, the tests should be repeated on much more questions, e.g. the QA@CLEF question sets from 2003, 2004, and 2005.

In the future, more deixis resolution rules need to be written, e.g. for resolving names of fixed and floating holidays (*during Christmas* and *21 days after Easter Sunday*), season names (*during this Summer*), and vague temporal expressions (*at the end of this year*). Temporal expressions are currently linked to individual sentences. Propagating temporal expressions from a sentence to parts of its cotext (e.g. by event ordering) will increase the amount of available temporal information dramatically.

References

Ahn, David; Valentin Jijkoun; Karin Müller; Maarten de Rijke; Stefan Schlobach; and Gilad Mishne (2005). Making stone soup: Evaluating a recall-oriented multi-stream question answering system for Dutch. In *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004* (edited by Peters, C.; P. Clough; J. Gonzalo; G. J. F. Jones; M. Kluck; and B. Magnini), volume 3491 of *Lecture Notes in Computer Science*, pp. 423–434. Berlin: Springer.

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, Revised Selected Papers* (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; Hermann Helbig; and Rainer Osswald (2006). Semantic interpretation of prepositions for NLP applications. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pp. 29–36. Trento, Italy.

Helbig, Hermann (2006). *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer.

Mani, Inderjeet and George Wilson (2000). Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 69–76. Hong Kong.

Nunberg, Geoffrey (1993). Indexicality and deixis. *Linguistics and Philosophy*, 16:1–43.

Pan, Feng and Jerry R. Hobbs (2005). Temporal aggregates in OWL-Time. In *Proceedings of the 18th Florida Artificial Intelligence Conference (FLAIRS-05)*, pp. 560–565.

Schilder, Frank and Christopher Habel (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of the ACL-2001 Workshop on Temporal and Spatial Information Processing*, pp. 65–72. Toulouse, France.