Increasing the coverage of answer extraction by applying anaphora resolution

Jori Mur

Humanities Computing
Faculty of Arts
University of Groningen
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen
j.mur@rug.nl

Abstract

Off-line answer extraction using patterns is a technique for corpus based Question Answering (QA) that has proven to be very effective. The results typically show a high precision score. However, the main problem with this technique is the lack of coverage of the extracted answers. One way to increase the coverage is to apply anaphora resolution. Using anaphora resolution we can turn information-poor extractions (e.g. she is the queen of Holland) into potential answers (e.g. Beatrix is the queen of Holland). In this paper we show that by applying a simple anaphora resolution technique the number of facts extracted from a Dutch newspaper corpus increased with more than 50% and although precision went down the increase in recall had a positive effect on the performance of a state-of-the art QA system.

Povečanje pokritja luščenja odgovorov z uporabo razrešitve anafore

Off-line luščenje odgovorov z uporabo vzorcev je metoda na korpusu temelječega sistema za odgovarjanje na vprašanja, ki se je izkazala za zelo učinkovito. Rezultati običajno dosegajo visoko točnost. Vendar pa je glavni problem metode pomanjkljivo pokritje izluščenih odgovorov. Ena od možnosti za povečanje obsega je razrešitev anafore. Z uporabo razrešitve anafore lahko pretvorimo slabo informativne izluščene izraze (npr. ona je nizozemska kraljica) v možne odgovore (npr. Beatrix je nizozemska kraljica). V članku prikažemo, da se z uporabo preproste tehnike razrešitve anafore število izluščenih dejstev iz nizozemskega časopisnega korpusa poveča za več kot 50%. Čeprav se točnost zmanjša, pa ima povečanje pokritja pozitivne učinke na delovanje sodobnega sistema za odgovarjanje na vprašanja.

1. Introduction

There is a need for tools which help users to find the information they are looking for. Search engines such as Google overcome this need by presenting on a user's query a ranked list of links to relevant documents. However, sometimes a user simply has a question and what he wants is an answer. He needs a system which analyses the relevant documents for him and which only returns the answer, rather than a list of documents, thus saving him a lot of time. Question Answering (QA) systems respond to this need. The task of a question answering system is to retrieve answers to questions posed in natural language, given a collection of documents.

A typical approach to question answering is the following: a question is analysed by a question classifier module that assigns a certain type to the question. Example types are 'function' (Who is the president of the United States?) or 'location of birth' (Where was Vincent van Gogh born?). Often this module also determines the expected answer type. For example, 'person name' or 'location name'. Keywords are selected from the question and they are fed into a document retrieval module. This module identifies relevant articles or paragraphs that are likely to contain the answer. Then terms of the same type as the type of the expected answer are extracted. The system uses further clues to rank the candidate answers. The answer ranked first is returned to the user.

A second approach, which some QA research teams have added as a module to their basic system is the technique of off-line answer extraction. Off-line methods have proven to be very effective in QA (Fleischman et al., 2003; Jijkoun et al., 2004; Bouma et al., 2005a). Before actual questions are known, a corpus is exhaustively searched for potential answers to questions of a specific question type (capital, abbreviation, inhabitants, year of birth, ...). Highly precise relations are extracted from the corpus off-line and stored as an answer repository for quick and easy access. This method typically results in a very high precision score.

However, one major drawback is the lack of coverage of the extracted answers. If the module finds an answer it is usually the correct one, but the module finds too few answers in general. Jijkoun et al. (2004) have shown for English that extraction patterns defined in terms of dependency relations can lead to significant improvements in recall over systems based on regular expression pattern matching. Yet, lack of coverage remains an issue (Tjong Kim Sang, 2005).

The aim of this paper is to address the low coverage problem of answer extraction by incorporating an anaphora resolution module into an answer extraction system. The task of anaphora resolution is to determine for a reference in the text the entity to which it refers, the antecedent. Patterns are often defined in such a way, that answers are only extracted when they appear within the matching sentence, whereas applying anaphora resolution will result in the extraction of information that is referenced anaphorically as well. The following example illustrates this:

(1) **Question:** Who is the Queen of Holland?
Text: Beatrix was invited to speak before the European Parliament. The Queen of Holland emphasised in her speech the equality of everyone

who lives in Europe.
Answer: Beatrix

If we know that *The queen of Holland* in the second sentence refers back to *Beatrix* in the first sentence, we can answer the question.

For our experiments we used a Dutch answer extraction module, which is part of our Dutch QA system. We developed an anaphora resolution system based on this module. The results indicate that anaphora resolution can be used effectively to increase the coverage of off-line answer extraction. Many more facts were extracted and in spite of a low precision score for the anaphora resolution system and the limited number of question types to which it was applied, the performance of the QA system overall increased as well.

The remainder of this paper has been organised in the following way. In the next section we provide details about the extraction methods. The anaphora resolution system is described in section 3. Section 4 contains a description of our experiments and section 5 shows the results. Finally, we discuss research of others related to our work in section 6 and we conclude in section 7.

2. Off-line answer extraction

In this section we describe the extraction method we used to create fact bases.

Since we want to define patterns in terms of dependency relations, we parsed the whole corpus. For this task we used the Alpino parser, a wide-coverage dependency parser for Dutch (Malouf and van Noord, 2004). Malouf and van Noord (2004) show that the accuracy of the system, when evaluated on a test-set of 500 newspaper sentences, is over 88%, which is in line with state-of-the-art systems for English. The Alpino system also incorporates a Named-Entity classifier to recognise and classify proper names.

The dependency analysis of a sentence gives rise to a set of dependency relations of the form $\langle \texttt{Head, Rel, Dep} \rangle$, where `Head` is the root form of the head of the relation, and `Dep` is the head of the constituent that is the dependent. `Rel` is the name of the dependency relation. For instance, the dependency analysis of sentence (2-a) is (2-b).

(2) a. Amsterdam telt 800.000 inwoners (*Amsterdam counts 800,000 inhabitants*)

b. $$\left\{ \begin{array}{l} \langle \texttt{tel, subj, amsterdam} \rangle, \\ \langle \texttt{tel, obj, inwoners} \rangle, \\ \langle \texttt{inwoners, det, 800.000} \rangle \end{array} \right\}$$

The module also accounts for syntactic variation. For instance, the subject of an active sentence may be expressed as a PP-modifier headed by *door* (*by*) in the passive. In (Bouma et al., 2005b) this module is described in more detail.

A dependency pattern is a set of (partially underspecified) dependency relations as in (3).

(3) $$\left\{ \begin{array}{l} \langle \texttt{tel, subj, } \mathbf{\texttt{<LOCATION>}} \rangle, \\ \langle \texttt{tel, obj, inwoners} \rangle, \\ \langle \texttt{inwoners, det, } \mathbf{\texttt{<NUMBER>}} \rangle \end{array} \right\}$$

To get a clear picture of the impact of anaphora resolution on the off-line construction of knowledge bases, we selected 12 question types that we expect to benefit from anaphora resolution.[1]

They are shown in table 1. We extracted for each type the number of facts listed in the second column of table 1 using the basic patterns, i.e. without applying anaphora resolution.

For our experiments we adjusted the patterns. We replaced the slot for the named entity with a slot for a pronoun. For instance, the pattern from the example above is changed into the following:

(4) $$\left\{ \begin{array}{l} \langle \texttt{tel, subj, } \mathbf{\texttt{<Pronoun>}} \rangle, \\ \langle \texttt{tel, obj, inwoners} \rangle, \\ \langle \texttt{inwoners, det, } \mathbf{\texttt{<NUMBER>}} \rangle \end{array} \right\}$$

It will match the parse of a sentence such as (5):

(5) Het telt slechts 1.000 inwoners.
(*It counts only 1,000 inhabitants*)

Similarly, we adjusted the patterns to match sentences with a definite noun. We considered noun phrases preceded by a definite determiner as definite noun phrases.

In the next section we describe the anaphora resolution technique we used to resolve the pronouns and definite NPs that we found with the patterns.

For the question types *capital* and *function* the anaphor could be part of the fact we want to extract. An example for the function type is given in example (1) on page 1. We extract both the antecedent *Beatrix* and the anaphor *The queen of Holland*.

For the other types and for the pronoun patterns we extract the antecedent and the answer terms, but not the anaphor. For example, if we should encounter the following text: *Beatrix was Juliana's first child. The Queen of Holland was born in 1938.* then we extract the antecedent *Beatrix* and the answer term *1938*.

3. Anaphora resolution

The pronouns and definite NPs we found during the pattern matching process have to be linked to the correct antecedent in order to extract complete and meaningful facts for the tables. We developed for both types of anaphoric NPs separate but similar techniques as described in the following two sections.

3.1. Resolving definite NPs

Our strategy for resolving definite NPs is based on knowledge about the categories of named entities, so-called instances (or categorised named entities). Examples are *Van Gogh* IS-A *painter*, *Seles* IS-A *tennis player*. Since the whole corpus was parsed we were able to acquire instances by scanning the corpus for apposition relations and predicate complement relations[2]. The apposition relation holds between *Van Gogh* and *painter* in:

[1] In total we defined 22 question types. The remaining types are: abbreviation, currency, date, firstname, location, measure, result, definition, which, what

[2] We limited our search to the predicate complement relation between named entities and a noun and excluded examples with

Question Type	# of facts	Clarification
Age	21669	Who is how old
Location of Birth	776	Who was born where
Date of Birth	2358	Who was born when
Capital	2220	Which city is the capital of which country
Age of Death	1160	Who died at what age
Date of Death	1002	Who died when
Cause of Death	3204	Who died how
Location of Death	585	Who died where
Founder	741	Who founded what when
Function	58625	Who full fills what function in life
Inhabitants	823	Which location contains how many inhabitants
Winner	334	Who won which Nobel prize when

Table 1: Question types for which we defined patterns together with the number of facts we extracted for each type

(6) Van Gogh, the famous Dutch painter.

We only consider the head of the definite NP for the instance list. And here is an example of a predicate complement relation:

(7) Van Gogh is a famous Dutch painter.

We extracted around 1 million appositions (tokens) and 0.3 million predicate complements (tokens) resulting in 1 million types overall.

Our strategy is as follows: We scan the left context of the definite NP for named entities from right to left (i.e. the closest named entity is selected first). For each named entity we encounter, we check whether it occurs together with the head of the definite NP as a pair on the instance list. If so, the named entity is selected as the antecedent of the NP. As long as no suitable named entity is found we select the next named entity and so on until we reach the beginning of the document. In a previous study (Mur and van der Plas, to appear) it was shown that this strategy leads to high precision, but low recall. So we decided to implement a fall back mechanism: if no suitable named entity is found, i.e., no named entity is found that forms an instance pair with the head of the definite NP, we select simply the first preceding named entity.

In order to explain our strategy for resolving definite NPs we will apply it to the next example:

He was the opponent of the quiet Ivanisevic in December 1995. Todd Martin who defeated the local hero Boris Becker a day earlier, was beaten by the 26-year old Croatian during the finals of the Grand Slam Cup in 1995 [...].

In the example above, the left context of the NP *the 26-year old Croatian* is scanned from right to left. The named entities *Boris Becker* and *Todd Martin* are each selected before the correct antecedent *Ivanisevic*. Neither *Boris Becker* nor *Todd Martin* is found in an instance relation with *Croatian*, so they are put aside as unsuitable candidates. Then *Ivanisevic* is selected and this candidate *is* found to be on the instance list with *Croatian*, so *Ivanisevic* is taken as the antecedent of *Croatian*. The fact `Ivanisevic, 26-year old` is added to the Age table.

3.2. Resolving Pronouns

We applied a similar technique for resolving pronouns. The pronouns we tried to resolve were the nominative forms of the singular pronouns *hij* (he), *zij/ze* (she), *het* (it) and the plural pronoun *zij/ze* (they). We chose to resolve only the nominative case, as in almost all patterns the slot for the name was the slot in subject position. The number of both the anaphor and the antecedent was determined by the number of the main verb.

Since we find the anaphors by matching patterns, we knew the named entity (NE) tag of the antecedent. For example, if we match a pattern defined for the location-of-birth type, we are looking for a person, if we match a pattern defined for the capital type, we are looking for a location and so on.

Again we scan the left context of the anaphor (now a pronoun) for named entities from right to left. We implemented a preference for proper nouns in the subject position. They were analysed before the other proper nouns in the same sentence. For each named entity we encounter, we check whether it has the correct NE-tag and whether its number corresponds to the number of the pronoun. If so and if it concerns a non-person NE-tag, the named entity is selected as the antecedent. If we are looking for a person name as the antecedent, we have to do another check to see if the gender of the name corresponds to the gender of the pronoun. To determine the gender of the selected name we created a list of boy's names and girl's names by downloading such lists from the Internet[3]. The female list contained 12,691 names and the male list 11,854 names. To be accepted as the correct antecedent, the proper name should not occur on the name list of the opposite sex of the pronoun.

After having resolved the anaphor, the fact was added to the appropriate table.

negation

[3]http://www.namen.info, http://www.voornamenboek.nl, http://www.babynames.com and http://prenoms.free.fr

baseline	pronouns	definite nouns	total anaphora
93,497 (86%)	+3,915 (40%)	+47,794 (33%)	+51,644 (34%)

Table 2: Number of added fact tokens (precision)

4. Experiment

The aim of the experiment is firstly to determine whether anaphora resolution on definite NPs and pronouns helps to acquire more facts and secondly to investigate if it improves the performance of a state-of-the-art QA system. To this end, we first create tables using the patterns based on anaphora resolution techniques and we compare these tables to the tables created by using the baseline patterns (i.e. the patterns that extract facts in a straightforward way). For both extraction modules we randomly selected a sample of around 200 extracted facts and we manually evaluated these facts on the following criteria:

- correctness of the fact;
- and in the case of anaphora resolution, correctness of the selected antecedent.

Secondly, we evaluated both extraction modules as part of a QA system. We measured the performance by counting how many questions were answered correctly.

4.1. Corpus and parser

We apply our answer extracting techniques to the Dutch CLEF corpus. This corpus is used in the annually organised CLEF evaluation track for Dutch question answering systems. It consists of newspaper articles from 1994 and 1995, taken from the Dutch daily newspapers *Algemeen Dagblad* and *NRC Handelsblad*. The corpus contains about 78 million words. The whole collection was parsed automatically using the Alpino parser described in section 2.

4.2. QA system and questions

For the experiments we used an open-domain corpus-based Dutch QA system, Joost (Bouma et al., 2005a). It achieved a score of 49,5% on the question answering track of CLEF-2005, the best result for the Dutch track. The system implements an IR-based approach as well as a table-lookup strategy. An incoming question is analysed and assigned to one of the predefined question types. If the type is one of the twelve types listed in table 1, the table lookup mechanism identifies knowledge bases where answers to the question can potentially be found. It uses keywords from the question to identify relevant entries in the selected knowledge bases and extracts candidate answers. Finally, the QA system re-ranks and sanity checks the candidates and selects the final answer. If the question was of a different type than the twelve listed above, an answer is found by the IR-based technique. We performed our experiments with the 200 Dutch questions of the CLEF-2005 data set.

5. Results

The numbers of extracted facts for each method are given in table 2. Between brackets you see the precision score of the extracted facts. 65 facts were extracted by applying both pronoun resolution and definite noun resolution. These facts are typical founders facts: *He founded the organisation*, where both *He* and *organisation* are anaphoric. They are included in the number of facts found by the pronoun patterns as well as by the number of facts found by the definite noun patterns.

The number of facts we extracted by the pronoun patterns is quite low. We did a corpus investigation on a subset of the corpus which consisted of sentences containing terms relevant to the 12 selected question types[4]. In only 10% of the sentences one or more pronouns appeared. This result indicates that is it not surprising that we only extracted 4% more compared to the baseline.

The precision of the new facts was a bit disappointing. In (Mur and van der Plas, to appear) we reported a very high precision for the facts added by applying definite noun resolution. There was a difference in precision between the original and the expanded tables of only 1%. The difference with the method used in the current experiment lies in the fact that we did not use a fall back method in the previous study. We resolved the anaphor if and only if an instance relation was found between the anaphoric NP and the candidate antecedent. There is also a difference in evaluation. In (Mur and van der Plas, to appear) we evaluated the tables containing all the facts together. In this experiment we evaluated them separately.

Nevertheless, if we assume this estimation of the precision to be correct we have added 17,559 valid facts to the original tables.

Since we were interested in the increase of coverage we also calculated the number of additional fact *types* we found with the new patterns, listed in table 3. If we had only used the pronoun patterns we would have found 3,627 new facts. On the other hand, if we had only used the definite noun patterns we would have found 35,687 new facts. Using both we extracted 39,208 additional facts.

baseline	pronouns	definite nouns	both anaphora
64,627	+3,627	+35,687	+39,208

Table 3: Number of added fact types

We also wanted to know what the effect of the extended tables would be on the performance of a state-of-the-art QA system. The results are shown in row (1) of table 4. Clearly, the low precision score of the added facts did not hurt performance. We believe that this effect is due to frequency counts. Incorrect answers are typically outnumbered by correct ones. In total two more questions are answered correctly. In fact, three more questions were answered cor-

[4] terms such as "geboren" (*born*), "stierf" (*died*), "hoofdstad" (*capital*) etc.

rectly, but for another question now an incorrect answer was found. This was due to an incorrectly chosen antecedent.

Further investigation showed that only 40 questions were assigned one of the twelve question types selected for anaphora resolution. Improvement was therefore only possible for those questions. Looking at this subset of questions we see an increase of 5% in performance.

	baseline	anaphora patterns
(1) Total	103/200 (51.5%)	105/200 (52.5%)
(2) 12 types	26/40 (65.0%)	28/40 (70.0%)

Table 4: Number of questions answered correctly

However, besides the low precision score for the anaphora resolution mechanism and the limited number of questions that fell into one of the twelve selected question types there was another issue that possibly caused the small improvement for QA. Question 107 in the question set was as follows: `Wie was piloot van de missie die de astronomische satelliet, de Hubble Space Telescope, repareerde?` (*Who was the pilot of the mission that repaired the astronomic satellite, the Hubble Space Telescope?*). The answer we found was extracted from a sentence in the Algemeen Dagblad of September 19th, 1994 which was formulated as follows: `Bowersox was piloot van de missie die de astronomische satelliet, de Hubble Space Telescope, repareerde.` (*Bowersox was the pilot of the mission that repaired the astronomic satellite, the Hubble Space Telescope.*).

In (Magnini et al., 2004) the authors claim that they created the questions independently from the document collection, thus avoiding any influence in the contents and in the formulation of the queries. However, the example above suggests otherwise. If questions are re-formulations of sentences in the newspaper corpus such as question 107, then it is not surprising that anaphora resolution has little effect.

6. Related work

In last QA tracks of TREC and CLEF (2005) 30 and 24 systems were evaluated respectively. After reviewing these systems, we can notice that only few systems model some co-reference relations between entities in the documents (Schone et al., 2005; Jiangping et al., 2005; Hartrumpf, 2005; Neumann and Sacaleanu, 2005; Laurent et al., 2005).

Schone et al. (2005) apply a symbolic method which tries to resolve pronouns and draw associations between definite NPs. This has a small positive effect on the performance of their QA system. Hartrumpf (2005)'s error analysis of his results for the QA track of CLEF 2004 indicated that the lack of co-reference resolution was a major source of errors. Therefore he incorporated a co-reference resolution system. This system, called CORUDIS, combines syntactico-semantic rules with statistics derived from an annotated corpus. Its results show an F-score of 66% for handling coreference relations between all kinds of NPs (e.g. pronouns, common nouns and proper nouns). The improvements for the QA system obtained by incorporating CORUDIS were unfortunately not significant due to the limited recall value of the co-reference resolution system. In other cases benefits of applying these reference techniques have not been analysed and measured separately.

There are also some systems from earlier years that have evaluated the contribution of reference resolution to the performance of their QA systems. (Watson et al., 2003; Stuckardt, 2003; Mollá et al., 2003). The earliest approaches that evaluate the contribution of reference resolution to QA are by Morton (2000) and Vicedo and Ferrández (2000a).

Morton (2000)'s approach models identity, definite NPs and non-possessive third person pronouns. For pronoun resolution and common noun resolution, he uses a set of features and a collection of annotated data to train a statistical model. For the resolution of coreferent proper nouns simple string-matching techniques were applied.He reports a small improvement, but his results do not quantify the effect of co-reference resolution effectively, since his baseline system includes terms from surrounding sentences.

Vicedo and Ferrández (2000a) analysed the effects of applying pronominal anaphora resolution to QA systems. They apply a knowledge-based approach, dividing the different kinds of knowledge (e.g. pos-tags,syntactic knowledge and morphological knowledge) into preferences and restrictions. Both the restrictions and the preferences are used to discard candidate antecedents. Contrary to our results their outcomes show a great improvement in QA performance. This difference in results can be explained by several aspects of both our experiments.

First, their anaphora resolution system achieved a higher success rate, 87% for Spanish and 84% for English. Our results could be improved. Hoste and Daelemans (2005), for example, reached a precision score for pronouns around 65% and for common nouns around 48% using a machine learning approach for Dutch co-reference resolution.

Second, they also applied anaphora resolution for query terms that are referenced pronominally in the target sentence (sentence containing the correct answer). We only looked at possible answers that were realised anaphorical.

Third, the authors consider an answer to a question to be correct if it appeared into the ten most relevant sentences returned by the system for each question, while we only evaluated the answer ranked first.

And last but not least, their experiment differs from our experiment in that they created their own question set. These questions were known to have an answer in the document collection. Moreover, for more than 50% of these questions the answer or a term in the query was referenced pronominally in the target sentence. This percentage depends on the corpus and the question set and it was probably lower for our data set. The authors define a well-balanced question set as a set that would have a percentage of target sentences that contain pronouns similar to the pronominal reference ratio of the text collection that is being queried. However, most question sets made available by the well-known evaluation fora TREC and CLEF seem to be not that well-balanced according to the definition of Vicedo and Ferrández (2000a). On the other hand, does such a well balanced question set represent a typical set of question set asked by users? It needs further investigation to decide what makes a good question set for the evaluation of the contribution of anaphora resolution to QA.

Vicedo and Ferrández (2000b) participated in the QA track of TREC 2000. The results achieved there were more similar to ours. Application of pronominal anaphora resolution produced only a small benefit, around a 1%. The authors argue there are two main reasons for this result. First, they noticed that the number of relevant sentences involving pronouns is very low. That is in line with our findings. Second, the authors observed that there were a lot of documents related to the same information: sentences in a document that contain the right answer referenced by a pronoun, can also appear in another document without pronominal anaphora. This observation affirms that further investigation to the question set and corpus is needed.

7. Conclusions and Future work

We can conclude from our results that applying anaphora resolution is a way to improve the coverage of answer extraction. In our experiments it resulted in an improvement of the performance of a state-of-the-art QA system for Dutch, in spite of three impediments. Firstly, the precision score for the anaphora resolution was quite low. Secondly, only a limited number of questions was assigned one of the twelve question types for which we applied anaphora resolution. Thirdly, it seemed that questions could be re-formulations of sentences in the corpus.

In the future we should investigate what happens if we improve the anaphora resolution technique and if the domain of question types on which anaphora resolution is applied is broadened. In addition, we need to examine the impact of certain corpora and question sets on the evaluation of the contribution of anaphora resolution to QA.

8. References

G. Bouma, J. Mur, G. van Noord, in L. van der Plas. 2005a. Question answering for dutch using dependency relations. V: *Proceedings of the CLEF 2005 Workshop*.

G. Bouma, J. Mur, in G. van Noord. 2005b. Reasoning over dependency relations for qa. V: *Proceedings IJCAI Workshop on Knowledge and Reasoning for Answering Questions*.

M. Fleischman, E. Hovy, in A. Echihabi. 2003. Offline strategies for online question answering: Answering questions before they are asked. V: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

S. Hartrumpf. 2005. University of hagen at qa@clef 2005: Extending knowledge and deepening linguistic processing for question answering. V: *Proceedings of CLEF 2005 Workshop*.

V. Hoste in W. Daelemans. 2005. Learning dutch coreference resolution. V: T. van der Wouden, Michaela Poß, Hilke Reckman, in Crit Cremers, ur., *Computational Linguistics in the Netherlands 2004*, str. 133–148, Utrecht. LOT.

C. Jiangping, Y. Ping, in G. He. 2005. Unt 2005 trec qa participation: Using lemur as ir search engine. V: *Proceedings of TREC 2005*.

V. Jijkoun, J. Mur, in M. de Rijke. 2004. Information extraction for question answering: Improving recall through syntactic patterns. V: *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

D. Laurent, P. Séguéla, in S. Nègre. 2005. Cross lingual question answering using qristal for clef 2005. V: *Proceedings of the CLEF 2005 Workshop*.

B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Pe nas, V. Peinado, F. Verdejo, in M. de Rijke. 2004. The multiple language question answering track at clef. V: C. Peters, M. Braschler, J. Gonzalo, in M. Kluck, ur., *Results of the CLEF 2003 Evaluation Campaign. Lecture Notes in Computer Science*, Berlin Heidelberg New York. Springer-Verlag.

R. Malouf in G. van Noord. 2004. Wide coverage parsing with stochastic attribute value grammars. V: *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*.

D. Mollá, R. Schwitter, F. Rinaldi, J. Dowdall, in M. Hess. 2003. Anaphora resolution in extrans. V: *International Symposium on Reference Resolution and its Application to Question-Answering and Summarisation (ARQAS)*.

T.S. Morton. 2000. Coreference for nlp applications. V: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*.

J. Mur in M.L.E. van der Plas. to appear. Anaphora resolution for off-line answer extraction using instances. V: *Workshop on Anaphora Resolution*.

G. Neumann in B. Sacaleanu. 2005. Dfki's lt-lab at the clef 2005 multiple language question answering track. V: *Proceedings of the CLEF 2005 Workshop*.

P. Schone, G. Ciany, R. Cutts, P. McNamee, J. Mayfield, in Tom Smith. 2005. Qactis-based question answering at trec-2005. V: *Proceedings of TREC 2005*.

R. Stuckardt. 2003. Coreference-based summarization and question answering: a case for high precision anaphor resolution. V: *International Symposium on Reference Resolution and its Application to Question-Answering and Summarisation (ARQAS)*.

E. Tjong Kim Sang. 2005. Developing offline strategies for answering medical questions. V: *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*.

J.L. Vicedo in A Ferrández. 2000a. Importance of pronominal anaphora resolution in question answering systems. V: *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*.

J.L. Vicedo in A Ferrández. 2000b. A semantic approach to question answering systems. V: *Proceedings of TREC-9*.

R. Watson, J. Preiss, in E.J. Briscoe. 2003. The contribution of domain-independent robust pronominal anaphora resolution to open-domain question-answering. V: *International Symposium on Reference Resolution and its Application to Question-Answering and Summarisation (ARQAS)*.