Experiments on Cross–Linguality and Question–Type Driven Strategy Selection for Open–Domain QA*

Günter Neumann and Bogdan Sacaleanu

LT–Lab, DFKI, Saarbrücken, Germany
neumann@dfki.de, bogdan@dfki.de

Abstract. We describe the extensions made to our 2004 QA@CLEF German/English QA-system, toward a fully German-English/English-German cross-language system with answer validation through web usage. Details concerning the processing of factoid, definition and temporal questions are given and the results obtained in the monolingual German, bilingual English-German and German-English tasks are briefly presented and discussed.

1 Introduction

The basic functionality of a cross–lingual open–domain question answering (abbreviated as ODQA) system is simple: given a Natural Language query in one language (for example German) find answers for that query in textual documents written in another language (for example English). In contrast to a standard cross-language IR system, the natural language questions are usually well-formed NL–query clauses (instead of a set of keywords), and the identified answers should be *exact* answer strings (instead of complete documents containing the answers).

Since 2003, cross-lingual systems are evaluated as part of a special track at Clef. This year, the task was to process 200 questions of type `factoid`, `temporally restricted`, and `definition`, and to return for each question one exact answer (together with the identifier of the document source from which the answer was extracted) or NIL, if no answer could be found. Last year only factoid and definition questions were tackled.

Starting from our 2004–system (cf. [1]), the major efforts we spend for the QA track at Clef 2005 were focused on:

- improving cross–lingual methods
- development of a component–oriented ODQA-core architecture

* The work presented in this paper has been funded by the BMBF project Quetal, FKZ 01 IW C02. Many thanks to Rob Basten for his support in the development of the component for handling temporally restricted questions, Yuan Ye for his support in data collection and annotation for the definition handlers, and Aljeandro Figuero for his support in the implementation of the web validation strategy.

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 429–438, 2006.

- processing definition and temporally restricted questions
- exploration of web-based answer validation

Beside that we also decided to take part in three different tasks:

1. monolingual German ODQA: here we we improved our result from last year from 23.5% to 43.5% this year
2. German-English ODQA: here we achieved with 25.5% accuracy a minor improvement compared with our 2004–result (23.5%)
3. English-German ODQA: this was our first participation in this task and we achieved a result of 23% accuracy

In all three tasks, we obtained the best results. We will now describe some interesting technical aspects of our 2005–system – named QUANTICO – before presenting and discussing the results in more detail.

2 System Overview

Based on a number of experiments we made during the development of our ODQA–technology, we developed the hypothesis that a structural analysis of un-structured documents towards the information needs of questions, will support the retrieval of relevant small textual information units through *informative* IR-queries. However, since we cannot foresee all the different users interests or questions especially in the *open–domain context*, a challenging research question is: How detailed can the structural analysis be made without putting over a "straitjacket" of a particular interpretation on the un-structured source? Thus, there is a trade–off between off-line and on-line document annotation. Questions and answers are somewhat related in that questions influence the information geometry and hence, the information view and access, cf. [2].

Based on this insights, we developed the ODQA–architecture as depicted in figure 1. The idea behind the specific design is the assumption that an off-line annotation of the data collection supports an answer type oriented indexing and answer extraction process through the selection of query–type specific strategies (cf. sec. 3 for more details; a similar approach is also used by [3]). Furthermore, a sentence–oriented preprocessing determining only sentence boundary, named entities (NE) and their co-reference, as well as NE–anchored tuples (see sec. 6) turned out to be a useful level of off–line annotation, at least for the Clef-type of questions.

In order to achieve a high degree of flexibility of the ODQA–core components in future applications, an important design decision was to a use a central QA-Controller: based on the result of the NL—question analysis component, the QAController decides which of the following strategies will be followed:

- Definition Question
- Temporal Question
- Factoid Question

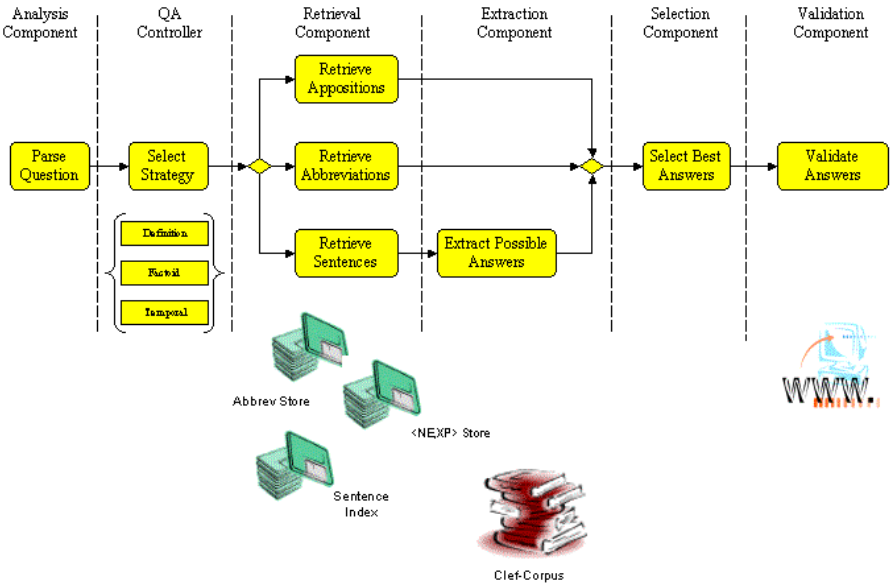

Fig. 1. The architecture of QUANTICO

For each of the above-mentioned tasks, a strategy corresponds to different settings of the components. For the Factoid Question strategy, for example, the Retrieval Component considers sentences as information units (see sec. 4 and 5 for more details); the Answer Extraction Component defines classes of instances for one of the entity types PERSON, ORGANIZATION, LOCATION, DATE and NUMBER; the Answer Selection Component considers relevant information as being the one more closed (distance metric) to the question keywords and with the most coherent context.

3 Question Analysis

The main purpose of the NL question analysis in the context of a open–domain QA-system is to determine the question–type, the expected answer type, the set of relevant keywords, and the set of recognized NE–instances in order to guide information search and answer processing. In our system, the question–type is used to select different answer strategies. For example, for a question of type *abbreviation*, possible answers are looked–up in special data bases (automatically filled with data from the Clef–corpus), where for questions of type *completion* the full–text search is activated. In a similar way, specific strategies for the treatment of definition and temporally restricted questions are handled (cf. 6). For more information on the syntactic and semantic aspects of our robust NL question analysis component, see [1].

4 Multi–layered Document Annotation

Beside word indexing and retrieval of raw text documents as information units relevant to a question, pre-emptive annotations have been done to the whole data collection. Driven by the controlled expected answer types of the potential questions, i.e. named entities types, a systematic annotation of named entities and co-reference resolution of both named entities and personal pronouns has been undertaken to the documents, in order to extend the IR-component with entity-based indices. Moreover, annotation of sentence boundaries, allowed us an accurate evaluation of IR-results along the information unit size. Based on

Table 1. Precision of retrieval for different unit types and top N units retrieved. We have alternatively considered the following retrieval units: documents, passages, sentences – and their NE-annotated correspondents (marked by *).

Unit–Type/#N	**1**	**5**	**10**	**20**	**30**	**40**	**50**	**100**
Sentences*	37.9	58.2	65.8	69.6	70	72.1	74	75.9
Sentence	28.4	53.1	60.1	67	70	72.7	72.7	74.6
Paragraph*	39.8	63.2	68.3	73.4	74	75.3	76.5	77.8
Paragraph	31.6	60.7	67.7	71.5	74	77.2	77.2	80.3
Document*	47.4	69.6	76.5	80.3	81	82.9	82.9	83.5
Document	46.2	68.3	77.8	82.2	82	83.5	84.1	85.4

experiments with the question set of previous CLEF competitions on the information retrieval unit and the indexation unit (see table 4), we have confined the first to the sentence level and added named entities and abbreviations, along words, as basic indexing units. By doing this, we could query the IR component not only by keywords extracted from the questions, but also by NE types corresponding to their expected answer types. This will not only narrow the amount of data being analyzed for answer extraction, but will also guarantee the existence of an answer candidate.

Even though we registered a decrease in precision of almost 10% with annotated sentences over raw documents as information units, we reduced the amount of "to be processed" data by a range of 30 and dispensed with the use of a passage retrieval component.

5 Treatment of Factoid Questions

Factoid questions require a single fact as answer, which has been restricted to a limited class of named entities (PERSON, ORGANIZATION, etc.) for the CLEF competition. Based on our named entities extended indices, a fixed number of sentences containing at least an instance of the expected answer type are being processed for answer extraction. Extracting the answers consists in gathering all those named entities corresponding to the expected answer type as possible answers to the question, whereby information from the retrieval component

(i.e., score, frequency of answer) is taken into account. Selection of best answers is based on a distance measure, which takes into consideration the number of overlapping words between the question words and the answers' context, the overlap cohesion (as distance between the question words) and the candidate cohesion (the distance between the answer and its most closed question words). The number of cross-document occurrences of the possible answers adds lastly to the weight to be computed for the best answer candidate.

6 Treatment of Definition and Temporally Restricted Questions

Definition Questions. Definition questions, asking about instances of PERSON and ORGANIZATION entity types, have been approached by making use of structural linguistic patterns known to be used with explanatory and descriptive goals. Both appositions:

> "Silvio Berlusconi, the Italian prime-minister, visited Germany."

and abbreviation-extension structural patterns:

> "In January 1994, Canada, the United States and Mexico launched the North American Free Trade Agreement (NAFTA) and formed the world's largest free trade area."

were used for this purpose.

Based on a corpus of almost 500 Mbytes textual data from the Clef corpus for every language taken into consideration (German and English), two indices were created corresponding to pairs of phrases of the form (see also fig. 1 where the (NE,XP) and abbreviation store memorize these indices).

> (Silvio Berlusconi, the Italian prime-minister)

and

> (NAFTA, North American Free Trade Agreement)

The Retrieval Component for the Definition Question strategy uses these indices and considers the phrases on the right hand as the information units containing the possible answer, if the corresponding matching left elements of such tuples have also been identified during the Query Analysis Component.

Temporally Restricted Questions. In order to fulfill the requirements of the 2005 qa@clef task description, we developed specific methods for the treatment of temporally restricted questions, e.g., questions like "Who was the German Chancellor in the year 1980?", "Who was the German Chancellor between 1970 and 1990?", or "Who was the German Chancellor when the Berlin Wall was opened?". It was our goal, to process questions of this kind on basis of our existing technology following a *divide-and-conquer* approach, i.e., by question decomposition and answer fusion. The highly flexible design of QUANTICO actually supported us in achieving this goal. Two methods were implemented:

1. The existing methods for handling factoid questions were used without change to get initial answer candidates. In a follow–up step, the temporal restriction from the question was used to check the answer's temporal consistency.
2. A temporally restricted question Q is decomposed into two sub–questions, one referring to the "timeless" proposition of Q, and the other to the temporally restricting part. For example, the question "Who was the German Chancellor when the Berlin Wall was opened?" is decomposed into the two sub–questions "Who was the German Chancellor‘" and "When was the Berlin Wall opened?". The answers for both are searched for independently, but checked for consistency in a follow–up answer fusion step. In this step, the identified explicit temporal restriction is used to instantiate the implicit time restriction.

The decomposition of such questions into sub–questions is helpful in cases, where the temporal restriction is only specified implicitly, and hence can only be deduced through application of specific inference rules. Note that the decomposition operation is mainly *syntax driven*, in that it takes into account the grammatical relationship of the sub– and main clauses identified and analysed by QUANTICO' parser SMES, cf. [4].

Through evaluation of a number of experiments, it turned out that processing of question with method 1.) leads to higher precision, and processing of questions using method 2.) leads to increased recall (see also [5]). An initial evaluation of our Clef–results also suggest, that the methods are critically dependant on the Named Entity recognizer's capability to properly recognize time and date expressions (see section 9).

7 Cross-Lingual Methods

Two strategies were used for answering questions asked in a language different from that used for documents containing the answer. Both strategies employ online translation services (Altavista, FreeTranslation, etc.) to solve the language barrier, but with different processing steps: before and after the Analysis Component (see also figure 2).

The **before–method** translated the question string in an earlier step, resulting in several automatic translated strings, of which the best one was then passed on to the Retrieval Component after having been analyzed by the Query Analysis Component. This was the strategy we used in the English–German task. To be more precise: the English source question was translated into several alternative German questions using online MT services. Each German question was then parsed with SMES, QUANTICO's German parser. The resulting query object was then weighted according to its linguistic well–formedness and its completeness wrt. query information (question type, question focus, answer–type). The assumption behind this weighting scheme is that "a translated string s_1 is of greater utility for subsequent processes than another translated string s_2, if the linguistic analysis of s_1 is more complete than the linguistic analysis of s_2."

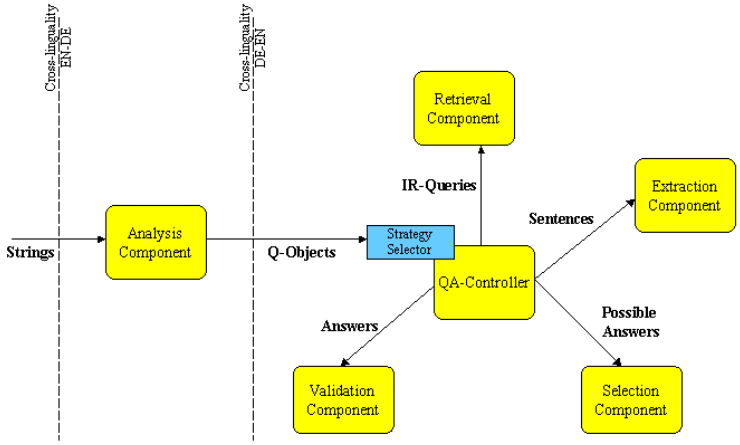

Fig. 2. The architecture of QUANTICO: cross–lingual perspective

The **after–method** translated the formalized result of the Query Analysis Component by using the question translations, a language modeling and a word alignment tool for creating a mapping of the formal information need from the source language into the target language. We used this strategy in the German–English task along two lines (using the following German query as example: *In welchem Jahrzehnt investierten japanische Autohersteller sehr stark?*):

1. translations as returned by the on-line MT systems are being ranked according to a language model

   In which decade did Japanese automakers invest very strongly? (0.7)
   In which decade did Japanese car manufacturers invest very strongly? (0.8)
2. translations with a satisfactory degree of resemblance to a natural language utterance (i.e. linguistically well-formedness), given by a threshold on the language model ranking, are aligned according to several filters: dictionary filter - based on MRD (machine readable dictionaries), PoS filter - based on statistical part-of-speech taggers, and cognates filter - based on string similarity measures (dice coefficient and LCSR (lowest common substring ratio)).

   In: [in:1] true 1.0
   welchem: [which:0.5] true 0.5
   Jahrzehnt: [decade:1] true 1.0
   investierten: [invest:1] true 1.0
   japanische: [japanese:0.5] true 0.5
   Autohersteller: [car manufacturers:0.8, automakers:0.1] true 0.8
   sehr: [very:1] true 1.0
   stark: [strongly:0.5] true 0.5

The CLEF evaluation gives evidence that both strategies are comparable in results, whereby the last one is slightly better, due to the fact of not being forced to choose a best translation, but working with and combining all the translations available. That is, considering and combining several, possible different, translations of the same question, the chance of catching a translation error in an earlier phase of the work–flow becomes higher and propagating errors through the whole system becomes less certain.

8 Web Validation

Our previous Clef–systems where "autistic" in the sense that we did not make use of the Web, neither for answer prediction nor for answer validation. Since we will fuse our current ODQA–technology with the Web in the near future, we started the development of web–based ODQA–strategies. Using the 2004 qa@clef as a testbed, we implemented an initial prototype of a web–validator realizing the following approach: Starting point are the M–best answer candidates found by QUANTICO using the Clef corpus only. Then, for each answer candidate a Google query is constructed from the answer and the the internal representation of the NL–query. The question–answer pair is sent to Google and the resulting total frequency count (TFC) is used to sort the set of answer candidates according to the individual values of TFC. The answer with the highest TFC is then selected as the best answer. The underlying assumption here is, that an IR–query consisting of the NL query terms and the correct answer term will have a higher redundancy on the Web, than one using a false answer candidate. Of course, applying such a method successfully presupposes a *semantic independency* between answer candidates. For this kind of answers, our method seemed to work quite well. However, for answer candidates, which stand in a certain "hidden" relationship (e.g., because a ISA–relation exists between the two candidates), the current method is not sufficient. This is also true for those answer candidates which refer to a different timeline or context than that, preferred by the Web search engine.

9 Results and Discussion

This year, we took part in three tasks: 1.) monolingual German (DE2DE), 2.) cross–lingual English/German (EN2DE), and 3.) cross–lingual German/English (DE2EN). at this point, we would like to stress, that in all different tasks, the *same* ODQA–core machinery was used, extended only for handling the cross–lingual aspects.

The results can be found in tables 2 (DE2DE), 3 (EN2DE), and 4 (DE2EN), respectively. For the tasks DE2DE and EN2DE we submitted two runs: one without web validation (the runs dfki051dede and dfki051ende) and one with web–validation (the runs dfki052dede and dfki052ende). For the task DE2EN, we only submitted one run without web validation. The system performance for the three tasks was as follows: for the task DE2DE, QUANTICO needs approx. 3

sec. for one question–answering cycle (about 10 minutes for all 200 questions); for the task EN2DE, QUANTICO needs approx. 5 sec. (about 17 minutes for all 200 questions), basically due to the extra time, the online machine translation needs. The task DE2EN needs the most computation resources due to online translation, alignment, language model use, etc. (actually approx. 50 minutes are used for all 200 questions).

Table 2. Results in the task German–German

	R		W	X	U	F	D	T
dfki051dede	87	43.50	100	13	-	35.83	66.00	36.67
dfki052dede	54	27.00	127	19	-	15.00	52.00	33.33

Table 3. Results in the task English–German

	R		W	X	U	F	D	T
dfki051ende	46	23.00	141	12	1	16.67	50.00	3.33
dfki052ende	31	15.50	159	8	2	8.33	42.00	0.00

Table 4. Results in the task German–English

	R		W	X	U	F	D	T
dfki051deen	51	25.50	141	8	-	18.18	50.00	13.79

As can be seen from the tables 2 and 3, applying the web validation component (for the best 3 answers determined by QUANTICO) does lead to a system performance loss. At the time of writing this report, we have not yet performed a detailed analysis, but it seems that the lack of contextual information causes the major problems, when computing the Google IR–query. Additional problems could be:

- the number of German web documents might be still too low, for taking into account redundancy effectively
- the correct answer extracted from the Clef–corpus does not exist on the web but a "alternative" answer candidate; in that case, the alternative answer candidate would get a higher rank
- the Clef corpus consists of newspaper articles from 1994 and 1995; thus, the Clef corpus might actually be too old for being validated by the Web, especially if questions referring not to historical events, but to daily news
- in case of EN2DE, web validation is performed with the German query terms, which resulted from automatic machine translation; errors through the translation of complex and long questions had a negative effect on the recall of the web search

However, a first comparison of the assessed results obtained for the task DE2DE, showed that the web validation is useful. Comparing the two runs

dfki051dede and dfki052dede (cf. table 2), a total of 51 different assignments were observed (e.g., an answer correct in run dfki051dede, was wrong in run dfki052dede). Actually, 13 questions (of which 8 are definition questions), which where answered incorrectly in dfki051dede, were now answered correctly in run dfki052dede. 28 questions, which were answered correctly in dfki051dede, were answered wrongly in dfki052dede. However, a closer look showed that about half of this errors, are due to the fact, that we actually performed web validation without taking into account the correct timeline. We assume that enhancing the Google IR–query with respect to Clef–corpus consistent timeline (1994/95) will improve the performance of our web validation strategy.

References

1. Neumann, G., Sacaleanu, S.: Experiments on robust nl question interpretation and multi-layered document annotation for a cross-language question/answering system. In: Clef 2004. Volume 3491., Springer-Verlag LNCS (2005) 411–422
2. Rijsbergen, C.V.: The Geometry of Information Retrieval. Cambridge University Pres (2004)
3. Moldovan, D., Harabagui, S., Clark, C., Bowden, M., Lehmann, J., Williams, J.: Experiments and analysis of lcc's two qa systems over trec 2004. In: Proceedings of The Thirteenth Text Retrieval Conference (TREC 2004), Gaithersburg, USA (2004)
4. Neumann, G., Piskorski, J.: A shallow text processing core engine. Computational Intelligence **18**(3) (2002) 451–476
5. Basten, R.: Answering open-domain temporally restricted questions in a multilingual context. Master's thesis, University of Twente and LT-lab DFKI (2005)