DFKI-LT at the CLEF 2006
Multiple Language Question Answering Track

Bogdan Sacaleanu and Günter Neumann
LT-Lab, DFKI, Saarbrücken, Germany
{bogdan,neumann}@dfki.de

Abstract

The paper describes QUANTICO, a cross-language open domain question answering system for German and English. The main features of the system are: use of preemptive off-line document annotation with syntactic information like chunk structures, apposition constructions and abbreviation-extension pairs for the passage retrieval; use of online translation services, language models and alignment methods for the cross-language scenarios; use of redundancy as an indicator of good answer candidates; selection of the best answers based on distance metrics defined over graph representations. Based on the question type two different strategies of answer extraction are triggered: for factoid questions answers are extracted from best IR-matched passages and selected by their redundancy and distance to the question keywords; for definition questions answers are considered to be the most redundant normalized linguistic structures with explanatory role (i.e., appositions, abbreviation's extensions). The results of evaluating the system's performance by CLEF were as follows: for the best German-German run we achieved an overall accuracy (ACC) of 42.33% and a mean reciprocal rank (MRR) of 0.45; for the best English-German run 32.98% (ACC) and 0.35 (MRR); for the German-English run 17.89% (ACC) and 0.17 (MRR).

Categories and Subject Headings

H.3 [**Information Storage and Retrieval**]: **H.3.1** Content Analysis and Indexing; **H.3.3** Information Search and Retrieval; **H.3.4** Systems and Software; **I.7** [**Document and Text Processing**]: **I.7.1** Document and Text Editing; **I.7.2** Document Preparation; I.2 [**Artificial Intelligence**]: **I.2.7** Natural Language Processing

General Terms

Algorithms, Design, Experimentation

Keywords

Open Domain Question Answering, Monolingual German, Cross-Language German/English

1. Introduction

QUANTICO is a cross-language open domain question answering system developed for both English and German open-domain question answering, cf. [2], [3]. It uses a common framework for both monolingual and cross-language scenarios, with different workflow settings for each task and different configurations for each type of question. For tasks with different languages on each end of the information flow (question and documents) we cross the language barrier rather on the question than on the document side by using free online translation services, linguistic knowledge and alignment methods. An important aspect of QUANTICO is the triggering of specific answering strategies by means of control information that has been determined by the question analysis tool, e.g., question type and expected answer type, see [3] for more details. Through the offline annotation of the document collection with several layers of linguistic information (chunks, appositions, named entities, sentence boundaries) and their use in the retrieval process, more accurate and reliable information units are being considered for answer extraction, which is based on the assumption that redundancy is a good indicator of information suitability. The answer selection component normalizes and represents the context of an answer candidate as a graph and computes its appropriateness in terms of the distance between the answer and question keywords.

We will begin giving a short overview of the system and presenting its working for both factoid and definition questions in monolingual and cross-language scenarios. We will then continue with a short description of each component and close the paper with the presentation of the CLEF evaluation results.

2. System Overview

QUANTICO uses a common framework for both monolingual and cross-language scenarios, but with different configurations for each type of question (definition or factoid) and different workflow settings for each task (DE2DE, DE2EN or EN2DE).

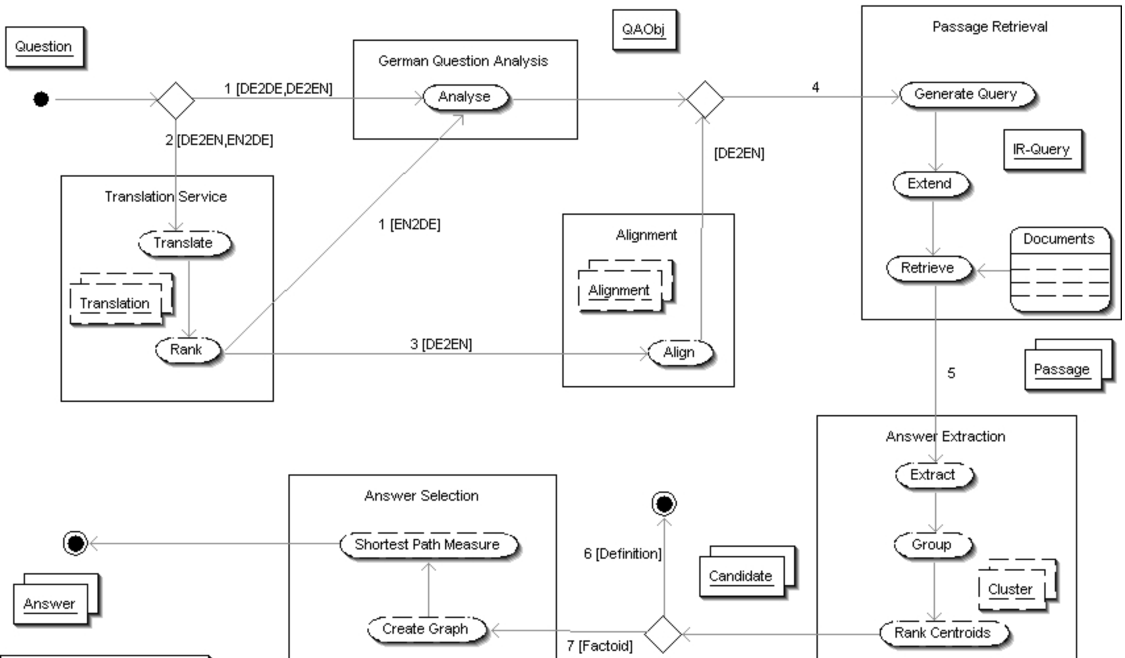

Figure 1. System Architecture

Concerning the workflow settings, the following things are to be mentioned. For the monolingual scenario (DE2DE) the workflow is as follows (according to the architecture in the Figure 1): 1-4-5-6/7 with the last selection depending on the question type. For a cross-language scenario, the workflow depends on the language of the question: for German questions and English documents (DE2EN) the workflow is 1-2-3-4-5-6/7, that is, the question is first analyzed, then translated and aligned to its translations, so that based on the generated *QAObj* and the alignments a new English *QAObj* is being computed; for English questions and German documents (EN2DE) the workflow is 2-1-4-5-6/7, that is, the question is first translated and then the best translation – determined according to linguistic completeness – is being analyzed resulting in a *QAObj*. The difference in the system's workflow for the cross-language scenario comes with our choice of analyzing only German questions, since our analysis component, based on the SMES parser [1], is very robust and accurate. In the presence of a *Question Analysis* component with similar properties for English questions, the workflow would be the same (1-2-3-4-5-6/7) independent of the question's language.

Regarding the component configurations for each type of question (temporal, definition or factoid) the difference is to be noted only in the *Passage Retrieval* and *Answer Extraction* components. While the *Retrieve* process for the factoid questions builds on classic Information Retrieval methods, for definition questions is merely a look-up procedure in a repository of offline extracted syntactic structures as appositions, chunks and abbreviation-extension pairs. For the *Answer Extraction* component the distinction consists in different methods of computing the clusters of candidate answers: for factoid question, where the candidates are usually named entities or chunks, is based on coreference (*John ~ John Doe*) and stop-word removal (*of death ~ death*), while for definition questions, where candidates can vary from chunks to whole sentences, is based on topic similarity (*Italian designer ~ the designer of a new clothes collection*).

3. Component Description

Following is a description of QUANTICO's individual components along with some examples.

3.1. Question Analysis

In context of a QA system or information search in general, we interpret the result of a NL question analysis as a *declarative description of search strategy and control information,* see [3]. Consider, for example, the NL question result in form of XML for the question "*In welcher Stadt fanden 2002 die olympischen Winterspile statt?" (The Olympic winter games took place 2002 in which town?)*, where the value of tag *a-type* represents the expected answer type, *q-type* the answer control strategy, and *q-focus* and *q-scope* additional constraints for the search space:


```
<QOBJ msg="quest" id="qId0" lang="de" score="1">
<NL-STRING id="qId0">
  <SOURCE id="qId0" lang="de">In welcher Stadt fanden 2002 die olympischen Winterspiele statt?</SOURCE>
  <TARGETS />
  </NL-STRING>
<QA-control>
  <Q-FOCUS>Stadt</Q-FOCUS>
  <Q-SCOPE>stattfind_winter#spiel</Q-SCOPE>
  <Q-TYPE restriction="TEMP">C-COMPLETION</Q-TYPE>
  <A-TYPE type="atomic">LOCATION</A-TYPE>
  </QA-control>
 <KEYWORDS>
 <KEYWORD id="kw0" type="UNIQUE">
  <TK pos="V" stem="statt#find">fanden</TK>
  </KEYWORD>
<KEYWORD id="kw1" type="UNIQUE">
  <TK pos="N" stem="stadt">Stadt</TK>
  </KEYWORD>
 <KEYWORD id="kw2" type="UNIQUE">
  <TK pos="NUMERAL" stem="2002">2002</TK>
  </KEYWORD>
 <KEYWORD id="kw3" type="UNIQUE">
  <TK pos="A" stem="olympisch">olympischen</TK>
  </KEYWORD>
 <KEYWORD id="kw4" type="UNIQUE">
  <TK pos="N" stem="winter#spiel">Winterspiele</TK>
  </KEYWORD>
  </KEYWORDS>
  <EXPANDED-KEYWORDS />
<NE-LIST>
  <NE id="ne0" type="DATE">2002</NE>
  </NE-LIST>
</QOBJ>
```


Parts of the information can already be determined on basis of local lexico-syntactic criteria (e.g., for the Wh-phrase *where* we can simply infer that the expected answer type is *location*). However, in most cases we have to consider larger syntactic units in combination with information extracted from external knowledge sources. For example for a definition question like "*What is a battery?"* we have to combine syntactic and type information from the verb and the relevant NP (e.g., combine definite/indefinite NPs together with certain auxiliary verb forms) in order to distinguish it from a description question like "*What is the name of the German Chancellor?"* In our QAS, we are doing this by following a two-step parsing schema:

- in a first step a full syntactic analysis is performed using the robust parser SMES (cf.[1]) and
- in a second step a question-specific semantic analysis.

During the second step, the values for the question tags *a-type*, *scope* and *s-ctr* are determined on basis of syntactic constraints applied on the dependency analysis of relevant NP and VP phrases (e.g., considering agreement and functional roles), and by taking into account information from two small knowledge bases. They basically perform a mapping from linguistic entities to values of the questions tags, e.g., trigger phrases like *name_of*, *type_of*, *abbreviation_of* or a mapping from lexical elements to expected answer types, like *town*, *person*, *and president*. For German, we additionally perform a *soft retrieval match* to the knowledge bases taking into account on-line compound analysis and string-similarity tests. For example, assuming the lexical mapping *Stadt* → *LOCATION* for the lexeme *town*, then automatically we will also map the nominal compounds *Hauptstadt* (capital) and *Großstadt* (large city) to *LOCATION*.

A main aspect in the adaptation and extension of the question analysis component for the Clef-2006 task concerned the recognition of the question type, i.e., simple factoid and list factoid questions, definition questions and the different types of the temporally restricted questions. Because of its high degree of modularity of the question analysis component, the extension only concerns the semantic analysis sub-component. Here, additional syntactic-semantic mapping constraints have been implemented that enriched the coverage of the question grammar, where we used the question set of the previous Clef campaigns as our development set.

3.2. Translation Services and Alignment

We are using two different methods for responding questions asked in a language different from the one of the answer-bearing documents. Both employ online translation services (Altavista, FreeTranslation, etc.) for crossing the language barrier, but at different processing steps: before and after formalizing the user information need into a *QAObj*.

The *a priori–method* translates the question string in an earlier step, resulting in several automatic translated strings, of which the best one is analyzed by the *Question Analysis* component and passed on to the *Passage Retrieval* component. This is the strategy we use in an English–German cross-lingual setting. To be more precise: the English source question is translated into several alternative German questions using online MT services. Each German question is then parsed with SMES, our German parser. The resulting query object is then weighted according to its linguistic well–formedness and its completeness with respect to the query information (question type, question focus, answer–type).
The assumption behind this weighting scheme is that "a translated string is of greater utility for subsequent processes than another one, if its linguistic analysis is more complete or appropriate."

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

- 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)*

- 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 based on 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:1.0] 1.0
welchem: [which:0.5] 0.5
Jahrzehnt: [decade:1.0] 1.0
investierten: [invest:1.0] 1.0
japanische: [Japanese:0.5] 0.5
Autohersteller: [car manufacturers:0.8, auto makers:0.1] 0.8
sehr: [very:1.0] 1.0

stark: [strongly:0.5] 0.5

3.3. Passage Retrieval

The preemptive offline document annotation refers to the process of annotating the document collections with information that might be valuable during the retrieval process by increasing the accuracy of the hit list. Since for factoid questions the expected answer type is usually a named entity type, annotating the documents with named entities provides for an additional indexation unit that might help to scale down the range of retrieved passages to those only containing the searched answer type. The same practice applies for definition questions given the known fact that some structural linguistic patterns (appositions, abbreviation-extension pairs) are used with explanatory and descriptive purpose. Extracting these kind of patterns in advance and looking up the definition term among them might return more accurate results than those of a search engine.

The *Generate Query* process mediates between the question analysis result *QAObj* (answer type, focus, keywords) and the search engine (factoid questions) or the repository of syntactic structures (definition questions) serving the retrieval component with information units (passages). The *Generate Query* process builds on an abstract description of the processing method for every type of question to accordingly generate the *IRQuery* to make use of the advanced indexation units. For example given the question "What is the capital of Germany?", since named entities were annotated during the offline annotation and used as indexing units, the *Query Generator* adapts the *IRQuery* so as to restrict the search only to those passages having at least two locations: one as the possible answer (*Berlin*) and the other as the question's keyword (*Germany*).

It is often the case that the question has a semantic similarity with the passages containing the answer, but no lexical overlap. For example, for a question like "*Who is the French prime-minister?*", passages containing "*prime-minister X of France*", "*prime-minister X … the Frenchman*" and "*the French leader of the government*" might be relevant for extracting the right answer. The *Extend* process accounts for bridging this gap at the lexical level, either through look-up of unambiguous resources or as a side-effect of the translation and alignment process (see [4]).

Whereas the *Retrieve* process for definition questions is straightforward for cases when the offline annotation repository lookup was successful, in other cases it implies an online search of the document collection and retrieval of only those passages that might bear a resemblance to a definition. The selection of these passages is attained by matching them against a lexico-syntactic pattern of the form:

<*Searched Concept*> <*definition verb*> .+

whereby <*definition verb*> is being defined as a closed list of verbs like "is", "means", "signify", "stand for" and so on.

3.4. Answer Extraction

The *Answer Extraction* component is based on the assumption that the redundancy of information is a good indicator for its suitability. The different configurations of this component for factoid and definition questions reflect the distinction of the answers being extracted for these two question types: simple chunks (i.e., named entities and basic noun phrases) and complex structures (from phrases through sentences) and their normalization. For factoid questions having named entities as expected answer type the *Group* (normalization) process consists in resolving cases of coreference, while for definition questions with phrases and sentences as possible answers more advanced methods are being involved. The current procedure for clustering definitions consists in finding out the focus of the explanatory sentence or the head of the considered phrase. Each cluster gets a weight assigned based solely on its size (definition questions) or using additional information like the average of the IR-scores and the document distribution for each of its members (factoid questions).

3.5. Answer Selection

Using the most representative sample (centroid) of the answer candidates' clusters, the *Answer Selection* component sorts out a list of top answers based on a distance metric defined over graph representations of the answer's context. The context is first normalized by removing all functional words and then represented as a graph structure. The score of an answer is defined in terms of its distance to the question concepts occurring in its context and the distance among these.

4. Evaluation Results

We participated in three tasks: DE2DE, EN2DE and DE2EN, where the summary of the results can be found in the table below.

Task	Overall accuracy	Factoid	Definition	Temporal	NIL (correct\returned)
dfki061dede	80/189=42.33%	59/156=37.82%	21/33=63.64%	0	9\32
dfki062dede	63/189=33.33%	47/156=30.13%	16/33=48.48%	0	8\29
dfki061ende	62/188=32.98%	44/156=28.21%	18/32=56.25%	0	12\57
dfki062ende	50/188=26.60%	34/156=21.79%	16/32=50.00%	0	13\58
dfki061deen	34/190=17.89%	26/150=17.33%	8/40=20.00%	0	8\65

For the tasks DE2DE and EN2DE we submitted two runs, for DE2EN only one. Compared to the results from last year, we were able to keep our performance for the monolingual German task (2005: 43.50%). For the task English to German we were able to improve our result (2005: 25.50%). But for the task German to English we observed a decrease (2005: 23.50%). Furthermore, although the question analysis component was able to identify the different types of temporal questions, in no cases we were able to correctly identify and extract answers for those questions. It is still unclear, why.

References

[1] G. Neumann and J. Piskorski. *A shallow text processing core engine.* Computational Intelligence, 18(3):451–476, 2002.

[2] G. Neumann and B. Sacaleanu. *Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering System.* In C. Peters et al. (Eds): Clef 2004, LNCS 3491, pp. 411-422, Springer Berlin Heidelberg, 2005.

[3] G. Neumann and B. Sacaleanu. *Experiments on Cross-Linguality and Question-Type Driven Strategy Selection for Open-Domain QA*. In C. Peters et al. (Eds): Clef 2005, LNCS 4022, pp. 429-438, Springer Berlin Heidelberg, 2006.

[4] B. Sacaleanu and G. Neumann: *Cross-Cutting Aspects of Cross-Language Question Answering Systems.* In Proceedings of the EACL workshop on Multilingual Question Answering - MLQA'06, Trento, Italy, 2006.