DFKI-LT at QA@CLEF 2007

Bogdan Sacaleanu, Günter Neumann and Christian Spurk

LT-Lab, DFKI, Saarbrücken, Germany
{bogdan, neumann, cspurk}@dfki.de

Abstract

This Working note shortly presents QUANTICO, a cross-language open domain question answering system for German and English document collections. The main features of the system are: use of preemptive off-line document annotation with information like Named Entities, sentence boundaries and pronominal anaphora resolution; online extraction of abbreviation-extension pairs and appositional constructions for the answer extraction; use of online translation services for the cross-language scenarios and of English as interlingua for language combinations not supported directly; 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 QA@CLEF 2007 were as follows: for the German-German run we achieved an overall accuracy (ACC) of 30%; for the English-German run 18.5% (ACC); for the German-English run 7% (ACC), for the Spanish-English run 10% (ACC) and for the Portuguese-German run 7% (ACC).

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

1 Introduction

QUANTICO is a cross-language open domain question answering system developed mainly for both English and German factoid and definition question. It uses a common framework for both monolingual and cross-language scenarios that crosses the language barrier rather on the question than on the document side by using free online translation services, linguistic knowledge and alignment methods. Through the offline annotation of the document collection with several layers of linguistic information (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. [1]

For this year's participation we have extended the system to deal with two further source languages (Spanish and Portuguese) and a new document collection (Wikipedia), and opted for first translating the questions and then interpreting them instead of analyzing the source question and using alignment methods to cross the language barrier as in our previous evaluations.

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 and the outcome of an incipient error analysis.

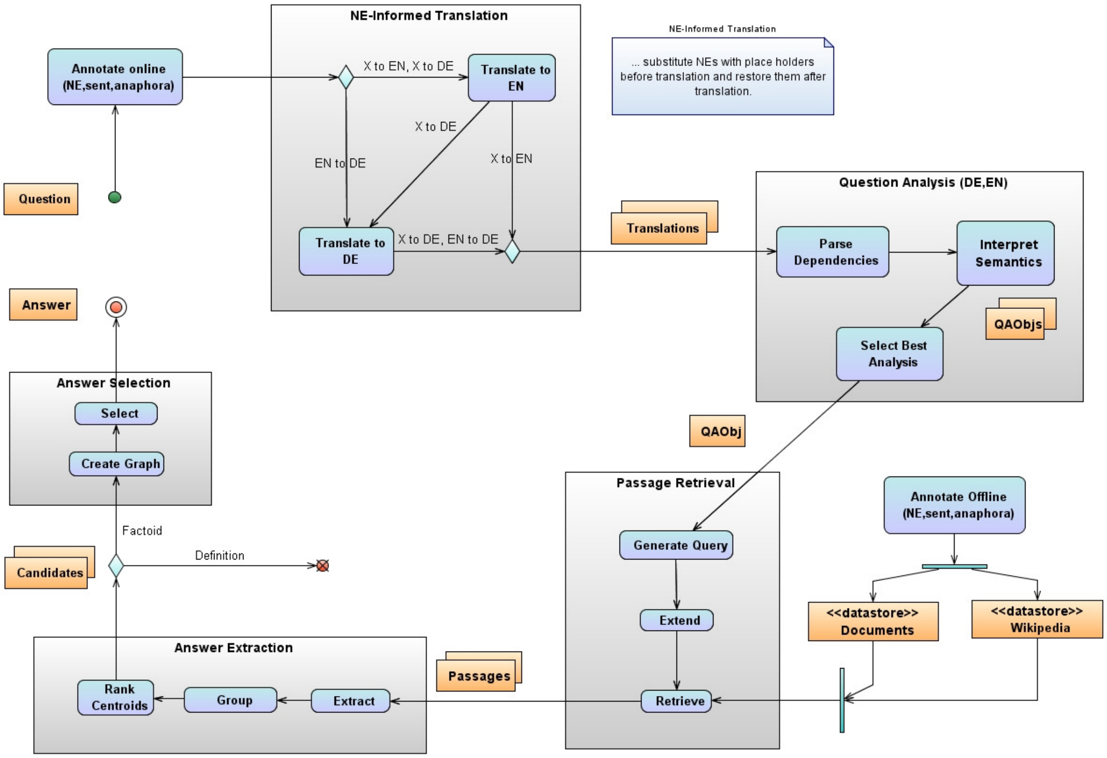

Figure 1. System Architecture

2 System Overview

QUANTICO uses a common framework for both monolingual and cross-language scenarios, with different workflow settings only for the translation component and different configurations of the extraction component for each type of question (definition or factoid).

The topics as provided by the evaluation exercise are first annotated with named entities (NE) and personal pronouns are linked to their NE-references. Every question is then translated into the target language resulting in a set of possible translations, which are individually interpreted. The outcome of the question analysis is ranked according to linguistic well-formedness and its completeness with respect to the query information (question type, question focus, answer–type) and the best alternative is considered for further processing. Passages relevant to this best formal representation of the question are retrieved and possible answer candidates are extracted and ranked based on their redundant occurrence. Finally, the best candidate is chosen as answer according to a distance metric based on features of the question's keywords and potential candidates.

The system is using online translation services (AltaVista[1], FreeTranslation[2] and VoilaTranslation[3]) for crossing the language barrier from the source language of the question to the target language of the document collection. The difference in workflow results from the lack of translation services for pairs of languages from any other language than English to German. For this cases (i.e. Portuguese to German) an additional step of translating to English and then to German has been considered, in which case English is used as an Interlingua.

[1] http://babelfish.altavista.com

[2] http:// ets.freetranslation.com

[3] http:// trans.voila.fr

For the *Answer Extraction* component the distinction consists in different methods of computing the clusters of candidate answers (*Group* process): for factoid question, where the candidates are usually named entities or chunks, is based on co-reference (*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 Descriptions

Following is a description of QUANTICO's individual components that have been used in this year's evaluation exercise along with some examples.

3.1 NE-Informed Translation

Since named entities can pose some problems in translation, especially proper names, by being translated when they should not be, the translation component has been developed with a substitution module that replaces some types of named entities with place holders before translating the question. The process is being reversed after translation, resulting in more accurate results. The outcome of this module is highly dependent on the accuracy of the named entity (NE) recognizer, since an inaccurate mark-up of the NE-terms might prevent from translating semantically relevant information.

For the case of inexistent or inaccurate online translation services for pairs of languages like Portuguese to German an Interlingua solution has been approached: that of using English as an intermediate translation and having the question first translated from Portuguese to English and then from English to German.

3.2 Question Analysis

The question parser computes for each question a syntactic dependency tree (which also contains recognized named entities) and semantic information like question type, the expected answer type, and the question focus, cf. [2] for details. The semantic information is determined on the basis of syntactic constraints applied on relevant NP and VP phrases of the dependency tree (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*.

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 the expected answer type for factoid questions 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 only to those containing the searched answer type.

The *Generate Query* process mediates between the question analysis result *QAObj* (answer type, focus, keywords) and the search engine 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*), as the following example shows:

+text:capital +text:Germany +neTypes:LOCATION +LOCATION:2.

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 through look-up of unambiguous resources.

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. Based on the control information supplied by the *Interpret Semantics* component (*q-type*), different extraction strategies are being triggered (noun phrases, named entities, definitions) and even refined according to the *a-type* (definition as sentence in case of an OBJECT, definition as complex noun phrase in case of a PERSON).

The *Extract* process for definition questions implies an online extraction of those passage-units only that might bear a resemblance to a definition. The extraction 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.

For factoid questions having named entities or simple noun phrases as expected answer type the *Group* (normalization) process consists in resolving cases of coreference, while for definition questions with complex 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' best-weighed 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.

In the context of the participation to CLEF a threshold of five best-weighed clusters has been chosen and all their instances, not only their centroids, have been considered for a thorough selection of the best candidate.

4 Evaluation Results

We participated in five tasks: DE2DE (German to German), EN2DE (English to German), DE2EN (German to English), ES2EN (Spanish to English) and PT2DE (Portuguese to German), with one run submitted for each of the tasks. A description of the achieved results can be seen in Table 1.

Table 1. System Performance

Run ID	Right		W	X	U
	#	%	#	#	#
dfki061dede$_M$	60	30	121	14	5
dfki061ende$_C$	37	18.5	144	18	1
dfki061deen$_C$	14	7	178	6	2
dfki062esen$_C$	10	5	180	10	0
dfki062ptde$_C$	5	2.5	189	4	2

A preliminary error analysis of the results has uncovered two weak places in our system:

- the English parser has a coverage too small as initially assumed to have and therefore created a bottleneck effect for those runs with English as target language,
- the *NE-Informed Translation* component, being highly dependent on the accuracy of the named entity (NE) recognizer, was disastrous for runs having Spanish and Portuguese as source language, for which the German NE-recognizer was used by default; in these cases a simple translation, without replacing any named entities, would have been more useful.

References

1. Sacaleanu, B., Neumann, G.: *DFKI-LT at the CLEF 2006 Multiple Language Question Answering Track*. In Working Notes for the CLEF 2006 Workshop, 20-22 September, Alicante, Spain.
2. Neumann, G., Sacaleanu, B.: *Experiments on Robust NL Question Interpretation and Multilayered 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)