LCC's PowerAnswer at QA@CLEF 2006

Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Jonathan Clark, Dan Moldovan
Language Computer Corporation
Richardson, Texas, 75080
United States of America
mitchell,marian,moldovan@languagecomputer.com

Abstract

This paper reports on Language Computer Corporation's first QA@CLEF participation. For this exercise, we integrated our open-domain PowerAnswer question answering system with our statistical machine translation engine. For 2006, we participated in the English-to-Spanish, French and Portuguese cross-language tasks. We took the approach of intermediate translation, only processing English within the QA system regardless of the input or source languages. The output snippets were then mapped back into the source language documents for the final output of the system and submission. What follows is a description of our system and methodology.

Categories and Subject Descriptors

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; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*; I.2 [**Artificial Intelligence**]: I.2.7 Natural Language Processing

General Terms

Measurement, Performance, Experimentation

Keywords

Open-domain Question Answering, Questions beyond factoids, Statistical machine translation

1 Introduction

For 2006, Language Computer's open-domain question answering system PowerAnswer [5] participated in QA@CLEF for the first time. PowerAnswer has previously participated in many other evaluations, notably TREC [1], however, this is the first Multilingual QA evaluation the system has entered. We have developed our own statistical machine translation system, which we integrated with PowerAnswer for this evaluation. Since PowerAnswer is a very modular and extensible system, we were able to make a minimum of modifications for this integration for our initial approach.

Our goals for this year's participation were (1) to examine how well the current QA system performs when given noisy data, such as that from automatic translation and (2) to examine the performance of the machine translation system in a question answering environment. To that end, we adopted an approach of intermediate translation instead of adapting the QA system to process target languages natively.

The paper presents a summary of the PowerAnswer system, our machine translation engine, the integration of the two for QA@CLEF 2006, and then follows with a discussion of our results and challenges in this year's CLEF question topics. Table 1 lists the cross-lingual tasks in which we participated.

Source	Target
English	French
English	Spanish
English	Portuguese

Table 1: LCC's QA@CLEF tasks

2 Overview of LCC's PowerAnswer

Automatic question answering requires a system that has a wide range of tools available. There is no one monolithic solution for all question types or even data sources. In realization of this, LCC developed PowerAnswer 2 as a fully-modular and distributed multi-strategy question answering system that integrates semantic relations, advanced inferencing abilities, syntactically constrained lexical chains, and temporal contexts. This section presents an outline of the system and how it was modified to meet the challenges of QA@CLEF 2006.

PowerAnswer comprises a set of strategies that are selected based on advanced question processing, and each strategy is developed to solve a specific class of questions either independently or together. A Strategy Selection module automatically analyzes the question and chooses a set of strategies with the algorithms and tools that are tailored to the class of the given question. PowerAnswer can distribute the strategies across workers in the case of multiple strategies being selected, alleviating the increase in the complexity of the question answering process by splitting the workload across machines and processors.

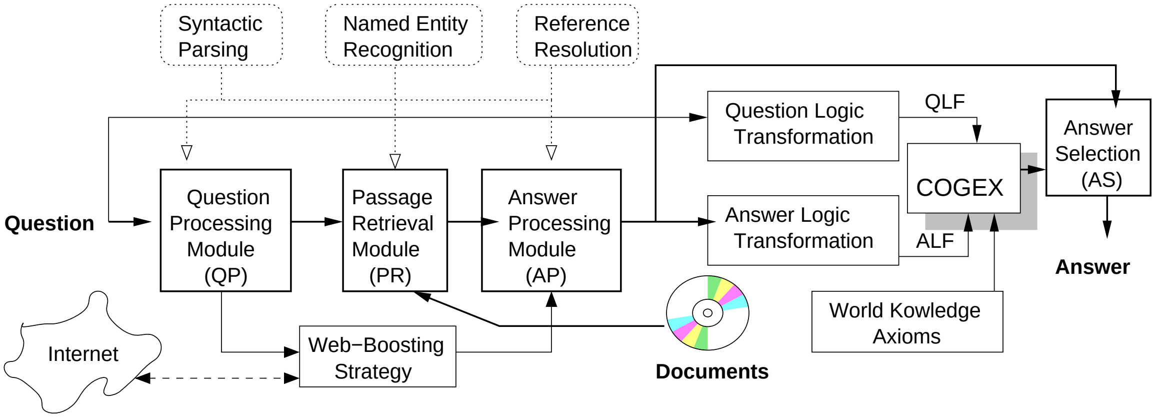

Figure 1: PowerAnswer 2 Architecture

Each strategy is a collection of components, (1) Question Processing, (2) Passage Retrieval, and (3) Answer Processing. Each of these components constitute one or more modules, which interface to a library of generic NLP tools. These NLP tools are the building blocks of the PowerAnswer 2 system that, through a well-defined set of interfaces, allow for rapid integration and testing of new tools and third-party software such as IR systems, syntactic parsers, named entity recognizers, logic provers, semantic parsers, ontologies, word sense disambiguation modules, and more. Furthermore, the components that make up each strategy can be interchanged to quickly create new strategies, if needed, they can also be distributed [10].

As illustrated in Figure 1, the role of the QP module is to determine (1) the expected answer type, (2) to select the keywords used in retrieving relevant passages, and (3) perform any preliminary questions as necessary for resolving question ambiguity. The PR module ranks passages that are retrieved by the IR system, while the AP module extracts and scores the candidate answers. All modules have access to a syntactic parser, a named entity recognizer and a reference resolution system through LCC's generic NLP tool libraries. To improve the answer selection, we take advantage of redundancy in large corpora, specifically in this case, the Internet. As the size of a document collection grows, a question answering system is more likely to pinpoint a candidate answer that closely resembles the surface structure of the question. These features have the role of correcting the errors in answer processing that are produced by the selection of keywords, by syntactic and semantic processing and by the absence of pragmatic information. Usually, the final decision for selecting answers is based on logical proofs from our inference engine COGEX. For this year's QA@CLEF, however, we disabled the logic prover in order to better evaluate the individual components of this QA architecture. COGEX's evaluation on multilingual data was performed in the CLEF Answer Validation Exercise [13].

3 Overview of Translation Engine

The translation system used at LCC implements phrase-based statistical machine translation [2], the core translation engine is the open-source Phramer [12] system, developed by one of LCC's engineers. Phramer in turn implements and extends the phrase-based machine translation algorithms implemented by Pharaoh [4]. A more detailed description of the MT solution that we adopted for Multilingual QA@CLEF can be found in [11]. We trained the translation system using the European Parliament Proceedings Parallel Corpus 1996-2003 (EUROPARL) [3], which provides between 600k and 800k pairs of sentences (sentences in English paired with the translation in another European language). We followed the training procedure described in the Pharaoh training manual[1] to generate the phrase table required for translation.

In order to translate entire documents, we augmented the core translation engine with (1) tokenization, (2) capitalization, and (3) de-tokenization.

The tokenization process was performed on the original documents (in French, Portuguese or Spanish), in order to convert the sentences to space-separated entities, in which the punctuation and the words are isolated. The step was required because the statistical machine translation core engine accepts only lowercased tokenized input.

The capitalization process follows the translation process and it restores the casing of the words. The capitalization tool uses three-gram statistics extracted from 150 million words from the English GigaWord Second Edition[2] corpus, augmented with two heuristics:

1. first word will always be uppercased
2. if the words appear also in the foreign documents, the casing is preserved (this rule is very effective for proper nouns and named entities)

4 PowerAnswer-Phramer Integration

Our cross-language solution for Question Answering was based on automatic translation of the documents in the source language (English). QA was performed on a collection consisting only of English documents. The answers were converted back into the target language (the original language of the documents) by aligning the translation with the original document (checking to see what was the original phrase in the original document that generated the answer in English); when this method failed, the system falls back to machine translation (source → target).

[1] http://www.iccs.inf.ed.ac.uk/ pkoehn/training.tgz

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2005T12

Passage Retrieval

Making use of PowerAnswer's modular design, we developed three different retrieval methods, settling on the first for our final experiment.

1. use an index of English words, created from the translated documents
2. use an index of foreign words (French, Spanish or Portuguese), created from the original documents
3. use an index of English words, created from the original documents in correlation with the translation table

The first solution is the default solution. The entire target language document collection is translated into English, processed through the set of NLP tools and indexed for querying. Its major disadvantage is the computational effort required to translate the entire collection. It also requires updating the English version of the collection when one improves the quality of the translation. Its major advantage is that there are no additional costs during question answering (the documents are already translated). This passage retrieval method is illustrated in Figure 2.

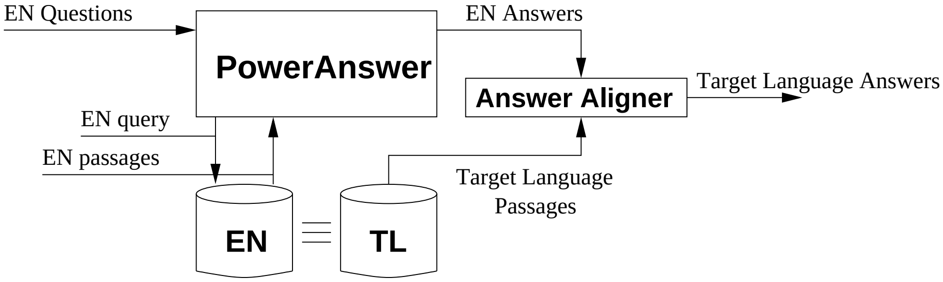

Figure 2: Passage Retrieval on English documents (default)

The second solution, as seen in Figure 3, requires minimum effort during indexing (the document collection is indexed in its native language). In order to retrieve the relevant documents, we translate the keywords of the IR query (the query submitted by PowerAnswer to the Lucene-based [3] IR system) with alternations as the new IR query (step 1). The translation of keywords is performed using Phramer, by generating n-best translations. This translated query is submitted to the target language index (step 2). The documents retrieved by this query are then dynamically translated into English using Phramer (step 3). We use a cache to store translated documents so that IR query reformulations and other questions that might retrieve the same documents will not need to be translated again. The set of translated documents is indexed into a mini-collection (step 4) and the mini-collection is re-queried using the original English-based IR query (step 5).

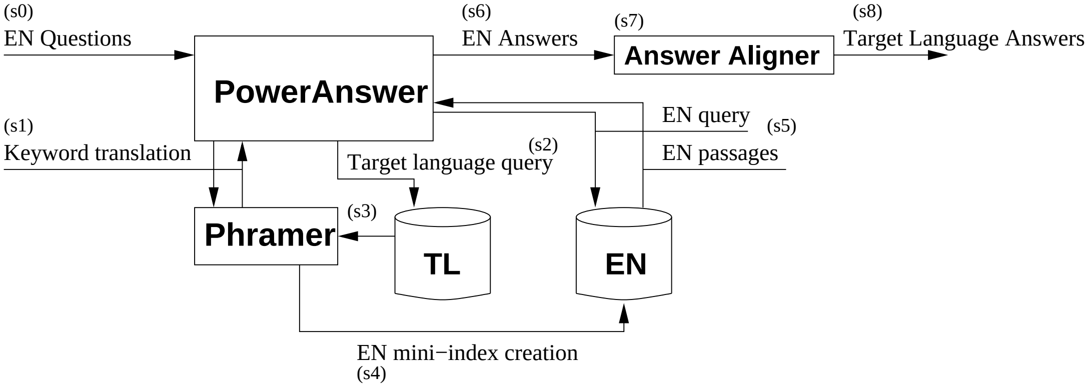

Figure 3: Passage Retrieval on Target Language documents

[3] http://lucene.apache.org/

For example, the boolean IR query in English *("poem" AND "love" AND "1922")* is translated into French as *("poeme" AND ("aiment" OR "aimer" OR "aimez" OR "amour") AND "1922")* with the alternations. This new query will return 85 French documents. Some of them do not contain "love" in their automatic translation (but the original document contains *"aiment"*, *"aimer"*, *"aimez"* or *"amour"*). Thus, by re-querying the translated sub-collection (that contains only the translation of those 85 documents) we retrieve only 72 English documents that will be passed to PowerAnswer.

The advantage of the second method is that minimum effort is required during collection preparation. Also, the collection preparation might not be under the control of the QA system (i.e. it can be web-based). Also, improvements in the MT engine can be reflected immediately in the output of the integrated system. The disadvantage is that more computation is required at run-time for translating the IR query and the documents dynamically.

The third alternative extracts during indexing the English words that might be part of the translation and indexes the collection accordingly. The process doesn't involve lexical choice - all choices are considered possible. The set of keywords is determined using the translation table, and collects all words that are part of the translation lattice ([4]). Determining only the words according to the translation table (semi-translation) is approximately 10 times faster than the full translation. The index is queried using the original IR query generated by PowerAnswer (with English keywords). After the initial retrieval, the algorithm is similar to the second method: translate the retrieved documents, re-query the mini-collection. The advantage is the much smaller indexing time when compared with the first method, besides all the advantages of the second method. Also, it has all the disadvantages of the second method, except that it doesn't require IR query translation.

Because preliminary testing proved that there aren't significant differences in recall between the three methods and because the first method is fastest after the document collection is prepared, we used only the first method for the final evaluation.

Answer Processing

For each of the above methods, PowerAnswer returns the exact answer and the supporting sentence answer the exact was extracted from (all in English). Then, these answers are then aligned to the corresponding text in the target language documents. The final output of the system is the converted responses in the target language with the appropriate supporting snippet. If the alignment method fails, the English answer is converted into the target language as the final response.

5 Results

Our integrated multilingual PowerAnswer system was tested on 190 English → Spanish, 190 English → French and 188 English → Portuguese factoid and definition questions and 10 English → Spanish, 10 English → French and 12 English → Portuguese list questions. For QA@CLEF, the main score is the overall accuracy, the average of SCORE(q), where SCORE(q) is defined for factoids and definition questions as 1 if the top answer for q is assessed as correct, 0 otherwise. Also included are the Mean Reciprocal Rank (MRR) and the Confidence Weighted Score (CWS) that judges how well a system returns correct answers higher in the ranked list of answers.

Table 2 illustrates the final results of Language Computer's efforts in our first participation at QA@CLEF for 2006.

Source	**Accuracy**	**CWS**	**MRR**
Spanish	20.00%	0.04916	0.2000
French	21.05%	0.04856	0.2623
Portuguese	8.51%	0.01494	0.1328

Table 2: LCC's QA@CLEF 2006 Factoid/Definition Results

6 Error Analysis and Challenges in 2006

There were several sources of errors in LCC's submission, including one major source that accounts for a 50% loss in accuracy. The sources of errors include: translation misalignments, tokenization errors, and data processing errors - questions and passages.

Translation misalignments
Because the version of PowerAnswer used this year is *mono*lingual, the design we used for *multi*lingual question answering involved translating documents dynamically for processing through the QA system and mapping the responses back into the source language documents. This resulted in many places where errors could occur. While the translation of the documents into English did introduce noise into the data such as mistranslations, words that were not translated and should have been or words that should not have been translated and were, it did not affect the QA system as much as we suspected. By far the greatest source of errors was the alignment between the English answers and the source documents which produced the final response list. This source accounts for roughly a 50% loss of accuracy for the tasks we participated in, as Table 3 shows. For Portuguese, there was also an error in the submission that accounts for the great difference in accuracy when compared to the Spanish and French results.

Source	**Position 1 Acc.**	**Top 5 Acc.**	Submission Acc.
Spanish	40.00%	63.16%	20.00%
French	40.53%	64.74%	21.05%
Portuguese	38.83%	62.23%	8.51%

Table 3: LCC's Factoid/Definition Results in English

Data errors
Questions that contained common nouns in the source language question were one example of data processing errors. Question 83 in the EN-PT task is *Who is the director of the film "Caro diario"?*. Here, the noun "Caro" was translated "expensive" when being indexed in English. Hence, the query keyword "Caro" as it is in the question could not be found in the translated collection by the IR system, causing the passage to receive a lowered score when retrieved on less keywords.

Sometimes the wording of the questions also affected the system in a negative way. There were some spelling changes that were unrecoverable by our system, such "Huan Karlos" for *Juan Carlos* (EN-PT #5). There were also a few instances of keywords for which PowerAnswer was unable to generate the correct alternation, for example "celebrated" in EN-ES #75 *In which year was the Football World Cup celebrated in the United States?*, where the correct alternation would have been a synonym for "to host".

The scoring of definition questions tends to be subjective, so there were cases where we believe an answer returned warranted at worst an inexact but was judged wrong, such as EN-PT #84 *What is the Unhcr?*. PowerAnswer responded with "O Acnur (Alto Comissariado das Nações Unidas para Refugiados) consultou o Brasil e mais 30 países sobre a possibilidade de acolher um grupo de 5.000 refugiados da ex-Iugoslávia", which was judged as wrong. Additionally, the definition question strategy often returned answer snippets that were a full sentence and were judged as inexact because of their length. One example is EN-PT question 34 *Who was Alexander Graham Bell?*. PowerAnswer returns the full sentence containing the important nugget *"A empresa foi fundada em 1885 e entre os sócios estava Alexander Graham Bell, o inventor do telefone"*, where the final part "inventor of the telephone" was all that was necessary.

7 Conclusions

While this year's performance was not what we hoped due to some major errors in the final stages of processing answers, we look forward to better results in the following years. One large step we will be taking is to make the core of PowerAnswer more language-independent. The English dependency comes from the NLP tools more than the modules of PowerAnswer, where the language dependence occurs primarily in Question Processing. By developing a set of multilingual NLP tools, including syntactic parsers and semantic relations extractors, and abstracting the English-dependent components of PowerAnswer out and building language-independent as well as some language-specific question processing, we hope to see great improvement in our multilingual QA capabilities. This work will be eased by the flexible design of PowerAnswer and our libraries of generic NLP tool interfaces. For next year's CLEF, we plan on submitting results from an improved and corrected system using the same method we have discussed in this paper, as well as from a more language-independent PowerAnswer that can process the source language text natively without intermediate translation.

References

[1] Sanda Harabagiu, Dan Moldovan, Christine Clark, Mitchell Bowden, Andy Hickl, Patrick Wang. Employing Two Question Answering Systems in TREC-2005. In *Text REtrieval Conference*, 2005.

[2] Philipp Koehn, Franz Josef Och and Daniel Marcu. Statistical phrase-based translation. *Proceedings of HLT/NAACL 2003 Edmonton, Canada*, 2003.

[3] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005*, 2005.

[4] Philipp Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004*, 2004.

[5] Dan Moldovan, Sanda Harabagiu, Christine Clark and Mitchell Bowden. PowerAnswer 2: Experiments and Analysis over TREC 2004. In *Text REtrieval Conference*, 2004.

[6] Dan Moldovan, Christine Clark, and Sanda Harabagiu. Temporal Context Representation and Reasoning. In *Proceedings of IJCAI*, Edinburgh, Scotland, 2005.

[7] Dan Moldovan, Christine Clark, Sanda Harabagiu, and Steve Maiorano. COGEX A Logic Prover for Question Answering. In *Proceedings of the HLT/NAACL*, 2003.

[8] Dan Moldovan and Adrian Novischi. Lexical chains for Question Answering. In *Proceedings of COLING*, Taipei, Taiwan, August 2002.

[9] Dan Moldovan and Vasile Rus. Logic Form Transformation of WordNet and its Applicability to Question Answering. In *Proceedings of ACL*, France, 2001.

[10] Dan Moldovan, Munirathnam Srikanth, Abraham Fowler, Altaf Mohammed, Eric Jean. Synergist: Tools for Intelligence Analysis. *NIMD Conference*, Arlington, VA, 2006.

[11] Marian Olteanu, Pasin Suriyentrakorn and Dan Moldovan. Language Models and Reranking for Machine Translation. In *NAACL 2006 Workshop On Statistical Machine Translation*, 2006.

[12] Marian Olteanu, Chris Davis, Ionut Volosen and Dan Moldovan. Phramer - An Open Source Statistical Phrase-Based Translator. In *NAACL 2006 Workshop On Statistical Machine Translation*, 2006.

[13] Marta Tatu, Brandon Iles, Dan Moldovan. Automatic Answer Validation using COGEX. *Cross-Language Evaluation Forum (CLEF)*, 2006.