Multilingual Question Answering through Intermediate Translation: LCC's PowerAnswer at QA@CLEF 2007

Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Thomas d'Silva, and Dan Moldovan

Language Computer Corporation
Richardson, Texas 75080, USA
mitchell,marian,moldovan@languagecomputer.com,
http://www.languagecomputer.com

Abstract. This paper reports on Language Computer Corporation's QA@CLEF 2007 preparation, participation and results. For this exercise, LCC integrated its open-domain PowerAnswer Question Answering system with its statistical Machine Translation engine. For 2007, LCC participated in the English-to-French and English-to-Portuguese cross-language tasks. The approach is that 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 the improved system and methodology and updates from QA@CLEF 2006.

1 Introduction

In 2006, Language Computer Corporation's open-domain question answering system PowerAnswer [6] participated in QA@CLEF for the first time [1], 2007 is a continuation of this exercise. PowerAnswer has previously participated in many other evaluations, notably NIST's TREC [7] workshop series, however, QA@CLEF is the first Multilingual QA evaluation the system has entered. Additionally, LCC has developed its own statistical machine translation system, which is integrated with PowerAnswer for this evaluation. Since PowerAnswer is a very modular and extensible system, the integration required only a minimum of modifications for the approach chosen.

The goals for participating in QA@CLEF are (1) to examine how well the QA system performs when given noisy data, such as that from automatic translation and (2) to examine and evaluate the performance and utility of the machine translation system in a question answering environment. To that end, LCC has 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, the machine translation engine, the integration of the two for QA@CLEF 2007, and then

follows with a discussion of results and challenges in the CLEF question topics. For 2007, LCC participated in the following bilingual tasks: English → French, and English → Portuguese.

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 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 2007.

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.

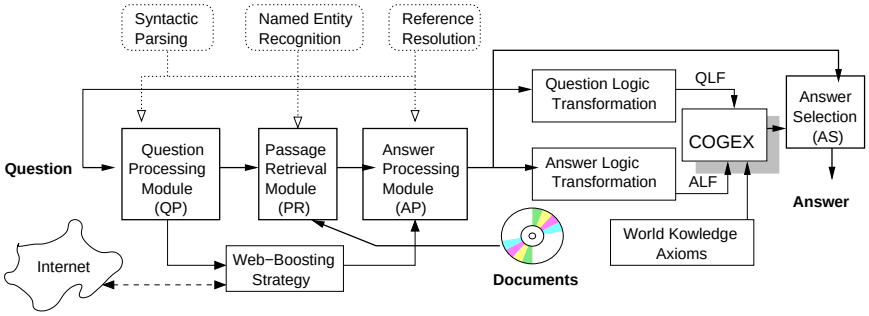

Fig. 1. PowerAnswer 2 Architecture

Each strategy is a collection of components, (1) Question Processing (QP), (2) Passage Retrieval (PR), and (3) Answer Processing (AP). 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 [12].

As illustrated in Figure 1, the role of the *QP* module is to determine (1) temporal constraints, (2) the expected answer type, (3) to process any question semantics necessary such as roles and relations, (4) to select the keywords used in retrieving relevant passages, and (5) 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 based on a number of syntactic and semantic features such as keyword density, count, proximity, semantic ordering, roles and entity type. All modules have access to a syntactic parser, semantic parser, a named entity recognizer and a reference resolution system through LCC's generic NLP tool libraries. To improve the answer selection, PowerAnswer takes 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 the inference engine COGEX [9]. For QA@CLEF, however, the logic prover is disabled in order to better evaluate the individual components of the QA architecture. COGEX's evaluation on multilingual data was performed in the 2006 CLEF Answer Validation Exercise [15], where the system was the top performer in both Spanish and English.

3 Overview of Translation Engine

The translation system used at LCC – MeTRe – implements phrase-based statistical machine translation [3]; the core translation engine is the open-source Phramer [14] system, developed by one of LCC's engineers. Phramer in turn implements and extends the phrase-based machine translation algorithms described by Koehn [3]. A more detailed description of the MT solution adopted for Multilingual QA@CLEF can be found in [13]. The translation system is trained using the European Parliament Proceedings Parallel Corpus 1996–2003 (EUROPARL) [4], which provides between 600,000 and 800,000 pairs of sentences (sentences in English paired with the translation in another European language). LCC followed the training procedure described in the Pharaoh [5] training manual[1] to generate the phrase table required for translation.

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

The tokenization process is performed on the original documents (in French or Portuguese), in order to convert the sentences to space-separated entities, in

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

which the punctuation and the words are isolated. The step is 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, due to using models trained on lowercase text. 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-MeTRe Integration

LCC's cross-language solution for Question Answering is based on automatic translation of the documents in the source language (English). QA is 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 (finding 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). While this fallback method provides excellent usability in a real-world situation, as discussed in the Errors discussion, the method produces answers judged *inexact* in an evaluation framework.

4.1 Passage Retrieval

Making use of PowerAnswer's modular design, for last year's QA@CLEF, LCC developed three different retrieval methods, settling on the first of these for the 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, and for 2007, the only method used. LCC selected this as the sole method this year because it gave the best performance in terms of quality versus runtime effort. Moreover, LCC has improved the speed of the automatic translator since the 2006 QA@CLEF. In addition to an algorithmic speed improvement of over 100% per execution core, and a decrease in the impact of network latency, the translator also now takes advantage of multiple processors, greatly increasing the time performance of the system. On dual-core machines, the translation speedup is more than 300%.

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

The entire target language document collection is translated into English, processed through the set of LCC's 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. For 2007, we created all new indexes of the collection. 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. As a main source of errors last year, for 2007 LCC made improvements to the Answer Aligner as described in Section 5. The second solution, as seen in Figure 3, requires minimum effort during

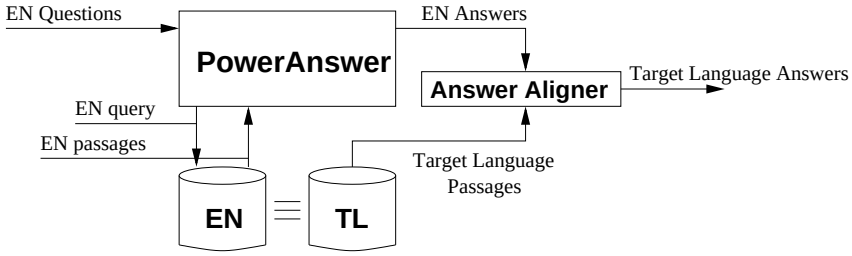

Fig. 2. Passage Retrieval on English documents (default)

indexing (the document collection is indexed in its native language). In order to retrieve the relevant documents, the system translates 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 MeTRe, 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 MeTRe (step 3). The system uses 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). 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) the system retrieves only 72 English documents that will be passed to PowerAnswer.

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

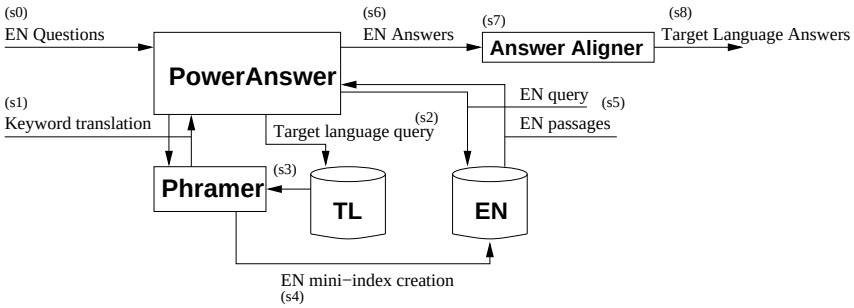

Fig. 3. Passage Retrieval on Target Language documents

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 ([5]). 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, only the first method was used for the final evaluation.

4.2 Answer Processing

For each of the above methods, PowerAnswer returns the exact answer and the supporting source sentence (all in English). These answers are aligned to the corresponding text in the target language documents. The final output of the system is the response list in the target language with the appropriate supporting snippets. If the alignment method fails, the English answers are converted directly into the target language as the final responses.

Source	Accuracy	CWS	Improv. from 2006
French	41.75%	0.22234	98.34%
Portuguese	29.32%	0.10484	244.54%

Table 1. LCC's QA@CLEF 2007 Results

5 Updates from QA@CLEF 2006

As 2006 was LCC's first year participating in CLEF, there were some substantial errors that were corrected for 2007 as well as some other improvements to various components of the system.

5.1 PowerAnswer improvements

Answer type detection
We extended PowerAnswer's answer type detection module by moving it to a hybrid system which takes advantage of precise heuristics as well as machine learning algorithms for ambiguous questions. A maximum entropy model was trained to detect both answer type terms and answer types. The learner's features for answer type terms include part-of-speech, lemma, head information, parse path to WH-word, and named entity information. Answer type detection uses a variety of attributes such as additional answer type term features and set-to-set lexical chains derived from eXtended WordNet[4] which links the set of question keywords to the set of potential answer type nodes.

Temporal processing
Dates for documents and the temporal context of the answer are maintained through question answering and after initial ranking, answers are given a boosting factor on top of their current relevance score that is intended to give greater priority to strong answers that are more recent than other strong answers. Answers that appear further down the response list and have lower relevance scores will not be affected by this boosting.

Because temporal answers can have a range of granularity, when pre-processing the data collection, the named entities stored in the IR index are extracted in a greedy fashion, so both "March 14, 1592" and "2000" will be tagged as *_date* to give PowerAnswer the best flexibility for entity selection. During answer processing, if the question is seeking just a month, or a year, then the excess information from the *_date* entity selected is removed after a more fine-grained NE recognition is performed on the answer nugget. EN → FR Q27 *In what year was Richard Nixon born?* demonstrates the utility of this method, where the answer is given in the text *... naît le 9 janvier 1913 ....* Otherwise, if a simple "When was ..." question is asked, the entity with the most detailed temporal information would be the final answer. This method operated on 4 EN → FR and 3 EN → PT questions seeking *year,* or *day.*

[4] `http://xwn.hlt.utdallas.edu`

5.2 Machine translation improvements

Since last year (QA@CLEF 2006 evaluation), we improved the Answer Aligner module: (1) we fixed bugs that altered the order of the answer in the output and (2) we improved the alignment heuristics.

In terms of Machine Translation quality, we added modules in MeTRe designed to better preserve the structure of the sentence. The add-ons were focused on rules that can be easily derived from punctuation: numeric values, currency amounts, insertions through quotation marks and through brackets, etc.

5.3 Wikipedia document conversion

PowerHarvest is a tool developed by Language Computer Corp. that is used for document harvesting and preprocessing for Question Answering. One of the features of PowerHarvest is to convert XML database dumps[5] into a format that is used by PowerAnswer's document collection indexing module.

Prior to QA@CLEF 07, PowerHarvest was limited to the English version of the Wikipedia collection – it only knew how to interpret English Wikipedia markup (e.g.: *Talk*, *User*, *User_talk*, *Template*, *Category*, ...). We extended PowerHarvest to work also on the targeted languages – French and Portuguese – by introducing support for French markup (e.g.: *Discuter*, *Utilisateur*, *Discussion_Utilisateur*, *Modèle*, *Catégorie*, ...) and Portuguese markup (e.g.: *Discussão*, *Usuário*, *Usuário_Discussão*, *Predefinição*, *Categoria*, ...).

The documents resulting from PowerHarvest (in French and in Portuguese) were translated using MeTRe and indexed, using the same procedures that were used for the Newswire parts of the collection (*Le Monde* and *French SDA* for French; *Público* and *Folha de São Paulo* for Portugese – according to the *Guidelines for Participants in QA@CLEF 2007*).

6 Results

The integrated multilingual PowerAnswer system was tested on 200 English → French and 200 English → Portuguese factoid, list and definition 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 is the Confidence Weighted Score (CWS) that judges how well a system confidently returns correct answers.

Table 1 illustrates the final results of Language Computer's efforts in its participation at QA@CLEF for 2007.

[5] http://download.wikimedia.org

7 Error Analysis and Challenges in 2007

While LCC saw a substantial improvement in errors over last year's results, there remain challenges that offer interesting research and engineering opportunities. The major sources of errors include: translation misalignments, tokenization errors, and data processing errors – questions and passages.

7.1 Translation misalignments

Because the version of PowerAnswer used is *mono*lingual, the system design for *multi*lingual question answering involves translating documents dynamically for processing through the QA system and later mapping the responses back into the source language documents. This results in several opportunities for error. 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, aggressive keyword expansion techniques diminish the impact of these mistranslations. Errors from misalignments still occured due to

For the French source results, PowerAnswer returned 14 inexact answers, and for Portuguese source 7 inexact, 7% and 3.5% of the total response. Many of these inexact responses are definition-style questions that either
(1) did not have enough information, such as EN → FR Q158: *Who is Amira Casar?, actrice née le 1erjuillet 1971 à Londres, d'une mère russe chanteuse d'opéra et d'un père d'origine kurde.* or (2) the alignment module was unable to correctly align the English answer within the given source language document, and so fell back to translating the English answer. While this particular default behavior is positive for the user since the answer is readable and still correct in nature, the language is not exact from the document and so warrants an inexact judgment in the evaluation. This failure is caused by translation errors when trying to map back from noisy text to the original source.

An example of this is EN → FR Q154: *Who is Allan Frederick Jacobsen?*. The source document is the Wikipedia "Allan Jacobsen" entry. The source language answer is *Allan Frederick Jacobsen, né le 22 septembre 1978 à Edimbourg (Écosse) est un joueur de rugby à XV qui joue avec l'équipe d'Écosse depuis 2002, évoluant au poste de pilier (1,78m et 109kg).*
The answer returned by PowerAnswer over the English translated Wikipedia article is *born on 22 September 1978 to Edinburgh (Scotland - is a player rugby to XV is playing with the team of Scotland since 2002 swimming as pillar (1.78 me and 109 kg).*
The final submitted result, which was translated as the default was *22 nés sur édimbourg à 1978 septembre un joueur - est (scotland est rugby xv à jouez avec écosse l ' équipe depuis 2002 de baigner (1.78 comme pilier 109 kg) moi et.* While the final answer is readable and comprehensible, it is not the answer as it appears in the source document.

7.2 Returning NIL as answer

The version of PowerAnswer used for QA@CLEF uses parameters that relax some of the semantic and syntactic restrictions on answers that PowerAnswer uses when running on more stable and less noisy data. A result of this is that zero NIL answers were returned because the system always attempts to return an answer. An example of this is EN → PT Q13: *When did the blue whale become extinct?*, the answer to which is NIL because the blue whale has never become extinct. PowerAnswer selected the translated answer *When the hunting of whale blue has finally been banned in the 1960s, 350000 whales Blue had been killed.* with the exact answer *the 1960s*, but with a low relative confidence score.

7.3 Other error sources

Other error sources are less specific to the methodology of intermediate translation and more general question answering errors such as answer type detection, keyword selection and expansion, passage retrieval and answer selection/ranking. An example of an answer selection error is EN → PT Q24 *What department is Caen the capital of?*. The correct answer string is *Caen é uma comuna francesa na região administrativa da Baixa-Normandia, no departamento* **Calvados** but PowerAnswer selected " Baixa-Normandia" as the correct answer instead of Calvados due to proximity.

7.4 English accuracy

As we also included for last year's results [1], Table 2 compares the PowerAnswer English accuracy versus the mapped submission accuracy. This table also demonstrates that the system did obtain the expected improvements after the correction of misalignment errors present in the submission for QA@CLEF 2006.

Source	**Submission Acc.**	**Eng. Position 1 Acc.**
French	41.75%	52.06%
Portuguese	29.32%	39.23%

Table 2. LCC's Factoid/Definition Results in English

8 Conclusions

QA@CLEF 2007 proved to be a valuable learning exercise. We have been able to correct some of the errors that were present in last year's results and achieve the kind of performance we expected from PowerAnswer. Intermediate translation for question answering provides the opportunity for additional errors in processing, but we believe that our results in this evaluation show that such a methodology can be practical and accurate.

References

1. Mitchell Bowden, Marian Olteanu, Pasin Suriyentrakorn, Jonathan Clark, Dan Moldovan. LCC's PowerAnswer at QA@CLEF 2006. In *CLEF 2006 Proceedings.* Springer (LNCS in press), 2007.
2. 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.
3. Philipp Koehn, Franz Josef Och and Daniel Marcu. Statistical Phrase-Based Translation. *Proceedings of HLT/NAACL 2003 Edmonton, Canada*, 2003.
4. Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. *MT Summit 2005*, 2005.
5. Philipp Koehn. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004*, 2004.
6. Dan Moldovan, Sanda Harabagiu, Christine Clark and Mitchell Bowden. PowerAnswer 2: Experiments and Analysis over TREC 2004. In *Text REtrieval Conference*, 2004.
7. Dan Moldovan, Mitchell Bowden, Marta Tatu. A Temporally-Enhanced PowerAnswer in TREC 2006. In *Text REtrieval Conference*, 2006.
8. Dan Moldovan, Christine Clark, and Sanda Harabagiu. Temporal Context Representation and Reasoning. In *Proceedings of IJCAI*, Edinburgh, Scotland, 2005.
9. Dan Moldovan, Christine Clark, Sanda Harabagiu, and Steve Maiorano. COGEX A Logic Prover for Question Answering. In *Proceedings of the HLT/NAACL*, 2003.
10. Dan Moldovan and Adrian Novischi. Lexical chains for Question Answering. In *Proceedings of COLING*, Taipei, Taiwan, August 2002.
11. Dan Moldovan and Vasile Rus. Logic Form Transformation of WordNet and its Applicability to Question Answering. In *Proceedings of ACL*, France, 2001.
12. Dan Moldovan, Munirathnam Srikanth, Abraham Fowler, Altaf Mohammed, Eric Jean. Synergist: Tools for Intelligence Analysis. *NIMD Conference*, Arlington, VA, 2006.
13. Marian Olteanu, Pasin Suriyentrakorn and Dan Moldovan. Language Models and Reranking for Machine Translation. In *NAACL 2006 Workshop On Statistical Machine Translation*, 2006.
14. 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.
15. Marta Tatu, Brandon Iles, Dan Moldovan. Automatic Answer Validation using COGEX. *Cross-Language Evaluation Forum (CLEF)*, 2006.