Experiments with LSA for Passage Re-Ranking in Question Answering

David Tomás[1] José L. Vicedo[1] Empar Bisbal[2] Lidia Moreno[2]
[1] Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, Spain
{dtomas,vicedo}@dlsi.ua.es
[2] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Spain
{ebisbal,lmoreno}@dsic.upv.es

Abstract

As in the previous QA@CLEF track, two separate groups at the University of Alicante participated this year using different approaches. This paper describes the work of *Alicante 1* group. We have continued with the research line established in the past competition, where the main goal was to obtain a fully data-driven system based on machine learning techniques. Last year an XML framework was established in order to obtain a modular system where each component could be easily replaced or upgraded. In this framework, a question classification system based on Support Vector Machines (SVM) and surface text features was included, achieving remarkable performance in this stage. The main novelties introduced this year are focused on the information retrieval stage. First, we employed Indri as our search engine for passage retrieval. Secondly, we developed a module for passage re-ranking based on Latent Semantic Analysis (LSA). This technique provides a method for determining the similarity of meaning between words by analysis of large text corpora. In our experiments, every question was compared with every passage returned by the search engine by means of LSA in order to re-rank them. Looking at the results, this technique increased the retrieval accuracy for definition questions but it decreased accuracy on factoid ones. To take advantage of the flexibility and adaptability of our machine learning based proposal, this year we extended our participation to monolingual Spanish task and bilingual Spanish-English task. We reach a best overall accuracy of 29.47% in the first task and 20.00% in the second one.

Categories and Subject Descriptors

H.3.2 [**Information Storage and Retrieval**]: Information Storage—*File organization*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Natural Language*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language parsing and understanding*

General Terms

Experimentation, Performance, Design

Keywords

Question answering, Information Retrieval, Multilingual, Machine Learning, LSA

1 Introduction

This paper is focused on the work of *Alicante 1* group at the University of Alicante. We have continued with the research line established in the past competition [1], where an XML framework was defined with the aim of easily integrate machine learning based modules on every stage of the Question Answering (QA) process: question analysis, information retrieval and answer extraction. In this framework and inside the question analysis stage, a question classification system based on Support Vector Machines (SVM) and surface text features [2] was included the last year, achieving remarkable performance at this point.

The main novelties introduced this year are focused on the information retrieval stage. First, we moved from Xapian[1] to Indri[2] as passage retrieval engine. Secondly, we developed a procedure for passage re-ranking based on Latent Semantic Analysis (LSA) [3]. This technique provides a method for determining the similarity of meaning between words by analysis of large text corpora. We employed LSA to compare every question formulated to the system with every passage returned by the Indri search engine. With this similarity measure our system can automatically re-rank the passages retrieved.

One of the main advantages that systems based on machine learning techniques offer is their flexibility and adaptability to new languages and domains. Thus, this year we extended our participation from the original monolingual Spanish task to bilingual Spanish-English task, as we wanted to test the system in a different language that the one it was originally intended for. For every task we submitted two runs, one with the baseline system and another with LSA passage re-ranking.

This paper is organized as follows: in section 2 we describe the system architecture detailing the novelties included this year; section 3 outlines the runs submitted; section 4 presents and analyzes the results obtained at QA@CLEF 2006 monolingual Spanish and bilingual Spanish-English task, paying special attention at the passage re-ranking module; finally, in section 5 we discuss the conclusions and main challenges for future work.

2 System Description

The main architecture of the system remains unchanged since the past competition [1]. We have a modularized structure where each stage can be isolated in an independent component so as to easily integrate and substitute new ones in an XML framework. This way, we want to evolve the system originally described in [4] and [5] to a fully data-driven approach, adapting every stage of the system with machine learning techniques.

The system follows the classical three-stage pipeline architecture mentioned above. Next paragraphs describe each module in detail.

2.1 Question Analysis

This first stage carries out two processes: question classification and query construction. The first one detects what type of information is claimed in the question, mapping it into a previously defined taxonomy. Otherwise, query construction detects meaningful terms from the question that help the search engine to locate the documents or paragraphs that are likely to contain the answer.

The question classification system is based on Support Vector Machines (SVM). The system was trained with a parallel annotated corpus in Spanish and English made up of question from

[1] http://www.xapian.org
[2] http://www.lemurproject.org/indri

Question Answering Track in TREC[3] 1999 to 2003 and CLEF 2003 to 2004, to sum up 2739 training questions. Thus, there is no need to manually tune the question classification system from one language to another since all the knowledge necessary to classify the questions is automatically acquired changing the training corpus. Anyway, as Spanish was the only source language in the tasks we took part, only classification in this language was carried out. Despite the corpus initially presented fifteen different categories, we reduced them to five so as to fit the needs of the answer extraction process. There is a more detailed description of this module in [2].

On the other hand, the query construction module remains the same for this year competition [5]. It employs hand made lexical patters in order to obtain the information needed to build the query for the Indri search engine. We have already developed a fully data-driven keyword detection module [6] to help on query construction although it has not being included in the global QA system yet. In the bilingual Spanish-English task, we just translated the question into English through SysTran[4] online translation service, and applied adapted patterns to obtain the query.

2.2 Information Retrieval

The main novelties introduced this year in our system are focused on this stage. First, we moved the search engine from Xapian to Indri due to the computational cost of Xapian indexing module. Indri offers a powerful and flexible query language, but as our keyword extraction and query construction approach are quite simple, we limited the queries to *ad hoc* retrieval with *query likelihood* ranking approach. The search engine performs passage retrieval over the entire Spanish or English corpora depending on the task. It returns the 50 topmost relevant passages with a length of 200 words which are provided to the answer extraction module.

On the other hand, we have developed a re-ranking component for the passages retrieved, trying to weigh them beyond query likelihood criterion. This component is based on LSA. This is a corpus-based statistical method for inducing and representing aspects of the meaning of words and passages reflected in their usage. In LSA a representative sample of text is converted to a term-by-passage matrix in which each cell indicates the frequency with which each term occurs in each passage. After a preliminary information-theoretic weighting of cell entries, the matrix is submitted to Singular Value Decomposition (SVD) and a 100 to 1500 dimensional abstract semantic space is constructed in which each original word and each original passage are represented as vectors. Thus, the similarities derived by LSA are not simple co-occurrence statistics, as the dimension-reduction step constitutes a form of induction that can extract a great deal of added information from mutual constraints among a large number of words occurring in a large number of contexts.

In order to test this module, Indri first retrieves the best 1000 passages extracted from the corpora. These passages are employed to construct the term-by-passage matrix with Infomap NLP Software[5] reducing the original matrix to 100 dimensions (the default value) by means of SVD. Then, the similarity between the question and every passage is calculated. All the passages are sorted in decreasing order of this value, keeping the 50 best valued passages. This process tries to better rank the passages semantically related to the question, overriding the problem of exact match of keywords from the query.

Unlike the previous competition, we removed the additional search performed in Google[6] as statistical indicator of answer correctness. We wanted to avoid introducing noise in the retrieval process in order to better evaluate the new search engine and the re-ranking module.

As Indri performs multilingual search without additional tuning and so does the re-ranking module, no special modification was necessary when working on Spanish or English corpora.

[3] Text REtrieval Conference, http://trec.nist.gov
[4] http://www.systransoft.com
[5] http://infomap-nlp.sourceforge.net
[6] http://www.google.com

2.3 Answer Extraction

This module remains almost unchanged since the past competition. The only remarkable change is due to Google search removal, as we described in the previous section. We had to modify the frequency value from the original formula [1] as we do not include the frequency of the web summaries retrieved by Google anymore.

At this point of the process the system keeps the following information: keywords and definitions terms from the query, and the set of relevant passages retrieved from de corpora. The system employs this information to extract a single answer from the list of relevant passages retrieved by the search engine. The set of possible answers is formed extracting all the n-grams (unigrams, bigrams and trigrams in our experiments) from these relevant passages. First, some heuristics are applied in order to filter non probable answers, like those that contains query terms or do not fit the question type. Next, remaining candidate answers are scored taking into account the sentence where they appear, the frequency of appearance, the distance to the keywords and the size of the answer.

As all these heuristics and features are language independent, the answer extraction module can be either applied to Spanish or English passages without additional changes.

3 Runs Submitted

As in the past three years, we have participated in the monolingual Spanish task. As a novelty, we also took part in the bilingual Spanish-English task. We were interested in testing the skills of our system retrieving and extracting answers in another language. We applied no special cross-lingual techniques as we just automatically translated the questions from Spanish to English to set a monolingual English environment to test the system.

We performed two different runs for every task. The difference between each run is established in the information retrieval stage. The first run represents the baseline experiment, where Indri is employed to retrieve the 50 topmost relevant passages related to the query. The second run introduces the re-ranking module described in section 2. Indri retrieves the best 1000 passages, which are then re-ranked via LSA to finally select the 50 best ones. The rest of the system remains the same for both runs.

4 Results

This year we submitted four different runs, two for the monolingual Spanish task and two more for the bilingual Spanish-English task. As a novelty, the QA@CLEF track presents this year the inclusion of list questions, where the systems are required to return not just one answer but a list of possible answers.

There were a total of 200 questions per run. Table 1 shows the overall and detailed results obtained for factoid, definition and temporally restricted factoid questions. This last type is not really representative as there were only two questions of this kind in the monolingual Spanish track, and none in the bilingual Spanish-English track. Experiment *aliv061eses* refers to the first run (baseline experiment) in the monolingual Spanish task. Experiment *aliv062eses* refers to the second run (passage re-ranking) in the monolingual Spanish task. Otherwise, experiment *aliv061esen* refers to the first run in the bilingual Spanish-English task, while *aliv062esen* refers to the second run.

Both runs in the monolingual Spanish task present the same overall accuracy (29.47%). Nevertheless, there are differences in the performance depending on the question type. For factoid answers, the first run offers better accuracy, while the second run results more precise for definition questions. This tendency is repeated in the bilingual Spanish-English task. So, the re-ranking module seems to perform better for definition questions while decreases the accuracy of the baseline system over factoid ones.

Table 1: Detailed results for monolingual Spanish and bilingual Spanish-English tasks

Accuracy (%)				
Experiment	**Factoid**	**Definition**	**Temporally restricted**	**Overall**
aliv061eses	28.08	35.71	0.00	29.47
aliv062eses	26.71	40.48	0.00	29.47
aliv061esen	19.33	22.50	-	20.00
aliv062esen	12.00	27.50	-	15.26

The experiments in bilingual Spanish-English task reflect a significant loss of performance with respect to the monolingual Spanish runs (20.00% of accuracy in the best run). As for this task we just employed automatic translations of the Spanish questions into English, the noise introduced in this early stage of the question answer process heavily damages the global performance of the system. Anyway, the query construction module relies on lexical extraction patterns so that grammatical mistakes in translations do not significantly affect the performance of the system. These problems provoke that the accuracy can not be correctly evaluated. We could override this problem by using perfect hand-made translations only for evaluation purposes.

With respect to list questions, there were ten different questions on every run. We did nothing special to deal with this type of questions in our system. We followed the normal procedure established for every question type, returning up to a maximum of ten answers as indicated in the guidelines. We obtained a P@N of 0.0100, 0, 0.0411 and 0.0200 for *aliv61eses*, *aliv62eses*, *aliv61esen* and *aliv62esen* respectively.

5 Conclusions and Future Work

In this paper we have described the novelties introduce in our Question Answering system for QA@CLEF 2006 competition. This year we have continued with the research line established in 2005, where the main goal is to obtain a fully data-driven system based on machine learning techniques. The main novelties this year are focused on the information retrieval stage. We defined a new passage re-ranking module based on LSA that seems to improve the performance of the system on definition questions, while decreasing the results for factoid ones. This suggest that we could set a good scheme for future work applying the baseline approach for factoid questions while employing re-ranking for definition ones.

Anyway, the results obtained with the re-ranking module were not completely satisfactory. The corpus employed to build de term-by-passage matrix (the 1000 best passages retrieve by Indri on every question) seems inadequate for this task. Probably a bigger amount of text would improve the results. Additionally, different tests varying the reduction of the dimensional space should be done in order to better tune the re-ranking module. The module is still in a preliminary stage and deserves deeper research.

This year we participated in the bilingual Spanish-English task to take advantage of the flexibility of machine learning based systems. The loss of performance with respect to the monolingual Spanish task may be due to the noise introduced by the automatic translation of questions, which makes a bit difficult to correctly evaluate the system.

As a future work, we plan to better test and tune the individual performance of the re-ranking procedure and continue integrating machine learning based modules in our system, going on with the answer extraction stage. Another challenge for the future is to test the system on new languages.

6 Acknowledgements

This work has been developed in the framework of the project CICYT R2D2 (TIC2003-07158-C04).

References

[1] D. Tomás, J. L. Vicedo, M. Saiz, and R. Izquierdo. An XML-Based System for Spanish Question Answering. Lecture Notes in Computer Science. Springer-Verlag, (4022):347-350, 2006.

[2] E. Bisbal, D. Tomás, J. L. Vicedo, and L. Moreno. A Multilingual SVM-Based Question Classification System. MICAI 2005: Advances in Artificial Intelligence. Lecture Notes in Computer Science, (3789):806-815, 2005.

[3] T. K. Landauer, P. W. Foltz, and D. Laham. Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284, 1998.

[4] J. L. Vicedo, R. Izquierdo, F. Llopis, and R. Muñoz. Question answering in Spanish. In: CLEF, editor, Proceedings CLEF-2003 Lecture Notes in Computer Science, Tronheim, Norwey, August 2003.

[5] J. L. Vicedo, M. Saiz, and R. Izquierdo. Does English help Question Answering in Spanish? In: CLEF, editor, Proceedings CLEF-2004 Lecture Notes in Computer Science, Bath, UK, September 2004.

[6] D. Tomás, J. L. Vicedo, E. Bisbal, and L. Moreno. Automatic Feature Extraction for Question Answering Based on Dissimilarity of Probability Distributions. FinTAL 2006: Advances in Natural Language Processing. Lecture Notes in Computer Science, (4139):133-140, 2006.