

Working Towards a Greek-English Cross-Language Question-Answering System

Spyros Methenitis and Leila Kosseim

CLaC Laboratory
Department of Computer Science
Concordia University
1455 de Maisonneuve Blvd. W.
Montreal, Quebec H3G 1M8, Canada

spiros@cu.gr and kosseim@cs.concordia.ca

Abstract

In this paper, we present a set of experiments towards a cross-language Greek-English Question-Answering (QA) system. After exploring previous work done both in monolingual and cross-lingual QA, we have applied the common “question reformulation” strategy used in monolingual systems along with different translation strategies and evaluated the results with the TREC question set. We have mainly focused on the passage retrieval component.

1 Introduction

In this paper we present several experiments performed towards a Greek-English cross-lingual passage retrieval module for QA using a question reformulation strategy and several machine translation mechanisms.

2 Previous Work

In this section we give a brief overview of the monolingual and cross-lingual systems which use similar strategies as the ones we have adopted.

2.1 Monolingual QA

In [1], Bennett et al. show the benefits of using the web as a corpus instead of other smaller document collections, like the TREC corpus which consist of less than 1 million documents. Because of its large size, the web has the advantage of containing redundant information which allows a QA system to use simple question re-writes based on word permutations to find the correct answer. In [1], these re-writes are sent to a search engine (Google) and the system fetches the X first relevant passages for each re-write to be passed as

input to the next component which tries to extract the exact answer.

2.2 Cross-Language QA

The first cross-language QA system that we have studied is the Diogene [2] system. Like other QA systems it is composed of three major components:

1. The Question Processing Component
2. The Search Component
3. The Answer Extraction Component

What is interesting in Diogene is that the first component translates keywords of the question word-by-word rather than using a translation mechanism to translate the whole question. First, DIOGENE extracts all the possible translations for each of the Italian keywords using the Collins Italian/English dictionary. If the translation is not found, MULTIWORDNET [4] is used to translate the word. In case no translation is found and the word is capitalized, it is left as it is. If the word is not capitalized and not found in the Collins dictionary, nor in MULTIWORDNET, it is skipped. The next step is to estimate the probability of every translation in order to find the most plausible. All possible combinations of translation are made and are searched in the target English corpus. Paragraphs that contain at least one English translated word per Italian word are returned. Then, from the paragraphs obtained, translation combinations and their frequency are extracted and the most probable translation is used. The probability of the translation is calculated by $\text{Frequency} / \text{NumberParagraphsInCorpus}$.

The second system that we have studied is Quantum [5] that exists also in monolingual version [6]. To make Quantum cross-lingual, the authors have decided to maintain the English core of the system and to translate only the question

from French into English and the answer back from English into French [5]. They have chosen this approach over having a French core since more resources are available for the English language than for French (WordNet, Annie Linguistic Tools, etc...) and since the other approach would require the translation of documents retrieved which is an error prone task. Nevertheless, the approach chosen required the modification of the first component, the question analysis component. This component now receives as input a French question, so the component should be modified to be able to identify the question type and the question focus in French. Once this is done, the question type and the English translation of the question focus (produced by an IBM2 statistical translation model) are passed to the next component, the answer extraction component. In order for relevant passages to be retrieved, the keywords of the French question should be translated into English too. This was done by using IBM1 translation model, which doesn't take the order of the words into account, and the translation was passed on to OKAPI¹ for passage retrieval. The translation of the extracted answer was not done since it was not required for the CLEF competition, but would be useful in order to make the system transparent to a French user.

BiQue [3] is the last system we have studied. The difference between this system and the previous ones is that instead of using a single translation engine, the system uses three different resources.

1. FreeTranslation (via <http://www.freetranslation.com/>)
2. Altavista (via <http://babel.altavista.com/>)
3. Logos (off-line)

The question is sent to all resources and 3 translations are constructed, one from each. The system then uses all open-class words from all translations to construct the keyword set.

For the question expansion task, the German and English WordNets were used. The goal was to extend the English keyword set collection with synonyms for the words that are present in WordNet.

3 Experiments

In order to evaluate the question re-write strategy for the cross-lingual case, several experiments were performed. This strategy was

¹ <http://www soi.city.ac.uk/~andym/OKAPI-PACK/index.html>

chosen as it does not require much NLP tools, which are lacking in less studied languages such as Greek.

The first experiment is meant to give us an upper bound on the performance we can expect in a Greek-English cross-lingual case. We thus created a monolingual Web-based passage retrieval module for QA based on question re-writes. In the next experiments, we used the previous monolingual module, but this time, we fed it questions that were automatically translated from Greek to English using different translation strategies

All experiments were performed on the TREC question sets.

3.1 The monolingual English experiment

As mentioned above, the first experiment is meant to give us an upper bound on the performance we can expect in the cross-lingual case. We have thus created a generic Web-based passage retrieval module for QA that takes a question, determines its type, removes the question words, generates a number of re-writes, sends these re-writes to a Web search engine and measures each result returned. This allows us in a first step to evaluate how re-writes are useful in a monolingual environment, thus giving us a comparison point for our cross-lingual experiments.

As a first step we have generated simple question re-writes just by moving the verb around. In order to identify the verb we should ideally use a POS tagger, but because this evaluation would be used to compare our Greek system and (at the time) we did not have a Greek POS tagger we have assumed that every word in the question could be a verb and have moved around every word to every position, one at a time.

For example, for question no 1 of TREC-8 (*Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?*) we have generated the following re-writes:

- 1 is the author of the book The Iron Lady ...
- 2 the is author of the book The Iron Lady ...
- 3 the author is of the book The Iron Lady ...
- 4 the author of is the book The Iron Lady ...
- 5 the author of the is book The Iron Lady ...
- 6 the author of the book is The Iron Lady ...

By using this method we have obtained 65,125 re-writes that we have sent to Google. For each re-write we have retrieved the first 10 and then the first 20 documents and snippets. By using TREC's answer patterns we have measured:

1. SN: The number of documents that contain an answer anywhere in the snippet.
2. DOC: The number of documents that contain an answer anywhere in the document.
3. SN-R: The number of documents that contain an answer in the snippet with respect to the re-write - that is, immediately before or immediately after the re-write.
4. DOC-R: The number of documents that contain an answer with respect to the re-write - that is, immediately before or immediately after the re-write.

Tables 1 and 2 show the results we have obtained with different values of X (the number of documents retrieved).

X=20				
TREC	SN	DOC	SN-R	DOC-R
ALL	22.6%	44.4%	0.01%	4.2%
8	8.8%	38.4%	0.0%	3.0%
9	15.3%	39.5%	0.0%	2.9%
10	28.1%	49.6%	0.02%	4.9%
11	35.4%	52.2%	0.0%	7.3%

Table 1: Results of the re-writes in the monolingual English experiment by taking the first 20 documents retrieved (X=20).

X=10				
TREC	SN	DOC	SN-R	DOC-R
ALL	21.5%	44.8%	0.01%	3.8%
8	15.6%	38.3%	0.0%	2.6%
9	15.0%	40.1%	0.0%	2.6%
10	27.1%	49.5%	0.02%	4.5%
11	34.9%	54.3%	0.0%	7.4%

Table 2: Results of the re-writes in the monolingual English experiment by taking the first 10 documents retrieved (X=10).

On average, about 1 document out of 2 contains the expected answer (see the DOC figure), while 1 snippet out of four contains the answer (see the SN figure). However, identifying the answer using only the re-write with no further checking (ex. NP or named entity tagging) seems useless (see the SN-R and DOC-R figures). Although the brute force approach that was used creates a large number of ungrammatical re-writes, we expected these not to be found in the document collection, hence, not affect the results that much.

As Tables 1 and 2 show, the number of documents retrieved (the value of X) does not seem to affect the results significantly.

3.2 Cross-lingual experiment using Systran

In order to test the performance of a Greek-English QA system that might use a translation mechanism, we have used the Systran Greek-English translation system (<http://Systran.otenet.gr>). Here, we first needed a corpus of Greek questions. We therefore translated the TREC-8 questions into Greek manually. Then we took the Greek questions and ran them through Systran to translate them back to English. Next we took the English questions that Systran had generated and ran them through the monolingual QA search-engine evaluation script described in section 3.1 with X=20. Finally, we compared the results we obtained with this cross-lingual version with the monolingual results of section 3.1.

Table 3 shows the results of this experiment. As the table shows, the cross-lingual experiment yielded much lower results. We therefore tried to use a different translation mechanism in order to identify if the low results were due to a poor translation from Systran.

	SN	DOC	SN-R	DOC-R
Systran	2.3%	6.4%	0.0%	0.0%
Original	8.8%	38.4%	0.0%	3.0%

Table 3: Results of the re-writes in the cross-lingual experiment with Systran (X=20) with the TREC-8 questions.

3.3 Trying to improve results using word for word translation

In this last experiment, we therefore tried another method of translation. We used a bilingual dictionary and sent every word of the question to the dictionary. We then used all possible translations of every word and created a boolean expression to be used by the search engine.

The dictionary selected was the online dictionary www.in.gr from www.in.gr/dictionary/lookup.asp. We chose this dictionary because it was easily available and because of its capability of providing synonyms of words if we wished to increase the recall.

Before sending all words of the Greek manually translated questions to the dictionary we had to do a number of tasks. First, we tagged all "Foreign Words", meaning words that after the manual translation in Greek, still remained in English, so we wouldn't have to send them for translation. Then, we tagged all number and date expressions of the question for the same reasons. We also removed the question word, since the dictionary was unable to provide a proper translation of question words. Due to the structural and

grammatical differences between the two languages, only a component specifically designed to translate the question word from English to Greek by using information from other words in the question would be able to perform such a task.

In order to perform these operations we used a Greek POS tagger developed by ILSP (www.ilsp.gr)² and we have tagged all TREC-8 questions that were manually translated in Greek. Then, a script took care of the operations performed before sending each word for translation.

Another script then sent every word of every question to the dictionary and obtained one or more possible translations from which a boolean expression was formulated. An example of this process is illustrated in Figure 1.

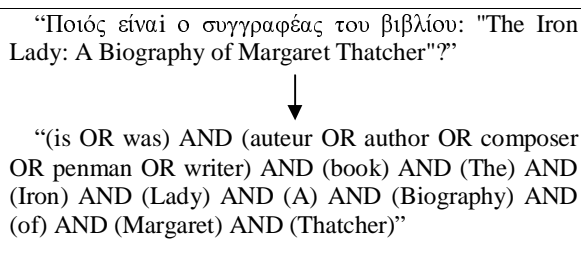


Figure 1: Example of a Boolean query formulated for the Greek version of the question *Who is the author of the book, “The Iron Lady: A Biography of Margaret Thatcher”?*

Once the query was formulated, we sent them to 3 search engines: Google, Altavista by using the AND operator and Altavista by using the NEAR operator.

3.3.1 Using Google

As in the previous experiments, we used Google as our search engine. However, Google had a serious disadvantage because it imposes a maximum of 10 keywords per query. However, since for every word in a question we used different possible translations and synonyms, the number of keywords per query is significant. In fact, the boolean queries have an average length of 25.75 words. For many of the questions, not all words of the query were therefore used by Google. In any case, we report the results with Google in Table 4.

3.3.2 Using Altavista with AND

Because Google did not consider all keywords in the query, we turned to Altavista. We used the AND operator to connect the different question words.

² This POS tagger became available too late for experiment 1.

3.3.3 Using Altavista with NEAR

Finally, we tried the same test with Altavista by using the NEAR operator to connect the different question words.

4 Overall comparison

Table 4 summarises the results of all experiments. Table 4 shows both the monolingual and the cross-lingual cases. Four measures are given:

1. SN: The number of documents that contain an answer anywhere in the snippet; as in the previous tables.
2. DOC: The number of documents that contain an answer anywhere in the document; as in the previous tables.
3. Q-DOC: The number of questions for which a correct answer was found in a document.
4. Q-SN: The number of questions for which a correct answer was found in a snippet.

DOC-R and SN-R are not given here because they assume that the query will be a re-write of the original question that can lead to a grammatically correct sentence (ex. *The author of the book “The Iron Lady: A Biography of Margaret Thatcher” is*). This is true in the case of the monolingual system and the Systran-based translation, but in the word-for-word translation, only queries based on a bag of keywords are produced. The SN-R and DOC-R measures are therefore irrelevant here.

	Mono-lingual	Cross-lingual			
		Systran	Word for Word translation		
			Google	Altavista using AND	Altavista using NEAR
DOC	38.4%	8.8%	12.4%	24.2%	16.2%
SN	8.8%	2.3%	3.4%	1.0%	1.0%
Q-DOC	91.1%	61.8%	37.0%	60.5%	49.0%
Q-SN	61.3%	7.2%	17.0%	4.0%	5.5%

Table 4: Comparison of the results for all experiments on TREC-8 with X=20

As the table shows, the word for word translation strategy scored overall better than the Systran translation at the document level (DOC). This means that more documents contain the correct answer for a question. The best results are obtained by Altavista using the AND operator for the documents (24.2%) and surprisingly for the snippets the best results are obtained by Google (3.4%) by using the AND operator.

Altavista’s NEAR operator did not improve the results as expected.

The fact that Systran had 61.8% of the questions with a correct answer found in some document while only 8.8% of the documents contained a correct answer indicates that when Systran makes a correct translation of the question, most of the documents returned for this question contain the correct answer. On the other hand, when Systran fails to make a correct translation, most of the documents returned from the search engine for this question do not contain the correct answer.

The fact that Altavista for the word for word translation strategy had 60.5% of the questions with a correct answer found in some document and 24.2% of the documents contained a correct answer indicates that there was a better distribution of documents containing correct answers with respect to the questions.

To obtain the optimal result, we should have a system that would use the word for word translation strategy and use Google to retrieve snippets and Altavista to retrieve documents.

5 Conclusion and Future Work

A lot of work could be done from here. First of all these experiments could have been performed by using a different search engine or different values of X, the number of documents retrieved per question (X=20 in our tests).

Also, since by using the word for word translation strategy we obtained a larger number of documents containing the correct answer per question, meaning more questions had documents that contained the correct answers, we could significantly increase the value of X (e.g. 50) and then on the retrieved documents we could run a different algorithm that would try to retrieve passages that contain a significant number of words of the translated question within a given distance.

Another interesting experiment would be to implement our own NEAR operator and run the experiment with different values of maximum distance between words for this operator (10 for Altavista). Then we could find out what the optimum maximum distance for the NEAR operator is to use in such a system.

Furthermore since Systran performs better when the translation succeeds and worse when it fails we could use Systran for passage retrieval and check how it performed by measuring the number of documents returned containing a correct answer. If Systran performed poorly the system would use the word for word translation strategy with Altavista

and the AND operator, else the system would use Systran.

We should continue to work hard in this field because there is much to be done, especially for less studied languages where there is a lack of tools or the tools available aren't as efficient as we need.

References

- [1] P. Bennett, S. Dumais and E. Horvitz (2002). Web question answering: Is more always better? In *Proceedings of SIGIR'02*, August, pp. 291-298.
- [2] M. Negri, H. Tanev, and B. Magnini (2003) Bridging Languages for Question Answering: DIOGENE at CLEF-2003. In *Proceedings of Cross-Language Evaluation Forum (CLEF-2003)*, Trondheim, Norway, August.
- [3] G. Neumann and B. Sacaleanu (2003) A Cross-Language Question/Answering-System for German and English. In *Proceedings of Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, August.
- [4] E. Pianta, L. Bentivogli and C. Girardi.(2002) MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, Mysore, India, January.
- [5] L. Plamondon and G. Foster (2003) *Quantum, a French/English Cross-language Question Answering System*. In *Proceedings of Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, August.
- [6] L. Plamondon, G. Lapalme and L. Kosseim (2002) The QUANTUM Question-Answering System at TREC-11. In *Proceedings of the 11th Text REtrieval Conference (TREC-11)*, pp. 750-757. November, Gaithersburg, Maryland, USA.