Answer Translation: An Alternative Approach to Cross-Lingual Question Answering

Johan Bos[1] and Malvina Nissim[2]

[1] Linguistic Computing Laboratory
Department of Computer Science
University of Rome "La Sapienza", Italy
bos@di.uniroma1.it
[2] Dept. of Language and Oriental Studies
Faculty of Humanities
University of Bologna, Italy
malvina.nissim@unibo.it

Abstract. We approach cross-lingual question answering by using a mono-lingual QA system for the source language and by translating resulting answers into the target language. As far as we are aware, this is the first cross-lingual QA system in the history of CLEF that uses this method—almost without exception, cross-lingual QA systems use translation of the question or query terms instead. We demonstrate the feasibility of our alternative approach by using a mono-lingual QA system for English, and translating answers and finding appropriate documents in Italian and Dutch. For factoid and definition questions, we achieve overall accuracy scores ranging from 13% (EN→NL) to 17% (EN→IT) and lenient accuracy figures from 19% (EN→NL) to 25% (EN→IT). The advantage of this strategy to cross-lingual QA is that translation of answers is easier than translating questions—the disadvantage is that answers might be missing from the source corpus and additional effort is required for finding supporting documents of the target language.

1 Introduction

Question Answering (QA) is the task of providing an exact answer (instead of a document) to a question formulated in natural language. Cross-lingual QA is concerned with providing an answer in one language (the target language) to a question posed in a different language (the source language). Most systems tackle the cross-lingual problem by translating the question (or query terms) posed in the source language in the target language, and then using a mono-lingual QA system developed for the target language for retrieving an answer.

Surprisingly little attention has been given to an alternative approach: translating the answer, instead of the question. The main advantage of this method is that answers are easier to translate than questions, due to their simpler syntactic structure. In fact, some types of answers (such as date expressions or names of persons) hardly need a translation at all. In addition, finding a document

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 290–299, 2007.

in the target language supporting the answer (as prescribed in the QA@CLEF exercise) is feasible: using either the translated question or keywords thereof and the translated answer, standard document retrieving tools can be used to find a document. This approach can work provided that the source and target language documents cover the same material, otherwise all bets are off.

In the context of CLEF, we tested this approach relying on the fact that the documents in the collection of newswire articles come from the same time period, and hence are likely to cover approximately the same topics. We ran our experiments for two language pairs, with English as source language, and Dutch (EN→NL) and Italian (EN→IT), as target language, respectively. We used an existing mono-lingual QA system for English and an off-the-shelf machine translation tool for translating the answers. In this report we describe and evaluate our method in the context of CLEF, and discuss the feasibility of this approach in general.

2 Related Work

Almost without exception, the method used in cross-lingual QA for crossing the language barrier is translating the question of the source language into the target language, and then using a mono-lingual QA system for the target language. A variant on this method is translating the query terms derived from the analysis of the source question into terms of the target language. Both methods apply the translation step very early in the pipeline of QA components.

To get an accurate overview of all the approaches, we carried out a survey on the methods used in all the editions of cross-lingual QA@CLEF, based on information gathered from the working notes of CLEF 2003 [7], CLEF 2004 [10], CLEF 2005 [8] and CLEF 2006 [9]. In these four years of CLEF we counted 44 systems participating in 68 (26 different ones) cross-lingual tasks. Specifically, the tasks considered were BU→EN (3x), BU→FR, DE→EN (5x), DE→FR, EN→DE (4x), EN→ES (6x), EN→IT (2x), EN→DN, EN→FR (5x), EN→NL (3x), EN→PT (2x), ES→EN (4x), ES→FR, ES→PT, FI→EN (2x), FR→EN (11x), FR→ES, IT→EN (3x), IT→ES, IT→FR (2x), NL→EN, NL→FR, PT→EN, PT→FR (3x), IN→EN (2x), and RO→EN.

In the majority of cases (63 of 68), the cross-lingual problem was addressed by using off-the-shelf translation software (such as Babelfish, Systran, Reverso, FreeTrans, WorldLingo, Transtool) to translate the question or the keywords of the query of the source language into the target language, followed by processing the translated question or query using a mono-lingual QA system.

The alternative method we propose applies the translation step at the last stage of the QA pipeline. Instead of translating the question, a mono-lingual QA system for the source language produces an answer which is subsequenlty translated in the target language. In the CLEF-2006 campaign we applied our method to the language pairs EN→IT and EN→NL. As far as we are aware, this is the first time, in the context of QA@CLEF, that such an approach was implemented and evaluated. A similar approach to ours was also presented at CLEF 2006 by LCC [3]. They ran a QA system for the source language on an automatically translated (from target to source language) document collection, and then aligned the answer found in the source language document with that of the answer and document in the target language. This was done for three language pairs: EN→FR, EN→ES, and EN→PT. Although our method is similar, it does not require the effort to translate the entire document collection, and focuses on just translating the obtained answer.

3 Method

Our cross-lingual QA system is based on a mono-lingual QA system for English extended with an answer translation module. It deals with factoid (including list questions) and definition questions, but it uses two different streams of processing for these two types of questions. The first component in the pipeline, Question Analysis, deals with all types of questions. Then, when the question turns out to be of type factoid, Document Analysis, Answer Extraction, and Answer Translation (and document support) will follow. Answers to definition questions are directly searched in the target language corpora (see Figure 1 and Section 3.5). What follows is a more detailed description of each component. An example of the different data structures of the system is shown in Fig. 2.

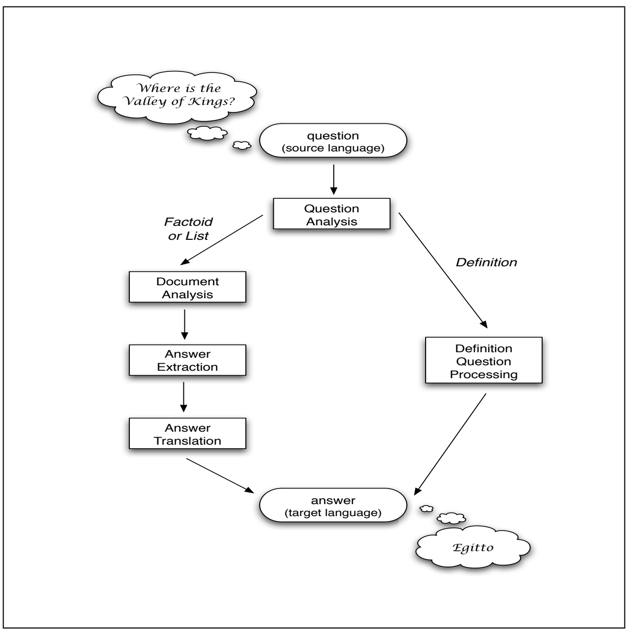

Fig. 1. Simplified architecture of the QA system

3.1 Question Analysis

The (English) question is tokenised and parsed with a wide-coverage parser based on Combinatory Categorial Grammar (CCG). We use the parser of Clark & Curran [4]. On the basis of the output of the parser, a CCG-derivation, we build a semantic representation in the form of a Discourse Representation Structure (DRS), closely following Discourse Representation Theory [5]. This is done using the semantic construction method described in [1,2]. The Question-DRS is the basis for generating four other sources of information required later in the question answering process: an expected answer type; a query for document retrieval; the answer cardinality; and background knowledge for finding appropriate answers.

We distinguish 14 main expected answer types which are further divided into subtypes. The main types are definition, description, attribute, numeric, measure, time, location, address, name, language, creation, instance, kind, and part. The answer cardinality denotes a range expressed by an ordered pair of two numbers, the first indicating the mininal number of answers expected, the second the maximal number of answers (or 0 if unspecified). For instance, 3–3 indicates that exactly three answers are expected, 2–0 means at least two answers. Background knowledge is a list of axioms related to the question—it is information gathered from WordNet or other lexical resources.

3.2 Document Analysis

In order to maximise the chance of finding an answer in the source language (English), we extended the English CLEF document collection with documents from the Acquaint corpus. All documents were pre-processed by applying sentence splitting and tokenisation and dividing them into smaller documents of two sentences each (taking a sliding window, so each sentence will appear in two mini-documents). These mini-documents were indexed with the Indri information retrieval tools (we used version 2.2, see `http://www.lemurproject.org/indri/`). The query generated by Question Analysis (see Section 3.1) is used to retrieve the best 1,000 mini-documents, again with the use of Indri.

Using the same wide-coverage parser as for parsing the question, all retrieved documents are parsed and for each of them a Discourse Representation Structure (DRS) is generated. The parser also performs basic named entity recognition for dates, locations, persons, and organisations. This information is used to assign the right semantic type to discourse referents in the DRS.

3.3 Answer Extraction

Given the DRS of the question (the Q-DRS), and a set of DRSs of the retrieved documents (the A-DRSs), we match each A-DRS with the Q-DRS to find a potential answer. This process proceeds as follows: if the A-DRS contains a discourse referent of the expected answer type (see 3.1) matching will commence attempting to identify the semantic structure in the Q-DRS with that of the A-DRS. The result is a score between 0 and 1 indicating the amount of semantic material that could be matched. The background knowledge (such as hyponyms from WordNet) generated by the Question Analysis (see 3.1) is used to assist in the matching. All retrieved answers are re-ranked on the basis of the match-score and frequency.

3.4 Answer Translation

The answer obtained for the source language is now translated into the target language by means of the Babelfish off-the-shelf machine translation tool. However, this is not done for all types of answers: we refrain from translating answers that are person names or titles of creative works. Indeed, person names are often not translated across languages. Names of creative works instead can be, but the machine translation software that we use does not perform well enough on some examples in our training data, so that we decided to leave these untranslated, too. In addition, titles of creative works often don't get a literal translation (a case in point is Woody Allen's "Annie Hall", which is translated in Italian as "Io e Annie") so a more sophisticated translation strategy would be required.

Document Support. Given the answer in the target language, we need to find a supporting document from the target language collection (as prescribed by the QA@CLEF exercise). Hence, we also translate the original question, independently of the answer found for it. This is used to construct another query for retrieving a document from the target language collection that contains the translated answer and as many as possible terms from the translated question. (As with the source language documents, these are indexed and retrieved using Indri.)

3.5 Definition Question Processing

We did not put much effort in dealing with definition questions. We simply adopted a basic pattern matching technique directly on the target language documents. Once a question is identified as a definition question (see Figure 1), all non-content words are removed from the question (wh-words, the copula, articles, punctuation, etc.). What is left is usually just the topic of the definition question. For instance, for the English question "Who is Radovan Karadzic?" we derive the topic "Radovan Karadzic".

Given a topic, we obtain interesting information about it expressed in the target language. We do this by searching an off-line dump of the Dutch and Italian Wikipedia pages for sentences mentioning the target. From this we generate an Indri query selecting all (one-sentence) documents containing the topic, and a combination of the terms found in Wikipedia (removing stop words). This yields a list of documents in the target language.

The last step is based on template matching using regular expressions to extract the relevant clauses of a sentence that could count as an answer to the question. There are only a few templates, but we use different ones for the different sub-types of definition questions (the Question Analysis component distinguishes between three sub-types of definition questions: person, organisation, and thing). We developed a few general patterns tested on definition questions from previous QA challenges, which were both valid for Dutch and Italian:

```
person:        /(^| )(\,|de|il|la|i|gli|l\’) (.+) $topic /i
person:        /$topic \( (\d+) \)/i
organisation:  /$topic \(([^\)]+)\)/i
thing:         /$topic , ([^,]+) , /
thing:         /$topic \(([^\)]+)\)/i
```

As an example of what these simple patterns can achieve, consider the following sentences (extracted answers are type-set in bold-face):

LASTAMPA94-041510	Il **leader** Radovan Karadzic non era reperibile.
LASTAMPA94-042168	Intanto il **leader serbo-bosniaco** Radovan Karadzic e il comandante in capo delle forze serbe, gen.

4 Evaluation

4.1 Performance at CLEF-2006

The question analysis component performed fairly well. Only 2 of 200 questions of the EN→IT set, and 7 of EN→NL set could not be parsed. For 189 questions of the EN→IT set, and 177 of the EN→NL set the system determined an expected answer type. This already shows that our system had more difficulties with the EN→NL set, which is reflected in the overall scores.

Relatively many questions did not have an answer in the English collection, or at least our system failed to find one. For 41 of the 200 EN→NL questions and for 43 of the 200 EN→IT questions no English answer (correct or incorrect) was found. Answer translation introduced only few errors (see Section 4.2), but finding a supporting document proved harder than we had hoped. So several correct answers were associated with wrong documents and therefore judged as “unsupported” (see below).

We submitted four runs—two for the EN→NL task, and two for the EN→IT task. The first runs contained one answer for factoid questions and up to ten for definition and list questions. The second runs contained up to ten answers for each type of question. We used this strategy because it was unclear, at the time of submission, what kind of evaluation would be used. So the number one runs would perform better on an accuracy score, and the number two runs better on scores based on the average mean reciprocal rank (MRR).

Eventually, both accuracy and MRR were used in the evaluation (see [6]). The results for definition and factoid questions are shown in Table 1, and the figures for the list questions in Table 2. As the tables illustrate, the scores for accuracy are the same for each run, which means that providing more than one answer for a factoid question was not punished. As expected, we achieved better scores for MRR on the number two runs. It is surprising that we did so well on definition questions, given that we paid very little effort in dealing with them.

Input Question:	[0040] What year did Titanic sink?
Q-DRS:	[x1 | year(x1)] ? [x2 x3 | named(x3,titanic,nam) sink(x2) agent(x2,x3) rel(x2,x1)]
Answer Type:	[tim:yea]
Cardinality:	1–1
English Query:	#filreq(Titanic #weight(1 Titanic 3 sink))
English Answer:	1912
English Context:	[APW19981105.0654] Speaking of bad guys: Cartoon News reprints a bunch of political cartoons published after the <u>Titanic</u> <u>sank</u> <u>in</u> **1912**.
A-DRS:	x0 x1 x2 x3 speak(x1) agent(x1,x0) bad(x2) guy(x2) proposition(x3) :(x2,x3) x3: [x4 x5 x6 x6 x7 x8 x9 named(x4,cartoon_news,org) bunch(x5) political(x6) publish(x7) patient(x7,x6) proposition(x8) after(x7,x8) x8: [x10 x11 x12 named(x10,titanic,nam) timex(x12)=1912-XX-XX sink(x11) agent(x11,x10) in(x11,x12)] cartoon(x6) of(x5,x6) reprint(x9) agent(x9,x4) patient(x9,x5)] of(x1,x2)
Italian Answer:	1912
Italian Query:	#filreq(1912 #combine(1912 anno Titanic affondato))
Italian Context:	[AGZ.941120.0028] I cantieri di Belfast hanno confermato di aver ricevuto la commessa più prestigiosa dopo la costruzione del "Titanic": la messa in mare del "Titanic 2". Una copia perfetta del fantasmagorico transatlantico affondato nell' oceano nel 1912 durante il viaggio inaugurale è stata ordinata da un consorzio giapponese presso la ditta "Harland and Wolff" e sarà completata entro il 1999.

Fig. 2. System input and output for a factoid question in the EN→IT task

Table 1. Number of right (R), wrong (W), inexact (X), unsupported (U) answers, accuracy measure of first answers (factoid, definition, overall, and lenient), and mean reciprocal rank score (MRR) for all four submitted CLEF 2006 runs

Run	R	W	X	U	Accuracy Fact. Acc.	Def. Acc.	Overall	Lenient	MRR
EN→IT 1	32	141	4	11	15.3%	24.4%	17.0%	25.0%	0.18
EN→IT 2	32	141	4	11	15.3%	24.4%	17.0%	25.0%	0.20
EN→NL 1	25	150	6	3	11.6%	20.5%	13.4%	18.5%	0.14
EN→NL 2	25	149	7	3	11.6%	20.5%	13.4%	19.0%	0.15

Table 2. Number of right (R), wrong (W), inexact (X), unsupported (U), unassessed (I) answers, P@N (percentage of correct answers in returned set) score for list questions, for all submitted CLEF 2006 runs

Run	R	W	X	U	I	P@N
EN→IT 1	12	63	1	6	0	0.10
EN→IT 2	15	70	1	7	0	0.15
EN→NL 1	6	29	0	7	24	0.18
EN→NL 2	9	40	0	5	42	0.15

4.2 Answer Translation Accuracy

Recall that depending on the kind of expected answer type, answers were translated into the target language or not. We did not translate names of persons, nor titles of creative works. Whereas this strategy worked out well for person names, it didn't for names of creative works. For instance, for question "0061 What book did Salman Rushdie write?", we found a correct answer "Satanic Verses" in the English documents, but this was not found in the Italian document collection, because the Italian translation is "Versetti Satanici". Babelfish wouldn't have helped us here either, as it translates the title into "Verses Satanic".

Table 3. Answer translation accuracy, divided over answer types (EN→NL). Quantified over the first translated answer, disregarding whether the answer was correct or not, but only for questions to which a correct expected answer type was assigned.

Answer Type	Correct	Wrong	Accuracy
location	30	0	100%
numeric	9	2	82%
measure	4	2	67%
instance (names)	14	5	74%
time	15	0	100%
other	7	1	88%

Overall, answer translation introduced little errors as many answers are easy to translate (Table 3). Locations are usually correcly translated by Babelfish, and so are time expressions. Difficulties sometimes arise for numeric expressions. For instance, for the question "0084 How many times did Jackie Stewart become world champion?" we found the correct English answer "three-time". However, this was wrongly translated in "drie-tijd", and obviously not found in the Dutch corpus. Similarly, many answers for measurements are expressed in the English corpus using the imperial system, whereas the metric system is used in the Dutch and Italian corpora. Finally, some things are just hard to translate. Although we found a reasonably correct answer for "0136 What music does Offspring play?", namely "rock", this was translated in Dutch as "rots". In itself a correct translation, but for the wrong context.

5 Discussion

What can we say about the answer-translation strategy to cross-lingual question answering, and how can we improve it? Generally speaking, we believe it is a promising approach, but success depends on improvement in three areas: translation itself, source document coverage, and target document support.

Translation. Generally, answers are easier to translate than questions since they are syntactically less complex. For many answer types, word sense ambiguity can cause erroneous translations. In addition, there are issues specific to certain answer types. Table 4 summarises and exemplifies some of them.

Table 4. Aspects of answer translation in different answer types

Answer Type	Aspect	Source	Target
numeric	punctuation	100,000 [EN]	100.000 [IT]
measure	unit conversion	26 miles [EN]	42 km [NL]
time	formatting	the 3rd of January, 1982 [EN]	3 Januari 1982 [NL]
location	spelling	London [EN]	Londra [IT]
		New York [IT]	New York [IT]
creation	translation	The Office [EN]	The Office [IT]
		Annie Hall [EN]	Io e Annie [IT]

As Table 4 shows, each answer type comes with specific problems in translation. For instance, for creative works and locations, sometimes terms are translated and sometimes they aren't. It seems that many difficulties in translating answers might be addressed by dedicated translation strategies and look-up tables for answers expressing measure terms and titles of creative works. In general, knowing the answer types is a great advantage for obtaining a high-quality translation.

Source Document Coverage. One problem that came to the surface was the omission of the answer in the source language document collection. There is only one way to deal with this situation, and that is getting a larger pool of documents. One option is to use the web for finding the answer in the source language.

Target Document Support. Another problem that arised was finding a supporting document for the target language. The system can certainly be improved with respect to this point: it currently only takes the translated question as additional information to find a target language document. This is interesting, as the machine translated question need not be grammatically perfect to find a correct document. One way to improve this is to translate the context found for the source language as well, and use it in addition to retrieve a target language document.

References

1. Bos, J., Clark, S., Steedman, M., Curran, J.R., Hockenmaier, J.: Wide-Coverage Semantic Representations from a CCG Parser. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING '04), Geneva, Switzerland (2004)
2. Bos, J.: Towards wide-coverage semantic interpretation. In: Proceedings of Sixth International Workshop on Computational Semantics IWCS-6, pp. 42–53 (2005)
3. Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J., Moldovan, D.: LCC's PowerAnswer at QA@CLEF 2006. In: Peters (ed.) Working Notes for the CLEF 2006, September 20–22, Workshop, Alicante, Spain (2006)
4. Clark, S., Curran, J.R.: Parsing the WSJ using CCG and Log-Linear Models. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04), Barcelona, Spain (2004)
5. Kamp, H., Reyle, U.: From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. Kluwer, Dordrecht (1993)
6. Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Osenova, P., Penas, A., Jijkoun, V., Sacaleanu, B., Rocha, P., Sutcliffe, R.: Overview of the CLEF 2006 Multilingual Question Answering Track. In: Peters (ed.) Working Notes for the CLEF 2006 Workshop,September 20–22, Alicante, Spain (2006)
7. Peters, C.: Results of the CLEF 2003 Cross-Language System Evaluation Campaign. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds.) CLEF 2003. LNCS, vol. 3237, pp. 21–22. Springer, Heidelberg (2004)
8. Peters, C.: Results of the CLEF 2005 Cross-Language System Evaluation Campaign. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 21–23. Springer, Heidelberg (2006)
9. Peters, C.: Results of the CLEF 2006 Cross-Language System Evaluation Campaign. In: Working Notes for the CLEF 2006 Workshop, September 20–22, Alicante, Spain (2006)
10. Peters, C., Borri, F.: Results of the CLEF 2004 Cross-Language System Evaluation Campaign. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 15–17. Springer, Heidelberg (2005)