Cross Lingual QA: A Modular Baseline

Lucian Vlad Lita, Monica Rogati, Jaime Carbonell
{llita, mrogati, jgc}@cs.cmu.edu
Carnegie Mellon University

Abstract

This work evaluates the applicability of currently available research tools to the problem of cross lingual QA. We establish a task baseline by combining a cross lingual IR system with a monolingual QA system in a very short amount of time. A higher *precision* strategy involves applying the monolingual QA system to an automatically translated question, assumed to be correct. A higher *coverage* strategy consists of a term weighted proximity measure with varied query expansion, tuned for each individual question type.

1 Introduction

In our CLEF 2003 participation we evaluate the application of existing research modules to Cross-Lingual Question Answering (CLQA). The obvious first step towards solving this task is to combine cross lingual information retrieval with monolingual question answering. In order to set up a baseline with very little effort – one week's worth of work – we glued two existing off-the-(authors' research)-shelf components: a cross lingual information retrieval system and a monolingual question answering system, and tuned them on available question/answer datasets.

We have participated with two runs in the cross-lingual French-to-English CLEF task and we focused on quickly obtaining a system based on available tools and components.

2 Overview

Our CLEF system consists of two pre-existing components: a cross lingual information retrieval system (CLIR) and a monolingual question answering system (MQA), plus the necessary glue.

2.1 The CLIR Component

The cross lingual information retrieval component [*4*] is a system trained for the CLEF 2003 cross lingual retrieval task. It uses a parallel corpus to train a translation model, which is then used for query translation. The system uses GIZA++ [*2*] to train the translation model and a retrieval system based on Lemur [*3*]. No proprietary machine translation systems including SysTran, Google etc, have been used and the parallel corpus is freely available.

The CLIR system produces both a list of relevant documents as well as a translated expanded query with corresponding weights for each word.

2.2 The MQA Component

The monolingual question answering component is a high precision, pattern based QA system that relies on very few resources. The system is trained on the TREC [*6*] QA task datasets and has a limited question type coverage.

The MQA system implements a simplified version of the widespread pipeline QA architecture. Initially, the questions are filtered and classified into question types and relevant question terms are extracted. A straightforward sentence-level retrieval follows, producing candidate sentences in tokenized form. In the answer extraction phase, high confidence finite state transducers (FSTs) are applied and candidate answers are produced with their corresponding confidence scores. Answers with similar surface forms are grouped and unified into a stronger representative answer with a higher score.

Currently, no answer verification is performed and no feedback loops are present. The MQA system was built to rely on as few resources as possible. Hence, the transducers are based on surface form, capitalization features, WordNet

[7] based features, and a short list of grammatical constructs. Named entity taggers and gazetteers are the two ubiquitous elements in QA architectures that are missing from our system.

2.3 Architecture

Our simple, ad-hoc architecture sets up an obvious baseline for the CLQA task. We approach the problem through two methods: a high precision method that is likely to answer few questions – especially given imperfect translations – and a higher recall method that covers most of the questions.

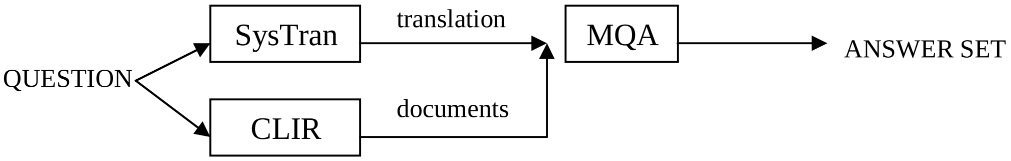

Figure 1. High precision approach

In our higher precision method, the source language question is first run through SysTran [5], a proprietary translation system with a free, limited, online interface. The un-altered English translation is then passed to the MQA component, which produces a list of answers, ordered by confidence scores. The documents used in answer extraction are produced by the CLIR component. If any answers are obtained via this strategy, the system offers them as the final set of answers. In case no answers were extracted, the higher recall method is activated.

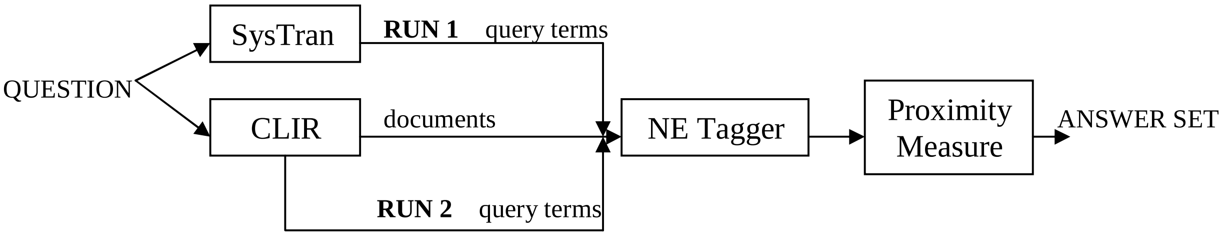

Figure 2. High coverage approach. In RUN 1, the query terms are produced directly from the SysTran translation. In RUN 2, the query terms are produced by the CLIR component.

Our higher coverage strategy also uses the CLIR component for document retrieval. A rudimentary question classification provides the mapping between the question text and answer types. The answer types correspond to entity types obtained by processing the relevant documents with a named entity tagger. Subsequently, given a set of question terms, we apply a term weighted proximity measure to candidate answers of the required type. Evidence for a particular answer is then combined with that of identical or similar answers in order to compute its final score.

The difference between our two CLEF runs consists in the way the question terms are produced. In the first run, the words in the SysTran translation are filtered using a stopword list, then used with the proximity measure. For the second run, the CLIR component takes over the query expansion and produces a weighted set of terms to be further applied in computing the proximity score.

3 Experiments

In the high precision approach experiments, the system first translates the question from French to English using the SysTran online interface. The un-altered, translated question is considered grammatically and semantically correct – i.e. a perfect translation – and is passed to the MQA component. Relevant documents are retrieved using the CLIR component and are also passed to the MQA component, which applies FSTs/patterns to identify candidate answers.

Using the higher coverage strategy, our system first classifies each question as one of the following types: location, temporal, person, numeric, quantity and other. We obtain the relevant documents using the CLIR component, we tokenize them, and we apply a simple sentence splitting algorithm. The BBN Identifinder [*1*] named entity tagger is applied to the processed documents. Entities corresponding to the required question type are identified as candidate answers. For each candidate answer, we compute a term weighted proximity score:

$$\text{score} = \Sigma_i (w_a * f(d_i) * w_{qi})$$

where w_a is the term weight for the candidate answer, $f(d_i)$ is a function based on the distance between the candidate answer and i^{th} question term in its proximity, and w_{qi} is the term weight for the i^{th} question term. The term weighting methods considered were: *okapi*, *idf*, and *ntc*. The distance functions explored were the linear, quadratic, and exponential functions.

3.1 Parameter tuning

The training set consisted of approximately one hundred questions selected from TREC 9 & 10 question sets. The original English questions were translated by two French native speakers[♣]. The automatic SysTran translations were corrected in order to provide our system with reasonable training data. We used the limited data for parameter tuning for each individual question type. *Table 1.* shows the final parameter set used for the CLEF runs.

Question Type	**Term Weighting**	**Distance function**	**# expansion terms**	**# documents**
Location	NTC	Linear	10	20
Person	Okapi	Linear	50	20
Temporal	Okapi	Exponential	10	20
Quantity	Okapi	Quadratic	10	20
Numeric	Okapi	Exponential	10	20

Table 1. Final parameters for the CLEF cross lingual QA task

The number of query expansion terms was varied from 5 to 200 for the second run. For the first run, the query terms considered were only the non-stopwords in the SysTran translation of the source question. The number of relevant documents retrieved by CLIR was varied from 1 to 30.

3.2 Performance at CLEF

The best MRR score we obtained was 0.1533 under the stricter policy, and 0.17083 under the looser. Our system did not offer any answer for nearly half of the questions, reflecting the fact that it is very conservative in terms of proposing answers with little evidence.

The monolingual component was trained on short, correct and meaningful English factoid questions. Out of 200 questions, the MQA approach worked on 11 questions, which were translated approximately correct. Out of these 11, it found correct answers for only 4 questions. For the other 7 questions it produced no answers.

	Strict MRR	**% questions w/ a correct answer**	**Loose MRR**	**% questions w/ a correct answer**	**# NILs proposed**	**# NILs correct**
Run 1	0.1533	19%	0.17083	21%	92	8
Run 2	0.1316	15.5%	0.14916	17.5%	91	7

Table 2. Final CLEF scores for both runs – test set contains 200 questions

[♣] many thanks to Antoine Raux

4 Discussion

Our CLEF 2003 system combines the CLIR and MQA research modules into a CLQA baseline. There are several modifications which could clearly improve the performance. We have employed no question analysis in French. Phrase detection, reformulation, classification by question type and answer type in the source language (French in our case) would clearly improve the performance. Since early errors propagate through the question answering pipeline, accurate question analysis before automatic translation would allow the system to select the appropriate answer extraction method.

The current question analysis in the target language (English) is a minimal classification into 6 question types. Since automatic translation is less than perfect, phrase identification and reformulation is almost always out of the question.

The training data consists of factoid questions and answers selected from TREC datasets. The CLEF question set appears to contain more complex questions and questions that contain more content words compared to the TREC style questions. However the fact that our system was tuned on *slightly* different data than the CLEF test data is not detrimental since it approaches a more likely real life test.

The retrieval step is tuned for the CLEF 2002 cross lingual IR task, and the query expansion is performed internally by the CLIR component. A document set of size 20 worked best for all question types. For fewer documents, there was not enough support for correct answers and for a larger document set size there was too much noise. The *person* question type required a wider query expansion.

The two runs showed that for this year's CLEF data, the CLIR-produced query expansion in English from the source question text in French is almost as good as using the question terms from the SysTran translation. This is particularly true for longer questions that involve more high-content words. The fact that proprietary components can play a more limited role in the system without drastic performance drops is certainly encouraging because it provides more control over the CLQA process.

The fact that the two runs were very similar in performance suggests that no question expansion – i.e. using the translated question terms in answer identification – is sufficient given our current basic methods. On the other hand, the lack of exact (SysTran) translation for the higher coverage strategy does not result in significant performance degradation.

5 Future Work

Our goal for this year's CLEF participation was to identify research issues for cross lingual question answering using a minimal baseline. In future research, we plan to explore issues such as language dependencies and semi-automatic question analysis for CLQA.

6 References

[1] Bikel D.M., Miller S., Schwartz R., and Weischedel R. (1997*). Nymble: A High-Performance Learning Name-Finder*. Proceedings of ANLP.

[2] Och F.J., Ney H. (2000). *Improved Statistical Alignment Models*. Proceedings of ACL.

[3] Ogilvie P., Callan J. (2001). *Experiments using the Lemur toolkit*. TREC.

[4] Rogati M., Yang Y. (2003) CONTROL: CLEF-2003 with Open, Transparent Resources Off-Line, CLEF.

[5] SysTran, http://babel.altavista.com/translate.dyn

[6] Voorhees E. (2002). *Overview of the TREC 2002 Question Answering Track*. Text Retrieval Conference.

[7] *WordNet: An Electronic Lexical Database*. MIT Press.