Using N-gram Models to Combine Query Translations in Cross-Language Question Answering

Rita M. Aceves-Pérez, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{rmaceves, mmontesg, villasen}@inaoep.mx

Abstract. This paper presents a method for cross-language question answering. The method combines multiple query translations in order to improve the answering precision. The combination of translations is based on their pertinence to the target document collection rather than on their grammatical correctness. The pertinence is measured by the translation perplexity with respect to the collection language model. Experimental evaluation on question answering demonstrates that the proposed approach outperforms the results obtained by the best translation machine.

1 Introduction

A question answering (QA) system is a particular kind of search engine that allows users to ask questions using natural language instead of an artificial query language. In a cross-lingual scenario the questions are formulated in a language different from the document collection. In this case, the efficiency of the QA system greatly depends on the way it confronts the idiomatic barrier. Traditional approaches for cross-lingual information access involve translating either the documents into the expected query language or the questions into the document language. The first approach is not always practical, in particular when the document collection is very large. The second approach is more common. However, because of the small size of questions in QA, the machine translation methods do not have enough context information, and tend to produce unsatisfactory question translations.

A bad question translation generates a cascade error through all phases of the QA process. This effect is evident in the results of cross-lingual QA reported on the last edition of CLEF [4]. For instance, the results from the best cross-lingual system (that uses the French as target language) were 64% of precision for the monolingual task, and 39.5% when using English as question language. In this case, the errors in the translation of the question cause a drop in precision of 61.7%.

Recent methods for cross-lingual information access attempt to minimize the error introduced by the translation machines. In particular, the idea of combining the capacities of several translation machines has been successfully used in cross-lingual

* This work was partially financed by the CONACYT (grants 43990 and 184663). We also like to thanks to the CLEF for the provided resources.

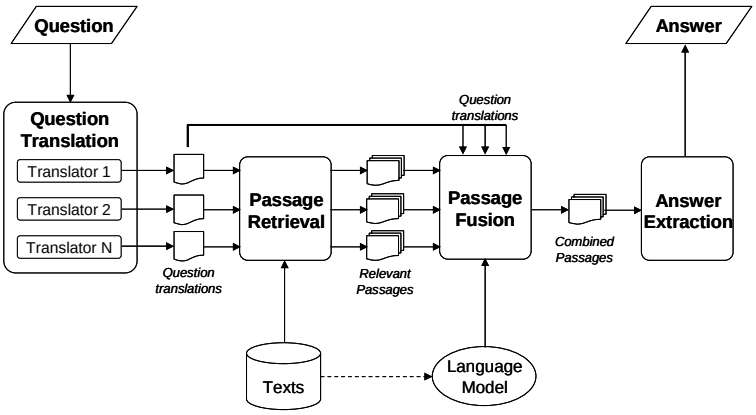

Figure 1. General scheme of the method

information retrieval [2]. In this field, most works focus on the selection of the best translation from a set of candidates [1]. In opposition, in this paper we propose a method that considers a weighted combination of the passages recovered from each translation in order to enhance the final precision of a cross-lingual QA system. In this way, all translations are treated as –possible– relevant reformulations of the original question.

2. Proposed method

The proposed method assumes that machine translation is not a solved task, and tries to face it by combining the capacities of different translators. Figure 1 shows the general scheme of the method. It considers the following procedures. First, the user question is translated to the target language by several different translators. Then, each translation is used to retrieve a set of relevant passages. After that, the retrieved passages are combined in order to form one single set of relevant passages. Finally, the selected passages are analyzed and a final question answer is extracted.

The main step of this method is the combination of the passages. This combination is based on the pertinence of the translations to the target document collection. The pertinence of a translation indicates its probability of being generated from the document collection. In other words, the pertinence of a translation expresses how it fits in the *n*-gram model calculated on the target document collection. The idea is to combine the passages favoring those retrieved by the more pertinent translations.

The following subsections describe the measuring of the pertinence of a translation to a target document collection, and the combination of the relevant passages in one single set.

2.1 Translation evaluation

As we mentioned, the pertinence of a translation to the target document collection is based on how much it fits in the collection *n*-gram model. In order to quantify this attribute we apply a general *n*-gram test on the translation. An *n*-gram test computes

the entropy (or perplexity) of some test data –the question translation– given an *n*-gram model. Basically, it is an assessment on how probable is to generate the test data from the *n*-gram model. The entropy is calculated as:

$$H = -\frac{1}{Q}\sum_{i=1}^{Q}\log P(w_i|w_{i-1}, w_{i-2},..., w_{i-N+1})$$

where w_i is a word in the *n*-gram sequence, $P(w_i)$ indicates the probability of w_i, Q is the number of words of the test data, and N is the order of the *n*-gram model.

The final score for a translation is expressed by its perplexity, defined as $B = 2^H$. In this case, a low perplexity value indicates a more predictable language, and therefore, that the translation is pertinent to the target collection.

2.2 Passage Fusion

This module combines the retrieved passages from each translation in one single set. Its purpose is to favor passages recovered by the more pertinent translations. The following formula is used to calculate the number of passages from a given translation that will be included in the combined passage set.

$$E_x = \frac{k}{\sum_{i=1}^{n}\frac{1}{B_i}\times B_x}$$

In this formula E_x indicates number of selected passages from the translator *x*, that is, the extension of *x* in the combined set. B_x is the perplexity of the translator *x*, *n* is the number of translation machines used in the experiment, and *k* indicates the number of passages retrieved by each translator as well as the total extension of the combined set. In the experiments we set $k = 20$, which corresponds to the best performance rate of our QA system [3].

3 Experiments

For the experimental evaluation of the method we considered a set of 141 factual questions extracted from the Multi-Eight Corpus of the CLEF[1]. We used the passage retrieval and answer extraction components of the TOVA question answering system [3], which was the second best in the Spanish QA task at the last edition of the CLEF.

The evaluation consisted of three bilingual experiments: English-Spanish, French-Spanish and Italian-Spanish. For the translation from English and French to Spanish we use four different translation machines[2]: Systran, Webtranslation, Reverso and Ya. For the translation from Italian to Spanish we used[3]: Systran, IdiomaX, Worldlingo, Zikitrake.

For the three experiments we measured the lost of precision in the answer extraction caused by the question translation in relation to the Spanish monolingual

[1] Cross-Language Evaluation Forum (www.clef-campaign.org).

[2] www.systranbox.com, www.imtranslator_webtranslation.paralink.com, elmundo.reverso.net, traductor.ya.com

[3] www.systranbox.com, www.idiomax.com, www.worldlingo.com, www.zikitrake.com

task. Table 1 shows the lost of precision, indicated as an error rate, for the three bilingual experiments. The first four columns indicate the error rates generated by each machine translation when they were used alone. The last column shows the error rates that were obtained when using the combined passages. In all cases, except for French, the proposed combination of the passages obtained lower error rates than the best translation machine. In addition, our method outperforms two other naïve approaches. One based on the selection of the translation with the lowest perplexity [1] (see column 5), and other one based on a uniform combination of the recovered passages (see column 6).

Table 1. Error rates in relation to the Spanish monolingual task

	MT1	MT2	MT3	MT4	Lowest perplexity	Uniform Combination	Proposed method
English-Spanish	17%	24%	17%	27%	14%	27%	***7%***
French-Spanish	***17%***	38%	27%	31%	31%	34%	27%
Italian-Spanish	52%	45%	41%	34%	41%	34%	***24%***

4 Conclusions

In this paper we presented a method for cross-lingual QA that tackles the problem of question translation by combining the capacities of different translators. The experiments demonstrated that the combination of passages retrieved by several translation machines tend to reduce the error rates introduced by the question translation process.

In the French-Spanish experiment, our method produced error rates higher than those from the best translation machine. This situation was caused by, on the one hand, the incorrect translation of several named entities from French to Spanish, and on the other hand, by the inadequate treatment of unknown words by our *n*-gram model.

As future work we plan to improve the *n*-gram test in order to handle unknown words, and to apply the method on different target languages.

6 References

1. Callison-Burch C., and Flournoy R. A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. In Proceedings of the Machine Translation Summit VIII, Santiago de Compostela, Spain, 2001.
2. Di Nunzio G. M., Nicola Ferro, Gareth J.F. Jones, Carol Peters. CLEF 2005: Ad Hoc Track Overview. CLEF 2005, Vienna, Austria, 2005.
3. Montes-y-Gómez, M., Villaseñor-Pineda, L., Pérez-Coutiño, M., Gómez-Soriano, J. M., Sanchis-Arnal, E. & Rosso, P. INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering. CLEF 2005, Vienna, Austria, 2005.
4. Vallin, A., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D. & Sutcliffe, R. Overview of the CLEF 2005 Multilingual Question Answering Track. CLEF 2005, Vienna, Austria, 2005.