Web-based Selection of Optimal Translations of Short Queries*

Paolo Rosso and **Davide Buscaldi**
DSIC, Universidad Politécnica de Valencia
Camino de Vera, s/n Valencia (Spain)
{prosso,dbuscaldi}@dsic.upv.es

Matteo Iskra
DISI, Università di Genova
Via Dodecaneso, 12 Genova (Italy)
2002s040@educ.disi.unige.it

Resumen: En este artículo se presenta una técnica para la selección de la mejor traducción de una pregunta entre un conjunto de traducciones obtenidas desde diferentes traductores automáticos. La técnica está basada en el cálculo de la entropía de la pregunta respeto a la web. La presente técnica se puede utilizar en aplicaciones multilingüe como la Búsqueda de Respuestas multilingüe.
Palabras clave: Traducción Automática, Búsqueda de Respuestas Multilingüe, Minería de Datos en la Web

Abstract: In this paper we present a technique for the selection of the best translation of a short query among a set of translation obtained from different translators. The technique is based on the calculation of the information entropy of the query with respect to the web. This technique may be used in multilingual applications such as the Cross-Lingual Question Answering.
Keywords: Machine Translation, Multilingual Question Answering, Web Mining

1 Introduction

Nowadays, it is possible to find in the web many Machine Translation (MT) tools that are commonly used to translate small pieces of text. Our assumption is that these tools can be used effectively in the Question Answering (QA) field, particularly for the Cross-Language task. If we consider QA as a specialized Information Retrieval (IR) task, the analogue of a user query in QA is a short, concise question. It has been demonstrated that the translations generated by typical web-based MT tools are more precise for short sentences than longer ones (Mellebeek et al., 2005). Therefore, the characteristics of shortness and conciseness of QA queries let us suppose that they can be translated effectively by a web MT tool, and subsequently improve the results of existing Cross-Language QA systems.

A great amount of the errors of multilingual QA systems are due to the translation phase. It has been observed that bad translations account for 15% up to 50% of the total number of errors, depending on the question type, in one of the best monolingual QA system (Laurent, Séguéla, y Nègre, 2006) that participated in the latest CLEF[1] evaluation exercise.

Various methods have been developed recently in order to minimize the error introduced by MT in IR-related fields. In particular, the idea of combining different MT systems has already been used succesfully for the cross-lingual Ad-Hoc retrieval task (Di Nunzio et al., 2005). The most common form of combination of different MT systems is the selection of the best translation from a set of candidates (Callison-Burch y Flournoy, 2001; Larosa et al., 2005), although there have been also proposals for the combination of fragments from different translations (Aceves-Pérez, Villaseñor-Pineda, y Montes, 2006).

The technique for the selection of the best translation that we present in this paper is based on the calculation of the entropy of the translations with respect to the language model in the web. It resembles a common practice among internet users, that is to exploit web search engines in order to check the spelling of a word or the correctness of a sequence of words; for instance, if someone has a doubt whether *"logic programming"* is more correct than *"logical programming"* or not, he can search the web and make a choice depending on the resulting page count. This can be done over the pieces of the translations in order to check their correctness against the "web English" language model.

In the following section we introduce the

* We would like to thank the TIN2006-15265-C06-04 research project for partially supporting this work.

[1] http://www.clef-campaign.org

ISSN: 1135-5948

adopted technique, in Section 3 we describe the experiments carried out and present the obtained results.

2 Description of the Technique

Given a translation X of a question q, let us define w as the sequence of n words that compose the translation:

$$w = (w_1, \ldots, w_n)$$

A *trigram chain* is, therefore, defined as the set of trigrams T:

$$T = \{(w_1, w_2, w_3), (w_2, w_3, w_4), \ldots \\ \ldots, (w_{n-2}, w_{n-1}, w_n)\}$$

For instance, let us consider the following Spanish translation of the English question *"Who is the Chairman of the Norwegian Nobel Committee?"*: *"Quién es el Presidente del Comité Nobel noruego?"*. Therefore, w =("Quién", "es", "el", "Presidente", "del", "Comité", "Nobel", "noruego"), and T = {("Quién es el"), ("es el Presidente"), ("el Presidente del"), ("Presidente del Comité"), ("del Comité Nobel"), ("Comité Nobel noruego")}.

The information entropy was introduced by Shannon (Shannon, 1948) and its general formulation is:

$$H(X) = -K \sum_{i=0}^{n} p(i) \log p(i) \qquad (1)$$

Where K is an arbitrary constant which depends on the problem, i is a fragment of a message X of length n, and $p(i)$ is the probability of the i-th fragment. In our case, the message is represented by the translation, and if we take into account trigrams, each fragment i corresponds to the i-th trigram of the translationd t_i.

We decided to calculate the probability of each trigrams by means of web counts. Let us name $c(x)$ the function that returns the number of pages that contain the text fragment x in the web. Let us define the i-th trigram $t_i = (w_i, w_{i+1}, w_{i+2})$ and its root bigram as $b_i = (w_i, w_{i+1})$. According to (Zhu y Rosenfeld, 2001), the probability $p(t_i)$ can be estimated as:

$$p(t_i) = \frac{c(t_i)}{c(b_i)} \qquad (2)$$

If we substitute $p(i)$ with Formula 2 in Formula 1, we obtain:

$$H(X) = -K \sum_{i=0}^{n} \frac{c(t_i)}{c(b_i)} (c(t_i) - c(b_i)) \qquad (3)$$

Due to the fact that in the web usually $c(b_i) >> c(t_i)$, we used the logarithmic scale for page counts, and used a linear normalization factor as K, obtaining the formula that we used to calculate the entropy of a translation X:

$$H(X) = -\frac{1}{n} \sum_{i=0}^{n} \frac{\log c(t_i)}{\log c(b_i)} (\log c(t_i) - \log c(b_i)) \qquad (4)$$

The selection of the best translation is made on the basis of the $H(X)$ calculated by means of Formula 4. Given M translations of a question q, we pick the translation $\bar{m}$ such that $\bar{m} = \arg\max_{m \in M} H(m)$.

For instance, consider the following translations of the example above:

1. *"Quién es el Presidente del Comité Nobel noruego?"*
2. *"Quién es el Presidente del Comité noruego Nobel?"*

The trigram counts obtained from the web (Google) are: The $H(X)$ calculated for the first translation (the right one) is 2.454 and 2.219 for the second one; therefore, the method correctly selects the best translation.

Trigram	Pages
Quién es el	271,000
es el Presidente	618,000
el Presidente del	8,560,000
Presidente del Comité	1,610,000
del Comité Nobel	468
Comité Nobel noruego	328
del Comité noruego	355
Comité noruego Nobel	73

Table 1: Web page counts for the trigrams of the two translations of the example.

3 Experiments and Results

The experiments were carried out using the set of 200 questions of the cross-lingual English-Spanish task of CLEF-2005[2]. In the

[2] http://clef-qa.itc.it/2005/

CLEF exercises, questions are the same if the target collection is the same; therefore, the right (reference) translation of each question was obtained by recurring to the monolingual Spanish question set.

3.1 MT Systems

The MT systems used for the experiments were Systran[3], FreeTrans[4], Linguatec[5], Promt[6] and Reverso[7].

The evaluation of the MT systems was carried out by means of the BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2001), a measure currently used for the evaluation of the MT systems at NIST[8]. Basically, the BLEU counts the n-grams shared by the candidate translation and the reference one. The degree of similarity returned by the BLEU is a number comprised between 0 (completely different) and 1 (perfect match). We calculated the average BLEU score for each of the MT systems on the 200 questions in the CLEF 2005 test set and over the DISEQuA corpus, consisting in 450 questions from CLEF 2003. Results are displayed in Table 2.

System	CLEF 2005	DISEQuA
Systran	0.346	0.282
Freetrans	0.371	0.333
Linguatec	0.391	0.311
Promt	0.420	0.363
Reverso	0.391	0.352

Table 2: Average BLEU scores obtained by each MT system over the 200 questions of the CLEF 2005 test set and the 450 questions of the DISEQuA corpus.

As it can be noticed from Table 2, the Promt system proved to be the more effective. Another remark that can be done is that the questions of the DISEQuA corpus seem to be more difficult to translate than the ones of the CLEF 2005.

The results grouped by question category (Table 3) show that some MT systems translate certain kinds of questions better than other ones.

Category	best BLEU	System
date	0.327	Promt
location	0.378	Promt
measure	0.317	Reverso
object	0.237	FreeTrans
organization	0.299	Reverso
person	0.513	Promt
not classified	0.308	Linguatec

Table 3: Best average BLEU scores, grouped by question category, and system that obtained the best score.

3.2 Evaluation of the Web-based Translation Selection

We used three different search engines to calculate the entropy of translations: Google[9], Yahoo[10] and Ask[11]. In order to compare the quality of the English of the Internet with the English of a static document collection, we used also Lucene[12] over the collection of documents used in the CLEF 2005 monolingual Spanish QA track.

We calculated the average entropy, obtained by means of Formula 4, for both the CLEF 2005 and DISEQuA test sets, using the above search engines to obtain the web count $c(x)$ for trigrams and bigrams. Results are shown in Table 4.

S.Engine	CLEF 2005	DISEQuA
Ask	0.381	0.325
Google	0.392	0.332
Lucene	0.378	0.313
Yahoo	0.355	0.344
Manual	0.462	n.a.

Table 4: Average BLEU score obtained with the proposed selection technique, using the different search engines for $c(x)$ over the 200 questions of the CLEF 2005 test set and the 450 questions of the DISEQuA corpus. *Manual*: average entropy obtained by selecting at hand the best translation of each question.

The "manual" heuristics can be considered as the maximum that could have been obtained if the entropy correctly helped to individuate the right translation for each ques-

[3] http://babelfish.altavista.com
[4] http://www.freetranslation.com
[5] http://www.linguatec.de
[6] http://www.e-promt.com
[7] http://www.reverso.net
[8] http://www.nist.gov
[9] http://www.google.com
[10] http://www.yahoo.com
[11] http://www.ask.com
[12] http://lucene.apache.org

tion. This is not the case, as we can observe how the manual selection obtains a 7% precision above the best result obtained with the web-based selection. Nevertheless, the manual selection does not reach the 50% of the translations, indicating that the translations of these questions is particularly problematic. We carried out an error analysis and discovered that in many cases the errors are due to the presence of Named Entities(NEs), particularly abbreviations and proper nouns. In many cases the NEs have to be translated (for instance "United Nations" is translated as "ONU" in Spanish), in other cases the translation is wrong (for instance, the Italian car manufacturer FIAT becomes "salsa de carne", "mandato" o "autorización" for some of the MT tools).

Notably, the best results obtained by means of the proposed technique are all inferior to the Promt MT system, although with the CLEF 2005 test set the web-based selection obtains a better average BLEU score than all the remaining MT systems.

4 Conclusions and Further Work

Although the best MT systems obtained better results than the web-based translation selection, some important conclusions can be drawn: the use of the web does actually prove better than a static collection, thanks to the redundancy of the data. Another result is that the selection of a search engine is important in order to obtain better results. We have observed that QA questions contain many Named Entities, and that MT tools often fail to translate properly these NEs. This can be addressed by recurring to specialized dictionary of abbreviations and proper nouns. A further work may be the combination of such a dictionary together with the selection technique improved by means of an interpolated model for probability estimation as proposed by (Zhu y Rosenfeld, 2001) for the modeling of language in the world wide web.

References

Aceves-Pérez, Rita M., Luis Villaseñor-Pineda, y Manuel Montes. 2006. Using N-gram Models to Combine Query Translations in Cross-Language Question Answering. *Lecture Notes in Computer Science, CiCLing 2006 Proceedings*, 3878:453–457.

Callison-Burch, Chris y Raymond Flournoy. 2001. A program for automatically selecting the best output from multiple translation engines. En *Proc. of the VIII Machine Translation Summit*, Santiago de Compostela, Spain.

Di Nunzio, Giorgio, Nicola Ferro, Gareth J.F. Jones, y Carol Peters. 2005. Ad hoc track overview. En *CLEF 2005 Working Notes*, Vienna, Austria.

Larosa, Sabatino, Manuel Montes y Gómez, Paolo Rosso, y Stefano Rovetta. 2005. Best Translation for an Italian-Spanish Question Answering System. En *Proc. Of Information Communication Technologies Int. Symposium (ICTIS)*, Tetuan, Morocco.

Laurent, Dominique, Patrick Séguéla, y Sophie Nègre. 2006. Cross lingual question answering using qristal for clef 2006. En *CLEF 2006 Working Notes*, Alicante, Spain.

Mellebeek, Bart, Anna Khasin, Josef Van Genabith, y Andy Way. 2005. Transbooster: Boosting the performance of wide-coverage machine translation systems. En *Proceedings of the EAMT 10th Annual Conference*, páginas 189–198, Budapest, Hungary.

Papineni, K., S. Roukos, T. Ward, y J.W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Informe técnico, IBM Research Division, Thomas J. Watson Research Center.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423.

Zhu, Xiaojin y Ronald Rosenfeld. 2001. Improving trigram language modeling with the World Wide Web. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*