Best Translation for an Italian-Spanish Question Answering System

S. Larosa[1], M. Montes-y-Gomez[2], P. Rosso[3], and S. Rovetta [1]

[1] *Dipartimento di Informatica e Scienze dell'informazione*
Università degli Studi di Genova, Italy
[2] *Laboratorio de Tecnologias de Lenguaje*
Instituto Nacional de Astrofisica, Óptica y Electrónica, Mexico
[3] *Departamento de Sistemas Informáticos y Computación*
Universidad Politécnica de Valencia, Spain

Emails: 2000s036@educ.disi.unige.it; prosso@dsic.upv.es; mmontesg@inaoep.mx; ste@disi.unige.it.

ABSTRACT. ***This paper shows the results currently achieved using four automatic translators in a Multilingual Question Answering context. The translators are used for the translation of questions (in our case from Italian to Spanish), with the purpose of selecting the best of a group and to pass it to a Question Answering system in Spanish. The choice of the translation is based on the redundancy of the words. The first results are promising considering the fact that the work was done on a small size.***

1. INTRODUCTION

Nowadays, almost all kinds of information (digital libraries, newspapers collections, etc.) in more than 1,500 languages is available on the Web, in electronic format. These documents may satisfy almost every informational need. But without suitable tools that help the user find the information such as *Information Retrieval* (IR), *Information Extraction* (IE) and recently *Question Answering* (QA), all this information would be useless. The purpose of a QA System is to provide precise answers, starting by questions formulated in natural language, instead of a collection of relevant documents. In particular, there are Multilingual QA Systems, which allow the user to get the answer by searching documents written in a language different than the one used in the query in order to exploit the redundancy of documents on the Web. In fact, searching for the same information in a different, but ample, body of documents in a common language will increase the possibilities of getting an exact answer. (Brill, 2001) (Solorio, 2004). An important and key step for a Multilingual QA system is the translation of a question from a language source to a destination one. At the moment, majority of QA systems use online translators. However, the quality of their translations is often not very good and this has a negative impact on the QA system efficiency. In this paper we focus on the problem related to the selection of the best translation if more than a translator is used. The two methods we propose are totally statistical and, therefore, they are language-independent. Particularly, we will concentrate on the translation from Italian to Spanish, because the documents written in the latter language present on the Web are greater in comparison to those written in Italian. (Approximately 2,658,631,000 Web pages in Spanish vs. 1,845,026,000 in Italian) (Kilgarriff, 2003). Finally, it has to be said that the task is difficult due to the presence of polysemic words inside the text penalizes the translation and makes the election difficult.

2. WORD-COUNT METHOD

2.1. The Scheme

With this method, which exploits the redundancy of terms in all the translations, the translation with the highest number of words in common (in other words the most similar) will be chosen. The more frequent a term, the more chances that the original word has been translated correctly. To establish the number of common words and calculate the similarity among the translations, two formulas have been chosen: the *Dice* and the *Cosine* formulae.

2.2. Word-count Implemented with the Dice formula

In order to find the number of common words, the intersection of the Spanish translations is taken into account.

Example of translated question with four different translators:

"Che cosa significa la sigla CEE?"
("What does the abbreviation EEC mean?")

1. ¿Qué significa la sigla CEE?
2. ¿Qué cosa significa siglas el EEC?
3. ¿Qué significa la CEE de la abreviación?
4. ¿Qué cosa significa la pone la sigla CEE?

Table 1. Intersection results for the previous example. "No" means that the intersection of the translation with itself is not considered.

	1 Tran.	*2 Tran*	*3 Tran.*	*4 Tran.*
1 Tran.	No	2	4	5
2 Tran.	2	No	2	3
3 Tran.	4	2	No	5
4 Tran.	5	3	5	No

The Dice formula is used to establish the degree of similarity among the translations and to create a hierarchy exploiting the information that they have in common, that is, the words.

$$Sim_{(t_i,t_j)} = \frac{2 * len(t_i \cap t_j)}{len(t_i) + len(t_j)} \qquad (1)$$

- t_i and t_j are the translations that we consider;
- $len(t_i \cap t_j)$ represents the intersection (number of words in common);
- $len(t_i)$ and $len(t_j)$ represent the number of words for every translation.

To get a corresponding similarity grade for every translation, the similarity between the translation reference and the others has to be calculated using the previous formula (1). The partial results will be added together to reach the desired similarity degree. For instance, to get the similarity grade of the first translation we do: ($Sim_{t_1,t_2} + Sim_{t_1,t_3} + Sim_{t_1,t_4}$). The translation with the highest value is chosen. To increase the accuracy in the choice of the best translation, N-Grams are used. The use of these N-Grams has, in reality, been used up to 3-Grams. If, for instance there are two translations which have the same identical words but with a different order, thanks to the N-Grams the degree of equality can be distinguished. The N-Grams are very useful in cases in which translations are formed by same identical words but in different order. In fact, thanks to them we can improve the precision of the similarity grade calculation.

- 1-Grams: the number of words in common is calculated and the *Dice* formula is applied (as described above);

- 2-Grams: the adjoining words of the translations are gathered into groups of two; thereafter, a new intersection among 2-Grams is made and the values obtained with those of the intersection of 1-Grams are added; finally, the Dice formula is applied on the obtained new data;

- 3-Grams: the adjoining words of the translations are gathered into groups of three; thereafter, a new intersection among 3-Grams is made and the values obtained with those of the 1-Grams and 2-Grams intersections are added; finally, the Dice formula is applied on the obtained the previous new data.

Example of 2-Grams of the phrase:

"Qué significa la sigla CEE?"
("What does the abbreviation EEC mean?")

"Qué significa" "significa la" "la sigla" "sigla CEE"

2.3. Word-count Implemented with the Cosine formula

We implement the word-count method previously described, using the cosine formula to calculate the similarity degree. In this model the translations are represented as vectors in a t dimensional space (t is the general number of index terms or keywords). To calculate the keyword weights the scheme TermFrequency-InverseDocumentFrequency (tf-idf) is used. All words that are in the translation are considered keywords (they are taken into consideration only once and without repetition).

Example of translated question with four different translators:

"Qual è la capitale della Repubblica del Sud Africa?"
("What is the capital of the Republic of South Africa?")

1. ¿Cuál es la capital de la República de la Sur África?
2. ¿Cuál es entendido ellos de la república de la África del sur?
3. ¿Cuál es la capital de la República del Sur una Africa?

4. ¿Cuál es el capital de la república del sur Africa?

We get the following list of keywords:

"cuál", "es", "la", "capital", "de", "república", "sur", "África", "entendido", "ellos", "del", "una", "africa", "el"

After that, we determine freq(i,j), that is, the frequency of every keyword (k_i) for every translation.
To calculate the weights for every translation we use the following formula:

$$f(i,j) * \log(1 + \frac{n_i}{N}) \qquad (2)$$

where:

- N = the total number of translations
- n_i = the number of documents that contain k_i
- $f(i,j) = freq(i,j) / max(freq(i,j))$
  It represents the frequency of the keywords in the translation, normalized w.r.t the maximum, calculated on all the keywords of that translation. The division is made to normalize the result.

The formula differs from the original one of Salton (Baeza, 1999) for the presence of the term 1 in the log. This is because if there are cases in which $N = n_i$ the log would not be equal to 0. We would note that the original formula is used with big collections of documents and the probabilities to have a term in all documents it is nearly null. But, in our case, we have a small group of translation and a term is frequently present in all documents. Another particularity is a n_i/N instead of N/n_i. We made this change to give more weight to the words that are in all translations.
At the end of this step, the vector containing the association weights to every keyword is obtained.

Example of weight keywords vectors:

t1: [1.33, 1.33, 4, 0.62, 2.6, 1.33, 1.33, 0.35, ..., 0]
t2: [2, 2, 4, 0, 4, 2, 2, 0.5, 0.3, 0.3, 0.93, 0, 0, 0]
etc...

Once the vectors have been found, the next step is the calculation of the similarity degree among translations by using the following formula:

$$Sim(t_j, t_q) = \frac{(\sum_i t_{ji} * t_{qi})}{\sqrt{\sum_i t_{ji}^2} * \sqrt{\sum_i t_{qi}^2}} \qquad (3)$$

In the formula t_{ji} and t_{qi} represent two generic vector weights. The final calculation is performed in this way:

Tran1 = $Sim(t_1,t_2) + Sim(t_1,t_3) + Sim(t_1,t_4)$
Tran2 = $Sim(t_2,t_1) + Sim(t_2,t_3) + Sim(t_2,t_4)$
Tran3 = $Sim(t_3,t_1) + Sim(t_3,t_2) + Sim(t_3,t_4)$
Tran4 = $Sim(t_4,t_1) + Sim(t_4,t_2) + Sim(t_4,t_3)$

The translation with the highest value is chosen.

3. DOUBLE TRANSLATION METHOD

3.1. The Scheme

Every question in Italian is translated into Spanish then retranslated back into Italian. Four translators are used and the translation whose results are more similar to the original question will be chosen. The *Dice* and the *Cosine* formulas are used in this case as well. The algorithms used are those previously illustrated in Section 2.2 and Section 2.3 with some little changes.

Example of original question and double translation:

"Che cosa significa la sigla CEE?"
("What does the abbreviation EEC mean?")

1. che cosa significa la sigla CEE?
2. Che cosa significa le abbreviazioni il EEC?
3. Che significa il CEE dell'abbreviazione?
4. che cosa ha importanza la mette la sigla di CEE?

3.2. Double Translation Implemented with the Dice formula

The algorithm for this method differs from the previous illustrated in Section 2.2 for the intersection between translations. In fact in this method we make an intersection between the original question and the retranslated questions. Then the formula (1) is used and the similarity grade is obtained. The translation with the highest similarity degree is chosen. Also with this method N-Grams, up to trigrams, are used.

3.3. Double Translation Implemented with the Cosine formula

Also in this algorithm we have some difference with respect to the first method. In this case we make a list of keywords including the original question. Then the steps to create a list of keywords are the same illustrated in Section 2.3. For the retranslated questions the formula (2) is used; but for the original question we use this formula:

$$(0.5 + [0.5 * f(i,j)]) * \log(1 + \frac{n_i}{N}) \quad (4)$$

The formula is used because the original question is compared to an Information Retrieval query. This formula differs from the original one of Salton & Bucley (Baeza, 1999) for the same reasons described in the previous section (2.3). After that the keywords vectors are obtained and the formula (3) applied on the original question weights keywords vector with all the others. The translation with the highest similarity degree is chosen. Also with this method n-grams, up to trigrams, are used.

4. EXPERIMENTAL RESULTS

4.1. Obtained Results

The (Cross-Language Evaluation Forum[1] (CLEF) is an European consortium that organizes an international competition which regards information retrieval systems and works on European languages in monolingual and multilingual environments. In our case we translate 450 factual questions (ie.g. who, when, where, what) written in Italian which have been derived from the CLEF 2003 competition. These questions are translated with 4 different translators and, for every translated question, 4 Spanish translations (word-count) or 4 Italian retranslations are obtained (double translation). The four translators used are: PowerTranslationPro[2], Idiomax[3], Google[4], FreeTranslation[5]. The first are software applications whereas the latter ones are available online. The online translators do not allow a direct translation from Italian to Spanish. Therefore, we needed to go through an intermediary translation into English. The following tables make a comparison of the obtained results using the different translators, applying the techniques previously explained.

Table 2. Word-count method with the *Dice* formula.

1-Gram	*2-Grams*	*3-Grams*
51,33 %	51,11 %	51,55 %
231 / 450	230 / 450	232 / 450

Table 3. Double Translation method with the *Dice* formula.

1-Gram	*2-Grams*	*3-Grams*
46,66 %	49,11 %	50,22 %
210 / 450	221 / 450	226 / 450

Table 4. Word-count method with the *Cosine* formula.

1-Gram	*2-Grams*	*3-Grams*
48,66 %	49,33%	50,00%
219 / 450	222 / 450	225 / 450

Table 5. Double Translation method with the *Cosine* formula.

1-Gram	*2-Grams*	*3-Grams*
45,77 %	48,44%	49,11%
206 / 450	218 / 450	221 / 450

Each of the previous tables shows the percentage of success and the number of questions which were properly translated in every experiment.

4.2. Results Discussion

From these experiments we have observed that some translators make a very bad translation. In particular we refer to Google and Freetranslation, the online translators. This is probably due to the fact that an intermediary translation in English is needed for obtaining a final Spanish translation. As a consequence, there are some cases where the bad translation and consequently the bad redundancy penalizes the election of the best translation especially in the double translation method.

The machine translator which obtained the best results is PowerTranslationPro (55.33%). This baseline was better than our best results (51.55%) which were obtained with the word-count method. Nevertheless, the preliminary results we obtained seem to be promising. In fact, an optimal combination among the *Word-count and Double Translation* methods could increase the percentage of success. We estimate that it should be possible to obtain approximately an increase of up to 20% of the system's performance. This is due to the fact that the choices obtained from two methods are not the same.

[1] www.clef-campaign.org
[2] Power Translator Pro: www.lec.com
[3] Idiomax: www.idiomax.com
[4] Google: www.google.it/language_tools
[5] FreeTranslation: www.freetranslation.com

5. CONCLUSIONS AND FURTHER WORKS

In this paper we have proposed a prototype of translation for a Multilingual QA System. In general the results of this experiment seem to be promising. Further experiments are needed to find an optimal combination of the two methods we investigated. In the future, the use of other translators is foreseen, in order to improve the quality of translations. Last, we need to make some further experiments with other sets of factual questions to make a comparison with the preliminary results we obtained.

6. ACKNOWLEDGEMENTS

The work was partially supported by the R2D2 (CICYTTIC2003-07158-C04-03), ICT EU-India (ALA/95/23/2003/077-054) research projects and Conacyt (J43990-Y).

7. REFERENCES

Baeza-Yates, R., and Ribeiro-Neto, B., 1999.,Modern Information Retrieval. Addison-Wesley.

Brill, E., Lin, J., Banko, M., Dumais, S., Ng, A., 2001, Data-intensive Question Answering. Proceedings of the Tenth Text REtrieval Conference, TREC-2001.

Kilgarriff, A., and Greffenstette, G., 2003, Introduction to the Special Issue on Web as Corpus, *Computational Linguistics,* 29(3), pp.1-15.

Lin., D., 1998, An information-theoretic definition of similarity. Proceedings 15th International Conf. on Machine Learning.

Solorio, T., Pérez, M., Montes, M., Villaseñor, L., and López, A., 2004, A Language Independent Method for Question Classification. In: Proc. of the 20th Int. Conf. on Computational Linguistics (COLING-04). Geneva, Switzerland.