The Negative Effect of Machine Translation on Cross-Lingual Question Answering^{*}

Sergio Ferrández and Antonio Ferrández

Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
{sferrandez,antonio}@dlsi.ua.es

Abstract. This paper presents a study of the negative effect of Machine Translation (MT) on the precision of Cross-Lingual Question Answering (CL-QA). For this research, a English-Spanish Question Answering (QA) system is used. Also, the sets of 200 official questions from CLEF 2004 and 2006 are used. The CL experimental evaluation using MT reveals that the precision of the system drops around 30% with regard to the monolingual Spanish task. Our main contribution consists on a taxonomy of the identified errors caused by using MT and how the errors can be overcome by using our proposals. An experimental evaluation proves that our approach performs better than MT tools, at the same time contributing to this CL-QA system being ranked first at English-Spanish QA CLEF 2006.

1 Introduction

At present, the volume of on-line text in natural language in different languages that can be accessed by the users is growing continuously. This fact implies the need for a great number of tools of Information Retrieval (IR) that permit us to carry out multilingual information searches.

Multilingual tasks such as IR and Question Answering (QA) have been recognized as an important issue in the on-line information access, as it was revealed in the the Cross-Language Evaluation Forum (CLEF) 2006 [6].

IR is the science that studies the search for information in documents written in natural language. QA is a more difficult task than IR task. The main aim of a QA system is to localize the correct answer to a question written in natural language in a non-structured collection of documents.

In the Cross-Lingual (CL) environments, the question is formulated in a different language from the one of the documents, which increases the difficulty. The multilingual QA tasks were introduced in the CLEF 2003 [7] for the first time. Since them, most of the CL-QA system uses MT systems to translate the

^{*} This research has been partially funded by the Spanish Government under project CICYT number TIC2003-07158-C04-01 and by the European Commission under FP6 project QALL-ME number 033860.

queries into the language of the documents. But, this technique implies a drop around 30% in the precision with regard to the monolingual task.

In this paper, a study of the negative effect of Machine Translation (MT) on the precision of Cross-Lingual Question Answering (CL-QA) is presented, designing a taxonomy of the identified errors caused by MT. Besides, a proposal to overcome these errors is presented which reduces the negative effects of the question translation on the overall accuracy of CL-QA systems.

The rest of the paper is organized as follows: section 2 describes the state of CL-QA systems. Afterwards, an empirical study of the errors of MT is described. And the analysis of the influence of MT errors on CL-QA is shown in section 3. Section 4 presents our approach for CL-QA that minimizes the use of MT. Besides, section 5 presents and discusses the results obtained using all official English questions of QA CLEF 2004 [8] and 2006 [6]. Finally, section 6 details our conclusions and future work.

2 State of the Art

Nowadays, most of the implementations of current CL-QA systems are based on the use of on-line translation services. This fact has been confirmed in the last edition of CLEF 2006 [6].

The precision of CL-QA systems is directly affected by its ability to correctly analyze and translate the question that is received as input. An imperfect or fuzzy translation of the question causes a negative impact on the overall accuracy of the systems. As Moldovan [9] stated, Question Analysis phase is responsible for 36.4% of the total of number of errors in open-domain QA.

Next, we are focusing on the bilingual English-Spanish QA task, because the CL-QA system used for the evaluation works in these languages. Nowadays, at CLEF 2006 [6], three different approaches are used by CL-QA systems in order to solve the bilingual task. The first one [10] uses an automatic MT tool to translate the question into the language in which the documents are written. This strategy is the simplest technique available. In this case, when compared to the Spanish monolingual task, the system loses about 55% of this precision in the CL task.

On the other hand, the system [1] translates entire documents into the language in which the question is formulated. This system uses a statistical MT system that has been trained using the European Parliament Proceedings Parallel Corpus 1996–2003 (EUROPARL).

Finally, the system BRUJA [5] translated the question using different on-line machine translators and some heuristics. This technique consults several web services in order to obtain an acceptable translation.

The previously described strategies are based on the use of MT in order to carry out the bilingual English-Spanish task, and all of them try to correct the translation errors through different heuristics. The low quality of MT provides a load of errors inside all the steps of the localization of the answer. These facts cause an important negative impact on the precision of the systems. And it can

be checked on the last edition of CLEF 2006 where the cross lingual system obtains less than 50% of correct answer compared to the monolingual task.

In the next section, a taxonomy of the identified errors caused by MT in CL-QA is shown, and how the errors are overcome using our proposals.

3 Taxonomy of the MT Errors for CL-QA

In this section, our classification of different errors caused by the use of MT is described. The taxonomy is designed using the CLEF 2004 [8] set of 200 English questions, this year the bilingual English-Spanish task was introduced for the first time.

The set of 200 English questions are translated into Spanish using on-line MT services¹. The errors noticed during the translation process are the following: wrong word-by-word translation, wrong translated sense, wrong syntactic structure, wrong interrogative particle, wrong lexical-syntactic category, unknown words and wrong proper name.

Table 1 shows the seven different types of error that compose our taxonomy, as well as the percentages of appearance in the English questions of CLEF 2004. Next, each type is described in detail as well as the problems that the wrong translations cause in the CL-QA process.

Table 1. Types of translation errors and percentage of appearance on CLEF 2004 set of question

Type of translation error	Percentage
Wrong Word-by-Word Translation	24%
Wrong Translated Sense	21%
Wrong Syntactic Structure	34%
Wrong Interrogative Particle	26%
Wrong Lexical-Syntactic Category	6%
Unknown Words	2.5%
Wrong Proper Name	12.5%

3.1 Wrong Word-by-Word Translation

This kind of error causes a lot of problems during the search of the correct answers in the CL-QA process, since it inserts words in the translation that should not be inserted.

In this type, the MT replaces words in the source language with their equivalent translation into Spanish, when there is not one-to-one correspondence between English language and Spanish language. Table 2 shows an example of this type of wrong translation in the question 002 at CLEF 2004.

¹ MT1: <http://www.freetranslation.com>, MT2: <http://www.systransoft.com> and MT3: <http://babelfish.altavista.com>

Table 2. Wrong Word-by-Word Translation

English Question	How much does the world population increase each year?
Right Spanish Question	¿Cuánto aumenta la población mundial cada año?
Translation into Spanish	¿Cuánto hace el aumento de población de mundo cada año?

In the previous example, the MT system inserts the verb “*hace*” which is not useful to the CL-QA process, this fact introduces a negative effect that does not permit the QA system to find out the answer.

The MT services produces this error in a 24% of the questions.

3.2 Wrong Translated Sense

Some wrong translations are produced when a single word has different senses according to the context in which the word is written. The MT service translates the 21% of the question with at least one wrong word sense. Table 3 shows an example, the question 065 at CLEF 2004 where the word “*sport*” is translated erroneously into the Spanish word “*luce*” (to show off).

Table 3. Wrong Translated sense

English Question	What sports building was inaugurated in Buenos Aires in December of 1993?
Right Spanish Question	¿Qué edificio deportivo fue inaugurado en Buenos Aires en diciembre de 1993?
Translation into Spanish	¿Qué luce edificio se inauguró en Buenos Aires en diciembre de 1993?

Sometimes these errors are able to modify completely the sense of the question and cause a great negative effect in the precision of the CL-QA system.

3.3 Wrong Syntactic Structure

In this case, the wrong translation produces changes in the syntactic structure of the question. This type of our taxonomy causes a lot of errors in the phase of question analysis within the QA process, since this phase is usually based on syntactic analysis. Table 4 shows an example, the question 048 at CLEF 2004 where the structure of the question has been strongly modified.

This kind of translation error is the most common error encountered during translation (34%).

When the CL-QA system is fundamentally based on syntactic analysis of the question and the documents, the changes in the syntactic structure of the

Table 4. Wrong Syntactic Structure

English Question	How many people died in Holland and in Germany during the 1995 floods?
Right Spanish Question	¿Cuántas personas murieron en las inundaciones de Holanda y Alemania en 1995?
Translation into Spanish	¿ Murieron cuántas personas en Holanda y en Alemania durante las 1995 inundaciones ?

question influence negatively producing errors that do not permit the system to localice the answers.

In the previous example, a syntactic analysis of the wrong translated question returns an erroneous noun phrase, “[1995 inundaciones]”, in which the year “1995” is tagged as a determinant. The right translation has to obtain two independent noun phrases : “[1995]” and “[inundaciones]” (floods).

3.4 Wrong Interrogative Particle

A QA system develops two main tasks in the phase of question analysis: 1) the detection of the expected answer type; and 2) the identification of the main syntactic blocks (SB) of the question.

Table 5. Wrong Particle Interrogative

English Question	What is the official German airline called?
Right Spanish Question	¿ Cómo se llama la aerolínea oficial alemana?
Translation into Spanish	¿ Qué se llama la linea aérea alemana oficial?

The wrong translation of the particle interrogative of the question causes a wrong detection of the expected answer type. This fact does not allow the QA system to carry out a correct run. The MT tool carries out this type of error in the 26% of the questions. In all these question, the detection of the expected answer type is, in most cases, erroneous.

In table 5, the question 025 at CLEF 2004 is shown, where the MT services makes a mistake in the particle interrogative.

3.5 Wrong Lexical-Syntactic Category

This kind of problem causes wrong translations, such as nouns that are translated into verbs. In this type of situations, the extraction of the correct answer is impossible to carry out. Table 6 describes an example, the question 092 at CLEF 2004 where the noun “war” is translated into the verb “Guerreó” (to fight).

Table 6. Wrong lexical-syntactic category

English Question	When did the Chaco War occur?
Right Spanish Question	¿Cuándo fue la Guerra del Chaco?
Translation into Spanish	¿Cuándo Guerreó el Chaco ocurre?

3.6 Unknown Words

In these cases, the MT service does not know the translation of some words. These words, because they are not translated, are not useful for CL-QA purposes. As it is shown in table 7, in the question 059 at CLEF 2004 the word “*odourless*” is unknown by the MT service.

Table 7. Unknown Words

English Question	Name an odourless and tasteless liquid.
Right Spanish Question	Cite un líquido inodoro e insípido.
Translation into Spanish	Denomine un odourless y líquido insípido.

3.7 Wrong Proper Name

This kind of wrong translations is the a typical error that encountered during translation using the MT service.

These problems do not allow the QA system to be able to find out the correct solution. For example, in the question 112 at CLEF 2004 (see table 8), the proper name “*Bill*” is translated into the common noun “*Cuenta*” (bill). This fact does not permit to know who is *Bill Clinton*.

Table 8. Wrong Proper Name

English Question	Who is Bill Clinton?
Right Spanish Question	¿Quién es Bill Clinton?
Translation into Spanish	¿Quién es Cuenta Clinton?

4 Our Approach to CL-QA

In this section, our approach to open domain CL-QA system called BRILI [2] is detailed. BRILI (Spanish acronym for “Question Answering using Inter Lingual

Index module”) introduces two improvements that alleviate the negative effect produced by MT:

- Unlike the current bilingual English–Spanish QA systems, the question analysis is developed in the original language without any translation. The system develops two main tasks in the phase of question analysis:
 - 1) the detection of the expected answer type, the system detects the type of information that the answer has to satisfy to be a candidate of an answer (proper name, quantity, date, ...).
 - 2) the identification of the main SB of the question. The system extracts the SB that are necessary to find the answers.
- The system considers more than only one translation per word by means of using the different synsets of each word in the Inter Lingual Index (ILI) Module of EuroWordNet (EWN).

Next, using the previous examples of wrong translations, the application of our approach to overcome the problems is described.

4.1 Solution to Wrong Word–by–Word Translation

In this problem, the MT service inserts words in the translation that should not be inserted. Our method resolves this mistake by making an analysis of the question in the original language. Afterwards, the system choose the SB that must be translated and that are useful to the extraction of the answer. In our case, these SB are referenced using the ILI.

- **English Question:** *How much **does** the world population increase each year?*
- **Wrong Translation:** *¿Cuánto **hace** el aumento de población de mundo cada año?*
- **SBs:**

[world population]
[to increase]
[each year]

- **Keywords to be referenced using ILI:** world population increase each year

In the previous example, the words “*How much does*” are discarded to the extraction of the answer phase in the QA process.

4.2 Solution to Wrong Translated Sense

This kind of error is produced when a single word has different senses according to the context in which the word is written. Our approach solves this handicap considering more than only one translation per word by means of using the different synsets of each word in the ILI module of EWN.

- **English Question:** *What **sports** building was inaugurated in Buenos Aires in December of 1993?*
- **Wrong Translation:** *¿Qué **luce** edificio se inauguró en Buenos Aires en diciembre de 1993?*
- **References of the word “sport” using ILI:** *deporte deporte coña socarronería mucación mutante deportista*

In the previous example, the method finds more than one Spanish equivalents for the word “*sport*”. Each ILI synsets is appropriately weighted by Word Sense Disambiguation (WSD) mechanisms [4]. In this case, the most valued Spanish word would be “*deporte*”.

4.3 Solution to Wrong Syntactic Structure

In this case, the MT tool produces changes in the syntactic structure of the question that cause errors within the question analysis phase. This mistake is solved by our method by making an analysis of the question in the original language without any translation. This behavior is shown in the next example:

- **English Question:** *How many people died in Holland and in Germany during the 1995 floods?*
- **Wrong Translation:** *¿Murieron cuántas personas en Holanda y en Alemania durante las 1995 **inundaciones**?*
- **SBs:**

[people]
[to die]
[in Holland and in Germany during the 1995 floods]

- **Keywords to be referenced using ILI:** people die Holland Germany floods

4.4 Solution to Wrong Interrogative Particle

The wrong translation of the particle interrogative of the question causes a wrong detection of the expected answer type. Our approach resolves this problem applying the syntactic patterns that determine the expected answer type to the question in the original language without any translation.

- **English Question:** *What is the official German airline called?*
- **Wrong Translation:** *¿Qué se llama la línea aérea alemana oficial?*
- **English Pattern:**

[WHAT] [TO BE] [AIRLINE]

- **Expected answer type:** group

4.5 Solution to Wrong Lexical-Syntactic Category

This kind of error causes wrong translations when, for example nouns are translated into verbs. Our method uses ILI to reference nouns and verbs independently, this strategy is shown in the next example.

- **English Question:** *When did the Chaco War occur?*
- **Wrong Translation:** *¿Cuándo **Guerreó** el Chaco ocurre?*
- **References of the noun “war” using ILI:** *guerra*
- **References of the verb “to occur” using ILI:** *acaecer tener-lugar suceder pasar ocurrir*

4.6 Solution to Unknown Words

In these cases, the MT service does not know the translation of some words. Our technique minimizes this handicap by using the ILI module.

On the other hand, the words that are not in EWN are translated using an on-line Spanish Dictionary². Furthermore, our method uses gazetteers of organizations and places in order to translate words that are not linked using the ILI module.

In the next question, it is shown an example of this solution:

- **English Question:** *Name an **odourless** and tasteless liquid.*
- **Wrong Translation:** *Denomine un **odourless** y líquido insípido.*
- **References of the word “odourless” using ILI:** -
- **Translated word using an on-line Spanish Dictionary:** *inodoro*

4.7 Solution to Wrong Proper Name

These wrong translations do not allow the QA system to be able to find out correct solutions. Our method, in order to decrease the effect of incorrect translation of the proper names, the achieved matches using these words in the search of the answer are realized using the translated word and the original word of the question. The found matches using the original English word are valued a 20% less.

This strategy is shown in the next example:

- **English Question:** *Who is Bill Clinton?*
- **Wrong Translation:** *¿Quién es **Cuenta** Clinton?*
- **References of the words “Bill Clinton” using ILI:** *cartel Clinton*
- **SBs used to the extraction of the answer:**

[cartel Clinton]
[Bill Clinton]

5 Evaluation

In this section, the experiments that prove the improvement of our method are shown. The evaluation has been carried out using a CL-QA system that is based on our approach and it has been compared with the QA system using the MT³ service.

² <http://www.wordreference.com>

³ <http://www.freetranslation.com/>

The main aim of this section is to value our CL-QA strategy. In order to make this, the CLEF 2004 and 2006 sets of 200 English question and the EFE 1994–1995 Spanish corpora are used.

Table 9 shows the achieved experiments, where the column 4 details the improvement in relation to the using of MT and in the case of the others participants at CLEF 2006, the column 4 shows the decrement in relation to our method. Besides, the obtained precision⁴ for each dataset is shown in the column 3.

Also, in table 9, we show the precision of the monolingual Spanish task of the used QA system (see rows 1 and 4) and the precision of the currents participants at CLEF 2006 (see rows 7, 8 and 9).

Table 9. Evaluation

	Approach	Dataset	Precision (%)	Improvement (%)
		CLEF 2004		
1	Spanish QA system	200 Spanish questions	38.5	+28.44
2	Our method + QA system	200 English questions	34	+19.12
3	MT + QA system	200 English questions	27.5	–
		CLEF 2006		
4	Spanish QA system	200 Spanish questions	36	+52.7
5	Our method + QA system	200 English questions	20.50	+17.07
6	MT + QA system	200 English questions	17	–
	Others participants at	CLEF 2006[6]		Decrement (%)
7	QA system [10]	200 English questions	6	–70.73
8	QA system [1]	200 English questions	19	–7.31
9	QA system [5]	200 English questions	19.5	–4.87

The experimental evaluation shows up the negative effect of the MT services on CL-QA. In the tests using the MT tool, the errors produced by the question translation (see rows 3 and 6) generate worse results than using our method (see rows 2 and 5: +19.12% at CLEF 2004 and +17.07% at CLEF 2006). Besides, our approach obtains better results than other participants at CLEF 2006 (see the decrement in relation to our method in rows 7, 8 and 9).

These experiments prove that our approach obtains better results than using MT (where the lost of precision in the CL task is around 29% at CLEF 2004 and 50% at CLEF 2006) and other current bilingual English–Spanish QA systems. Furthermore, this affirmation is corroborated checking the official results on the last edition of CLEF 2006 [6] where our method [3] has being ranked first at the bilingual English–Spanish QA task.

⁴ Correct answers return on the first place.

6 Conclusion and Future Work

This paper presents a taxonomy of the seven identified errors caused using MT services and how the errors can be overcome using our proposals in order to solve QA task in cross lingual environments.

Our method carries out two tasks reducing the negative effect that is inserted by the MT services. Our approach to CL-QA tasks carries out the question analysis in the original language of the question without any translation. Besides, more than one translation per word is considered by means of using the different synsets of each word in the ILI module of EuroWordNet.

The tests on the official CLEF set of English questions prove that our approach generates better results than using MT (+19.12% at CLEF 2004 and +17.07% at CLEF 2006) and than other current bilingual QA systems [6].

Further work will study the possibility to take into account a Name Entity Recognition to detect proper names that will not be translated in the question. For instance, using the question 059 at CLEF 2006, *What is Deep Blue?*, the words “Deep Blue” should not be translated.

Furthermore, the gazetteers of organizations and places will be extended using multilingual knowledge of extracted from Wikipedia⁵ that is a Web-based free-content multilingual encyclopedia project.

References

1. M. Bowden, M. Olteanu, P. Suriyentrakorn, J. Clark, and D. Moldovan. LCC’s PowerAnswer at QA@CLEF 2006. In *Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2006.
2. S. Ferrández, A. Ferrández, S. Roger, P. López-Moreno, and J. Peral. BRILI, an English-Spanish Question Answering System. *Proceedings of the International Multiconference on Computer Science and Information Technology. ISSN 1896-7094*, 1:23–29, November 2006.
3. S. Ferrández, P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alavarado, E. Noguera, and F. Llopis. AliQAn and BRILI QA System at CLEF-2006. In *Workshop of Cross-Language Evaluation Forum (CLEF)*, 2006.
4. S. Ferrández, S. Roger, A. Ferrández, A. Aguilar, and P. López-Moreno. A new proposal of Word Sense Disambiguation for nouns on a Question Answering System. *Advances in Natural Language Processing. Research in Computing Science. ISSN: 1665-9899*, 18:83–92, February 2006.
5. M.A. García-Cumbreres, L.A. Ureña-López, F. Martínez-Santiago, and J.M. Perea-Ortega. BRUJA System. The University of Jaén at the Spanish task of CLEFQA 2006. In *Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2006.
6. B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osevana, A. Peñas, , P. Rocha, B. Sacaleanu, and R. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answering Track. In *Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2006.

⁵ <http://www.wikipedia.org/>

7. B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. Rijke. The Multiple Language Question Answering Track at CLEF 2003. *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Lecture Notes in Computer Science*, 2003.
8. B. Magnini, A. Vallin, C. Ayache and G. Erbach, A. Peñas, M. Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 Multilingual Question Answering Track. *Comparative Evaluation of Multilingual Information Access Systems: 5th Workshop of the Cross-Language Evaluation Forum, CLEF, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers. Lecture Notes in Computer Science*, 2004.
9. D.I. Moldovan, M. Pasca, S.M. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.*, 21:133–154, 2003.
10. E.W.D. Whittaker, J.R. Novak, P. Chatain, P.R. Dixon, M.H. Heie, and S. Furui. CLEF2005 Question Answering Experiments at Tokyo Institute of Technology. In *Workshop of Cross-Language Evaluation Forum (CLEF)*, September 2006.