BRUJA: Question Classification for Spanish. Using Machine Translation and an English Classifier.

Miguel Á. García Cumbreras
SINAI Group
Computer Sciences
University of Jaén. Spain
magc@ujaen.es

L. Alfonso Ureña López
SINAI Group
Computer Sciences
University of Jaén. Spain
laurena@ujaen.es

Fernando Martínez Santiago
SINAI Group
Computer Sciences
University of Jaén .Spain
dofer@ujaen.es

Abstract

Question Classification is an important task in Question Answering Systems. This paper presents a Spanish Question Classifier based on machine learning, automatic online translators and different language features. Our system works with English collections and bilingual questions (English/Spanish). We have tested two Spanish-English online translators to identify the lost of precision. We have made experiments using lexical, syntactic and semantic features to test which ones made a better performance. The obtained results show that our system makes good classifications, over a 80% in terms of accuracy using the original English questions and over a 65% using Spanish questions and machine translation systems. Our conclusion about the features is that a lexical, syntactic and semantic features combination obtains the best result.

1 Introduction

A Question Answering (QA) system seeks and shows the user an accurate and concise answer, given a free-form question, and using a large text data collection.

The use of Cross Language Information Retrieval Systems (CLIR) is growing, and also the application of these ones into other general systems, such as Question Answering or Question Classification.

A CLIR system is an Information Retrieval System that works with collections in several languages, and extract relevant documents or passages (Grefenstette, 1998).

We have proposed a Multilingual Question Answering System (BRUJA - in Spanish "Busqueda de Respuestas University of Jaen") that works with collections in several languages. Since there are several languages, tasks such as obtaining relevant documents and extracting the answer could be accomplished in two ways: using NPL tools and resources for each language or for a pivot language only (English) and translating to the pivot language the rest of the relevant information when it is required. Because of the translation step, the second approach is less accurate but more practical since we need only NPL resources for English. The central question is the noise, because of the translation process, is too high in order to use this approach in spite of their practical advantages.

The first step of this system is a Question Classifier (QC). Given a query, a question classification module obtains the class of such question. This information is useful for the extraction of the answer. For example, given the query "Where is Madrid?, the QA system expects a location entity as answer type. The proposed QA module works with questions in several languages, translates them into English using different online translators, and obtains the type of questions and some features, such as the focus, the keywords or the context. In this work we aim to find out whether a multilingual QC module is possible by using translation tools and English as pivot language or not.

2 Question Classification

Question Classification is the task that, given a question, classifies it in one of k semantic classes.

Some QC systems are based on regular expressions and manual grammatical rules (Van Durme et al., 2003).

Recent works in QC have studied different machine learning methods. (Zhang and Lee, 2003) propose a QC system that uses Support Vector Machine (SVM) as the best machine learning algorithm. They compare the obtained results with other algorithms, such as Nearest Neighbors, Naive Bayes, Decision Tree or Sparse Network of Winnows (SNoW).

(Li and Roth, 2002) propose a system based on SNoW. They used five main classes and fifty fined classes. Other systems have used SVM and modified kernels.

QC systems have some restrictions (Hacioglu and Ward, 2003), such as:

- Traditional question classification uses a set of rules, for instance "questions that start with *Who* ask about a person". These are manual rules that have to be revised to improve the results.
- These rules are very weak, because when new questions arise, the system has to be updated to classify them.

Most of the QC systems use English as the main language, and some of the best and standard resources are developed for English.

It would be possible to build a question classifier for every language based on machine learning, using a good training corpus for each language, but is something expensive to produce. For this reason we have used Machine Translation Systems.

Machine Translation (MT) systems are very appreciated in CLIR (McNamee et al., 2000). Last years these systems have improved the results, but there are not translators for each language pair and the quality of the result depends on this pair.

The reason of using MT and not a Spanish classifier is simple: we have developed a multilingual QA system that works in this moment with three languages: English, Spanish and French. Because it is too complex for us to work with resources into these three languages and also to manage the information into three languages, our kernel system works into English, and we use MT to translate information when it is necessary.

We have developed a QC system that covers three tasks:

- It uses machine learning algorithms. We have tested methods based on Support Vector Machine, for instance SVMLight or LibSVM, and TiMBL. TiMBL [1] is a program that implements several Memory-Based Learning techniques. It stores a representation of the training set explicitly in memory, and classifies new cases by extrapolation from the most similar stored cases.
- To classify Spanish questions we have checked two online machine translators. Our proposal is to study how the translation can affect in the final results, compared to the original English results.
- Finally, we would obtain different results applying different levels of features (lexical, syntactic and semantic). In the next section well explain them and in results chapter we will see these differences.

Our QC system has three independent modules, so it will be easy to replace each one with other to improve the final results. In Figure 1 we can see them.

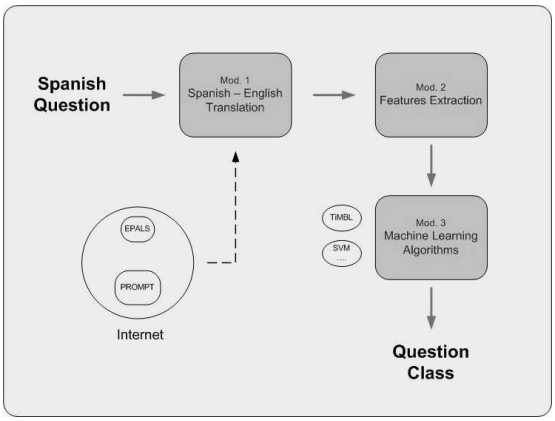

Figure 1: QC system Modules.

The first module translates the question into other languages, Spanish in this case. We have used two machine translation systema that work well for the language pair Spanish-English: Epals and Prompt. This module could work with other machine translation systems and other languages if there would be a good translator for the language pair used.

The second module extracts some relevant features (see next section) using the original or the translated English questions. Some of these features will be used by the machine learning module (lexical, syntactic and semantic features) and the

[1] ILK Research Group, Tilburg University and CNTS Research Group, University of Antwerp

others will be used later in the answers extraction phase. Take into account that the second module also extracts important features such as the context of the question, the focus or the keywords that we would use in next steps of the Question Answering system.

The final module applies the machine learning algorithm and returns the question category or class. In our first experiments we used Library for Support Vector Machines (LibSVM) and Bayesian Logistic Regression (BBR), but for this one we have used Tilburg Memory Based Learner (TiMBL).

TiMBL (Daelemans et al., 2004) implements several Memory-Based Learning techniques, classic k-NN classification kernel and several metrics. It implements Stanfill modified value difference metric (MVDM), Jeffrey Divergence and Class voting in the k-NN kernel according to the distance of the nearest neighbors. It makes classification using heuristic approximations, such as the IGTREE decision tree algorithm and the TRIBL and TRIBL2 hybrids. It also has optimizations for fast classification.

2.1 Features in Question Classification

We have analyzed each question in order to extract the following features:

- Lexical Features
  - The two first words of the question
  - All the words of the question in lowercase
  - The stemming words
  - Bigrams of the question
  - Each word with its position in the question
  - The interrogative pronoun of the question
  - The headwords of the nouns and verbs
- Syntactic Features
  - The interrogative pronoun and the Part of Speech (POS) of the rest of the words
  - The headword (a word to which an independent meaning can be assigned) of the first noun phrase
  - All POS
  - Chunks
  - The first verb chunk
  - The length of the question
- Semantic Features
  - The question focus (a noun phrase that is likely to be present in the answer)
  - POS with the named entities recognized
  - The type of the entity if the focus is one of them
  - Wordnet hypernyms for the nouns and Wordnet synonyms for the verbs

We have used some English resources such as the POS tagger TreeTagger (Schmid, 1994), Lingpipe [2] to make Named Entity Recognition, and the Porter stemmer (Porter, 1980). We have also used Wordnet to expand the queries.

3 Experiments and Results

3.1 Experimental Method

The experiments are made using some public datasets available by USC (Hovy et al., 1999), UIUC and TREC [3] as training and test collections.

These datasets have been labeled manually by UIUC group by means of the following general and detailed categories:

ABBR: abbreviation, expansion.

DESC: definition, description, manner, reason.

ENTY: animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word.

HUM: description, group, individual, title.

LOC: city, country, mountain, other, state.

NUM: code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight.

For instance the question "*What does NATO mean?*" is an ABBR (abbreviation) category, "*What is a receptionist?*" is a DESC (definition) category or "*When did George Bush born?*" is a NUM (numeric) category.

The training data are a set of 5500 questions and the test data are a set of 500 questions. All questions were labelled for the 10th conference Cross-Language Evaluation Forum of Question Answering (CLEF-QA).

[2] LingPipe is a suite of Java tools designed to perform linguistic analysis on natural language data, available in http://www.alias-i.com/lingpipe

[3] http://trec.nist.gov

The same dataset has been used in other investigations, such as in (Li and Roth, 2002).

The distribution of these 5500 training questions, with respect to its interrogative pronoun or the initial word is showed in Table 1.

Likewise, the distribution of categories of these 5500 training questions is showed in Table 2.

Table 1: Training questions distribution according with its interrogative pronoun

Type	Number
What	3242
Who	577
How	764
Where	273
When	131
Which	105
Why	103
Name	91
In	67
Define	4
Whom	4
Others	91

Table 2: Training questions distribution according with its general category.

Category	Number
ABBR	86
DESC	1162
ENTY	1251
HUM	1223
LOC	835
NUM	896

The distribution of the 500 test questions, with respect to its interrogative pronoun or the initial word, is showed in Table 3, and the distribution of categories of these 500 test questions is showed in Table 4.

Table 3: Test questions distribution according with its interrogative pronoun.

Type	Number
What	343
Who	47
How	35
Where	26
When	26
Which	6
Why	4
Name	2
In	5
Others	6

In our experiments we try to identify the general category. Our proposal is to try a detailed classification later.

Table 4: Test questions distribution according with its general category.

Category	Number
ABBR	9
DESC	138
ENTY	94
HUM	65
LOC	81
NUM	113

We have used the Accuracy as a general measure and the Precision of each category as a detailed measure.

$$Accuracy = \frac{\sharp ofcorrectpredictions}{\sharp ofpredictions} \quad (1)$$

$$precision(\mathbf{c}) = \frac{\sharp ofcorrectpredictionsofthecategoryc}{\sharp ofpredictionsofthecategoryc} \quad (2)$$

Other measure used is the F-score, defined as the harmonic mean of precision and recall (Van Rijsbergen, 1979). It is a commonly used metric to summarize precision and recall in one measure.

$$F - score = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

3.2 Results

We have made some experiments changing the machine translation systems:

- 5500 training questions and 500 test questions, all into English. This is the basic case.
- 5500 training questions into English and 500 test questions translated from Spanish using the MT Epals.
- 5500 training questions into English and 500 test questions translated from Spanish using the MT Prompt.

The MT resources are available in the following URLs:

- *Epals*

  http://www.epals.com

- *Prompt*

  http://translation2.paralink.com

According to the lexical, syntactic and semantic features we have made seven features sets. Our proposal here is to check which ones increase the final results. These features sets are the following:

1. Lexical Features: interrogative pronoun *(lex1)*
2. Lexical and Syntactic Features: Two first words of the question + All the words of the question in lowercase + Stemming words + Headwords *(lexsyn2)*
3. Lexical and Syntactic Features: previous four + Each word with its position in the question + interrogative pronoun + The first verb chunk *(lexsyn3)*
4. Semantic Features: The question focus + POS with the named entities recognized + The type of the entity if the focus is one of them *(sem4)*
5. Syntactic Features: The interrogative pronoun and the Part of Speech (POS) of the rest of the words + All POS + Chunks + The length of the question *(sin5)*
6. All Lexical + all Syntactic + all Semantic *(lexsemsin6)*
7. Lexical Features: Two first words of the question + interrogative pronoun ; Syntactic Features: + The headwords of the nouns and verbs + The first verb chunk + the interrogative pronoun + the Part of Speech (POS) of the rest of the words + The length of the question; Semantic Features: POS with the named entities recognized *(lexsemsin7)*

We can see in the Table 5 the obtained results in terms of global accuracy.

Table 5: Results in terms of Accuracy.

Features	English original	Epals	Prompt
lex1	0,458	0,334	0,414
lexsyn2	0,706	0,656	0,632
lexsyn3	0,718	0,638	0,612
sem4	0,675456	0,59798	0,629555
sin5	0,608	0,438	0,518
lexsemsin6	0,839757	0,662626	0,722672
lexsemsin7	0,8	0,678	0,674

Note that the average loss of precision is around 17% if we use Epals, and around 12% if we use Prompt.

(Li and Roth, 2002) obtain a better performance for English, around a 92.5% in terms of accuracy.

The best results are obtained when we use a combination of all lexical, syntactic and semantic features. The main reason is that the classifier works better when the number of features, which can be different to each category, is increased.

For future work, it will be also necessary to study the time consumption for each features set, to decide which ones can be used.

Table 6 shows the results in terms of F-score.

Table 6: Results in terms of F-score.

Features	English original	Epals	Prompt
lex1	0,476077	0,319793	0,441075
lexsyn2	0,708444	0,669692	0,6455
lexsyn3	0,721258	0,644813	0,614353
sem4	0,649405	0,593019	0,620068
sin5	0,576356	0,404038	0,48739
lexsemsin6	0,827789	0,664122	0,726667
lexsemsin7	0,795897	0,680039	0,68014

As an example in Table 7 we show detailed results for the best case, where the result for each general category is showed.

Table 7: Detailed results for each category, using the combination *lexsemsin6* and the original English questions and the translated questions by using Prompt

Class	English original		Prompt	
	Precision	*F-score*	*Precision*	*F-score*
ABBR	0.857	0.750	1	0.611
DESC	0.8442	0.906	0.695	0.806
ENTY	0.731	0.727	0.595	0.737
HUM	0.839	0.825	0.898	0.914
LOC	0.847	0.867	0.680	0.859
NUM	0.935	0.843	0.798	0.856

As we have seen there are no important differences between categories. In addition, this table shows that the translation results are reliable since for every category the lost of precision is similar (about 15%).

There are some reasons for the lost of precision. Some of them are the following:

1. Bad translation of synonym words. For instance we can compare an English original sentence: “What are the animals that don’t have backbones called?”, and its Prompt translation: “How are they called the animals that have no spines?”. The word *backbone* has been replaced with *spine*, so the IR system cannot find the same lists of relevant documents.
2. Translation of Named Entities. For instance we can compare an English original sentence: “Who was Galileo?”, and its Prompt translation: “Who was Galilean?”.

3. General bad translations. For instance we can compare an English original sentence: "Who discovered x-rays?", and its Prompt translation: "Who discovered the beams the Xth?".

4 Conclusions

Multilingual Question Answering systems have opened a new investigation task, where the question classification is an important first phase to know the type of answer and some relevant information about this question.

Our option is to use some standards resources for English and translate Spanish questions.

Of course we could develop a multilingual QC system using good training corpus for every language, but it is expensive to produce.

The use of machine translation systems is, then, very important, so the study of different online translators is a main task. In our case we have applied them to translate questions from Spanish into English.

We have made a complete investigation using the two datasets of training and test questions that have been used by other groups, all labelled manually. Different parameters have been the test file used (originally in English or translated from Spanish with the MT Epals or Prompt), the machine learning algorithm, some TiMBL parameters and the lexical, syntactic or semantic features.

The best results have been obtained using the original English questions and a combination of lexical, syntactic and semantic features. The best MT has been Prompt.

We have some conclusions:

- Applying machine learning with a complete set of training questions we obtain good results, over 0,8 in terms of accuracy.
- The use of machine translation systems decreases the results around 15%, but it will be possible to increase the performance using other models based on machine learning or a voting system for instance.
- A combination of all lexical, syntactic and semantic features obtains the best results.

As future work we want to check the system with other training and test datasets. We also want to design a voting system using different QC models; models based on patterns (to detect the class for some types of questions); models based on rules (filtering non-redundancy types of questions. For instance all questions with "who" are related to a person).

It could be also interested to test the combination between a better QC system, the current one by Li and Roths for instance (Li and Roth, 2002), and our machine translation method.

Finally, we want to study types of questions with poor results in order to improve them applying other techniques, such as question expansion.

Acknowledgement This work has been supported by Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

References

W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch. 2004. Timbl: Tilburgmemory based learner, version 5.1, reference guide. ilk technical report 04-02.

G. Grefenstette, editor. 1998. *Cross-Language Information Retrieval*, volume 1. Kluwer academic publishers, Boston, USA.

K. Hacioglu and W. Ward. 2003. Question classification with support vector machines and error correcting codes. In *Proceedings of Human Language Technology conference (HLT-NAACL)*.

E. Hovy, L. Gerber, U. Hermjakob, C. Lin, and D. Ravichandran. 1999. Towards sematics-based answer pinpointing. In *Proceedings of the DARPA Human Language Technology conference (HLT)*.

X. Li and D. Roth. 2002. Learning question classifiers. In *In COLING'02*.

P. McNamee, J. Mayfield, and C. Piatko. 2000. The jhu/apl haircut system at trec- 8. In *Proceedings of the Eighth Text Retrieval Conference (TREC8)*.

M. F. Porter. 1980. An algorithm for suffix stripping. In *Program 14*.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing.*

B. Van Durme, Y. Huang, A. Kupsc, and E. Nyberg. 2003. Towards light semantic processing for question answering. In *Proceedings of Human Language Technology conference (HLT-NAACL)*.

C.J. Van Rijsbergen. 1979. Information retrieval.

D. Zhang and W. Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.