Question Classification in Spanish and Portuguese

Thamar Solorio[1], Manuel Pérez-Coutiño[1], Manuel Montes-y-Gómez[1,2],
Luis Villaseñor-Pineda[1], and Aurelio López-López[1]

[1] Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica Óptica y Electrónica
Santa María Tonantzintla, Puebla, México 72840
[2] Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, España
{thamy,mapco,mmontesg,villasen,allopez}@inaoep.mx

Abstract. We present in this work a method for question classification in Spanish and Portuguese. The method relies on lexical features and attributes extracted from the Web. A machine learning algorithm, namely Support Vector Machines is successfully trained on these features. Our experimental results show that this method performs consistently well over two different languages.

1 Introduction

Question Classification (QC) is concerned with assigning a semantic category to questions posed in natural language. This semantic category corresponds to the type of answer needed for satisfying the user query. For instance, the question *In which European city is the Eiffel Tower?* belongs to the semantic class of "LOCATION". Most approaches to Question Answering systems perform some type of question classification given that the search space of possible answers is greatly reduced, also it has been shown that a poor performance in this stage of the system can provoke over one third of the errors [1]. However, most of these approaches are targeted to specific languages, this is because they use complex linguistic tools that are language dependent. Unfortunately for most languages these resources, such as part-of-speech taggers, named entity extractors, parsers, and so on, are not very well developed. Then, the adaptability of these methods to a different language is limited to those languages for which the linguistic tools are readily available.

In previous work we presented a language independent method for question classification were evaluation was performed on three languages: English, Spanish and Italian [2]. Although we achieved high accuracies we believe that considerable improvements can be attained by modifying some of the weakest features of this method, namely the set of heuristics chosen in order to construct the Internet queries. In this paper we present results of some modifications to this approach applied to questions written in Portuguese and Spanish. Our motivation is to provide a method for question classification that can be applied to different languages without requiring additional linguistic tools, such as parsers, named entity extractors and the like.

We first summarize some of the previous related work and describe the data sets used in our evaluation. Then we introduce the problem of question classification, we describe the lexical features used in the learning process and how the Web can be successfully used in this problem. We present some evaluation results and conclude this article with the findings of this work and interesting directions of future research.

2 Related Work

Li and Roth reported a hierarchical approach for question classification in English based on the SNoW (Sparse Network of Winnows) learning architecture [3]. This hierarchical classifier discriminates among 5 coarse classes, which are then refined into 50 more specific classes. The learners are trained using lexical and syntactic features such as part-of-speech tags, chunks and head chunks together with two semantic features: named entities and semantically related words. They reported question classification accuracy of 98.80% for a coarse classification, using 5,500 instances for training.

A different approach, used for Japanese question classification, is that of Suzuki *et al.* [4]. They used SVM with a new kernel function, called Hierarchical Directed Acyclic Graph, which allows the use of structured data. They experimented with 68 question types and compared performance of using bag-of-words against using more elaborated combinations of attributes, namely named entities and semantic information. Their best results, an accuracy of 94.8% at the first level of the hierarchy, were obtained when using SVM trained on bag-of-words together with named entities and semantic information.

In [5] Zhang and Sun Lee present a new method for question classification using Support Vector Machines targeted to English. They compared accuracy of SVM against Nearest Neighbors, Naive Bayes, Decision Trees and SNoW, with SVM producing the best results. In their work, accuracy is improved by introducing a tree kernel function that allows to represent the syntactic structure of questions. Their experimental results show that SVM using this tree kernel function achieves an accuracy of 90%, however, a parser is needed in order to acquire the syntactic information.

The idea of using the Internet in a QA system is not new. What is new, however, is that we are using the Internet to obtain values for features in our question classification process, as opposed to previous approaches where the redundancy of information available on the Internet has been used in the answer extraction process [6–9].

3 Data sets

The data set used in this work consists of the questions provided in the DISEQuA Corpus [10]. Such corpus was made up of simple, mostly short, straightforward and factual queries that sound naturally spontaneous, and arisen from a real desire to know something about a particular event or situation. The DISEQuA Corpus contains 450 questions, each one formulated in four languages: Dutch, English, Italian and Spanish. The questions are classified into seven categories: *Person*, *Organization*, *Measure*, *Date*, *Object*, *Other* and *Place*. The experiments performed in this work used the Spanish versions of these questions.

For Portuguese questions we use a data set consisting of 180 questions taken from the data sets used in CLEF 2004, the categories of the questions are the same as for the Spanish corpus.

4 Learning Question Classifiers

4.1 Lexical Features

With the aim of developing a flexible method we decided to use for learning only lexical features that can be automatically extracted from the questions. The most frequently used lexical features are bag-of-words and n-grams. Since we are using a machine learning technique, the n-grams approach seems the less desirable one, given that our training examples are limited and n-grams require a large training set. Then we opted for the bag-of-words approach, we also made a comparison of results between bag-of-words and prefixes.

As mentioned before, the lexical features are used as attributes for training a classifier. In this work, we used Support Vector Machines (SVM) as they have proved to perform well over natural language related problems such as text classification [11].

SVM use geometrical properties in order to compute the hyperplane that best separates a set of training examples [12]. When the input space is not linearly separable SVM can map, by using a kernel function, the original input space to a high-dimensional feature space where the optimal separable hyperplane can be easily calculated. This is a very powerful feature, because it allows SVM to overcome the limitations of linear boundaries. They also can avoid the over-fitting problems of neural networks as they are based on the structural risk minimization principle. The foundations of these machines were developed by Vapnik, for more information about this algorithm we refer the reader to [13, 14].

Table 1. Question classification accuracies when training SVM with words and prefixes of size 5 and 4

Language	Words	Prefix-5	Prefix-4
PORTUGUESE	75.14%	**75.73%**	74.55%
SPANISH	76.44%	**78.44%**	71.55%

In Table 1 we show experimental results of using SVM trained on three different sets of attributes: bag-of-words, prefixes of size 4 and prefixes of size

5. As we can see accuracies are very similar, with prefixes of size 5 achieving the best results for both languages. All the results reported here are the overall average of several runs of 10-fold cross-validation.

4.2 Using the Web

Previous results are encouraging considering that the only information needed to achieve these accuracies can be automatically extracted from the questions. However, we wanted to see if we can further improve these results by making use of the Web. The Web has become the greatest information source available worldwide, and although English is the dominant language represented on it, it is very likely that one can find information in almost any desired language. Considering this, and the fact that the texts are written in natural language, it is immediate to develop new methods exploring the use of the Web to solve natural language related problems [15]. Following this trend, we propose using the Web in order to acquire information that can be used as attributes in our classification problem. This attribute information can be extracted automatically from the web and the goal is to provide an estimate about the possible semantic class of the question.

The procedure for gathering this information from the web is as follows: we use a set of heuristics to extract from the question a word w, or set of words, that will complement the queries submitted for the search. We then go to a search engine, in this case Google, and submit queries using the word w in combination with all the semantic classes of interest for our purpose. For instance, for the question *Who is the President of the French Republic?* we extract the word *President* using our heuristics, and submit 5 queries in the search engine, one for each possible class. These queries take the following form:

- "President is a person"
- "President is a place"
- "President is a date"
- "President is a measure"
- "President is an organization"

We count the number of results returned by Google for each query and normalize them by their sum. The resultant numbers are the values for the attributes used by the learning algorithm. As can be seen, it is a very straightforward approach, but as the experimental results show, this information gathered from the Web is quite useful. In Table 2 we present the figures obtained from Google for the example question above, column *Results* show the number of hits returned by the search engine and in column *Normalized* we present the number of hits normalized by the total of all results returned for the different queries. It can be seen that Google returned hits for all the categories except for the "DATE" category, but the highest number of hits were returned for the category "PERSON", which is the real class of the question in our example.

Table 2. Example of using the Web to extract features for question classification

Query	Results	Normalized
“President is a person”	259	0.8662
“President is a place”	9	0.0301
“President is an organization”	11	0.0368
“President is a measure”	20	0.0669
“President is a date”	0	0

An additional advantage of using the Internet is that by approximating the values of attributes in this way, we take into account words or entities belonging to more than one class (polysemy).

Now that we have introduced the use of the Internet in this work, we continue describing the set of heuristics that we use in order to perform the web search.

Heuristics We begin by eliminating from the questions all words that appear in our stop lists. These stop lists contain the usual items: articles, prepositions and conjunctions plus all the interrogative adverbs and all lexical forms of the verb “to be”. The remaining words are sent to the search engine in combination with the possible semantic classes, as described above. If no results are returned for any of the semantic classes we then start eliminating words from right to left until the search engine returns results for at least one of the semantic categories. As an example consider the question posed previously: *Who is the President of the French Republic?* we eliminate the words from the stop list and then formulate queries for the remaining words. These queries are of the following form: *“President French Republic is a* s_i*”* where $s \in \{Person, Organization, Place, Date, Measure\}$. The search engine did not return any results for this query, so we start eliminating words from right to left. The query is now like this: *“President French is a* s_i*”* and given that again we have no results returned we finally formulate the last possible query: *“President is a* s_i*”* which returns results for all the semantic classes except for *Date.*

These were the heuristics used in previous experiments [2], in addition to these, in this work we run queries eliminating words in the reverse direction. That is, if no hits are returned after eliminating the stop words, we eliminate the first word to the left and continue repeating this process until we have results.

Being heuristics, we are aware that in some cases they do not work well. Nevertheless, for the vast majority of the cases they presented surprisingly good results, in the two languages, as shown in Table 3. What we did on this experiments was to compare results of training an SVM on the attributes from the Web. In column *Web RL* the heuristics used are those from the previous work, eliminating words from right to left. Column *Web LR* shows results eliminating words in the reverse order. These two columns seem to show that eliminating words from right to left yield more informative queries. In column *Web RL+LR* we present results of using both sets of attributes. That is, we combine the information from eliminating the words in both directions. These combination of attributes achieved the best results. These results also show that for Portuguese

Table 3. Experimental results of accuracy when training SVM with attributes extracted form the Web

Language	Web RL	Web LR	Web RL+LR
PORTUGUESE	59%	52.07%	**60.35%**
SPANISH	65.77%	44.66%	**67.11%**

the accuracies are much lower than for Spanish. We believe that this is due to the lower availability of Portuguese documents on the Web than for Spanish. The number of words available in Portuguese was near 1.3 billions whereas for Spanish was 2.6 billions [15].

4.3 Combining Web-extracted attributes with Lexical Features

So far we have shown that, on one hand, lexical features can provide enough information to build an automated classifier. On the other hand, information from the Web is not sufficient to provide accurate classifiers, the lack of language representation might be one reason for this. Yet, another possibility that we have to explore is a combination of these two types of features. Then, we performed new experiments combining the lexical attributes with the Web information in order to discover if we can further improve accuracy. Table 4 shows experimental results of this attribute combination and Figure 1 shows a graphical representation of these results. By comparing results presented in Tables 1 and 4 we can see that the best results are acquired using a combination of features. Even though the Web-based attributes did not seem to provide very interesting results at first, combining them with the lexical features did yield higher classification accuracy.

Table 4. Accuracies of combining the Web-extracted attributes (Web) with lexical features. The web extracted attributes are the combination of Left-to-Right and Right-to-Left presented in section 4.2

Language	Words+Web	Prefix-5+Web	Prefix-4+Web
PORTUGUESE	76.33%	75.53%	**78.1%**
SPANISH	78.33%	**79.11%**	78.22%

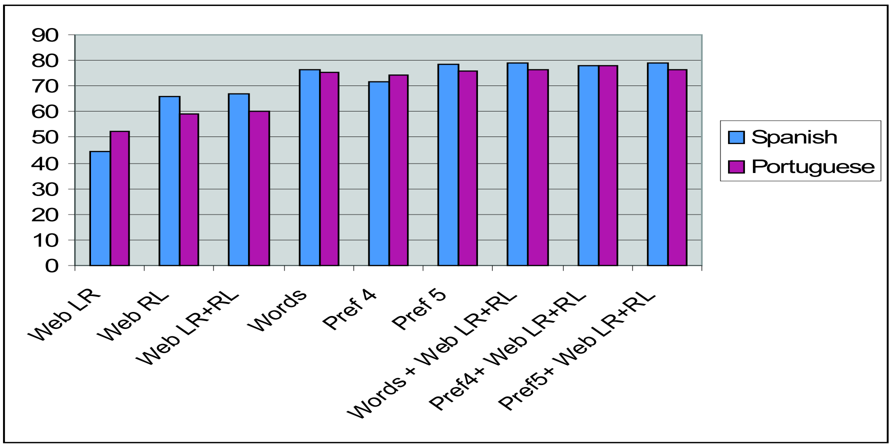

Fig. 1. Graphical comparison of question classification accuracies

5 Conclusions

We have presented here experimental results of a very flexible method for question classification. The method is claimed to be language independent to a good degree since the features used as attributes in the learning task can be extracted from the questions in a fully automated manner; we do not use semantic or syntactic information because otherwise we will be restricted to work on languages for which we do have parsers that can extract this information. We believe that this method can be successfully applied to other languages, such as Romanian, French and Catalan.

We are currently working on improving the heuristics used, we believe that better queries, formulated in a more careful manner, will help increase classification accuracy.

Another interesting line for future work is exploring the advantage of using mixed languages corpora lo learn question classification. The Romance languages, for instance, such as Italian, French and Spanish have stems in common. Then it is feasible that questions for several languages may help to train a classifier for a different language. The advantage of this idea will be the availability of larger corpora for languages for which a large enough corpus is not available, counting in favor of languages that are under-represented on the Web.

Acknowledgements

We would like to thank CONACyT for partially supporting this work under grants 166934, 166876, U39957-Y and 43990A-1, and Secretaría de Estado de Educación y Universidades de España.

References

1. D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.*, 21(2):133–154, 2003.
2. T. Solorio, M. Pérez-Coutiño, M. Montes y Gómez, L. Villaseñor-Pineda, and A. López-López. A language independent method for question classification. In *The 20th International Conference on Computational Linguistics, COLING-04*, Geneva, Switzerland, 2004.
3. X. Li and D. Roth. Learning question classifiers. In *COLING'02*, 2002.
4. J. Suzuki, H. Taira, Y. Sasaki, and E. Maeda. Question classification using HDAG kernel. In *Workshop on Multilingual Summarization and Question Answering 2003*, pages 61–68, 2003.
5. D. Zhang and W. Sun Lee. Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 26–32, Toronto, Canada, 2003. ACM Press.
6. E. Brill, S. Dumais, and M. Banko. An analysis of the AskMSR question-answering system. In *2002 Conference on Empirical Methods in Natural Language Processing*, 2002.
7. J. Lin, A. Fernandes, B. Katz, G. Marton, and S. Tellex. Extracting answers from the web using knowledge annotation and knowledge mining techniques. In *Eleventh Text REtrieval Conference (TREC 2002)*, Gaithersburg, Maryland, November 2002.
8. B. Katz, J. Lin, D. Loreto, W. Hildebrandt, M. Bilotti, S. Felshin, A. Fernandes, G. Marton, and F. Mora. Integrating web-based and corpus-based techniques for question answering. In *Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, Maryland, November 2003.
9. A. Del-Castillo, M. Montes y Gómez, and L. Villaseñor-Pineda. Qa on the web: A preliminary study for spanish language. In *Encuentro Internacional de Ciencias de la Computación (ENC'04)*, Colima, Mexico, September 2004.
10. B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke. Creating the DISEQuA corpus: a test set for multilingual question answering. In Carol Peters, editor, *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway, August 2003.
11. T. Joachims. *Learning to Classify Text using Support Vector Machines: Methods Theory and Algorithms*, volume 668 of *The Kluwer International Series in Engineering and Computer Science.* Kluwer Academic Publishers, 2002.
12. M. O. Stitson, J. A. E. Wetson, A. Gammerman, V. Vovk, and V. Vapnik. Theory of support vector machines. Technical Report CSD-TR-96-17, Royal Holloway University of London, England, December 1996.
13. V. Vapnik. *The Nature of Statistical Learning Theory.* Number ISBN 0-387-94559-8. Springer, N.Y., 1995.
14. B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, 2002.
15. A. Kilgarriff and G. Grefenstette. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347, 2003.