Question Classification Transfer

Anne-Laure Ligozat
LIMSI-CNRS / BP133, 91403 Orsay cedex, France
ENSIIE / 1, square de la résistance, Evry, France
firstname.lastname@limsi.fr

Abstract

Question answering systems have been developed for many languages, but most resources were created for English, which can be a problem when developing a system in another language such as French. In particular, for question classification, no labeled question corpus is available for French, so this paper studies the possibility to use existing English corpora and transfer a classification by translating the question and their labels. By translating the training corpus, we obtain results close to a monolingual setting.

1 Introduction

In question answering (QA), as in most Natural Language Processing domains, English is the best resourced language, in terms of corpora, lexicons, or systems. Many methods are based on supervised machine learning which is made possible by the great amount of resources for this language.

While developing a question answering system for French, we were thus limited by the lack of resources for this language. Some were created, for example for answer validation (Grappy et al., 2011). Yet, for question classification, although question corpora in French exist, only a small part of them is annotated with question classes, and such an annotation is costly. We thus wondered if it was possible to use existing English corpora, in this case the data used in (Li and Roth, 2002), to create a classification module for French.

Transfering knowledge from one language to another is usually done by exploiting parallel corpora; yet in this case, few such corpora exists (CLEF QA datasets could be used, but question classes are not very precise). We thus investigated the possibility of using machine translation to create a parallel corpus, as has been done for spoken language understanding (Jabaian et al., 2011) for example. The idea is that using machine translation would enable us to have a large training corpus, either by using the English one and translating the test corpus, or by translating the training corpus. One of the questions posed was whether the quality of present machine translation systems would enable to learn the classification properly.

This paper presents a question classification transfer method, which results are close to those of a monolingual system. The contributions of the paper are the following:

- comparison of train-on-target and test-on-source strategies for question classification;
- creation of an effective question classification system for French, with minimal annotation effort.

This paper is organized as follows: The problem of Question Classification is defined in section 2. The proposed methods are presented in section 3, and the experiments in section 4. Section 5 details the related works in Question Answering. Finally, Section 6 concludes with a summary and a few directions for future work.

2 Problem definition

A Question Answering (QA) system aims at returning a precise answer to a natural language question: if asked ”How large is the Lincoln Memorial?”, a QA system should return the answer ”164 acres” as well as a justifying snippet. Most systems include a question classification step which determines the expected answer type, for example *area* in the previous case. This type can then be used to extract the correct answer in documents.

Detecting the answer type is usually considered as a multiclass classification problem, with each answer type representing a class. (Zhang and

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 429–433,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

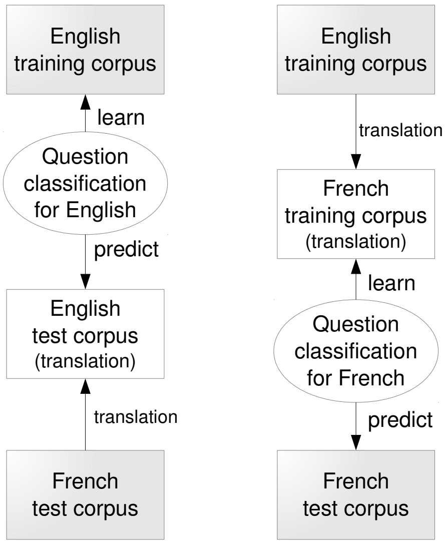
English training corpus
learn
Question classification for English
predict
English test corpus (translation)
translation
French test corpus
English training corpus
translation
French training corpus (translation)
learn
Question classification for French
predict
French test corpus

Figure 1: Methods for transfering question classification

Lee, 2003) showed that a training corpus of several thousands of questions was required to obtain around 90% correct classification, which makes it a costly process to adapt a system to another language than English. In this paper, we wish to learn such a system for French, without having to manually annotate thousands of questions.

3 Transfering question classification

The two methods tested for transfering the classification, following (Jabaian et al., 2011), are presented in Figure 1:

- The first one (on the left), called *test-on-source*, consists in learning a classification model in English, and to translate the test corpus from French to English, in order to apply the English model on the translated test corpus.

- The second one (on the right), called *train-on-target*, consists in translating the training corpus from English to French. We obtain an labeled French corpus, on which it is possible to learn a classification model.

In the first case, classification is learned on well written questions; yet, as the test corpus is translated, translation errors may disturb the classifier. In the second case, the classification model will be learned on less well written questions, but the corpus may be large enough to compensate for the loss in quality.

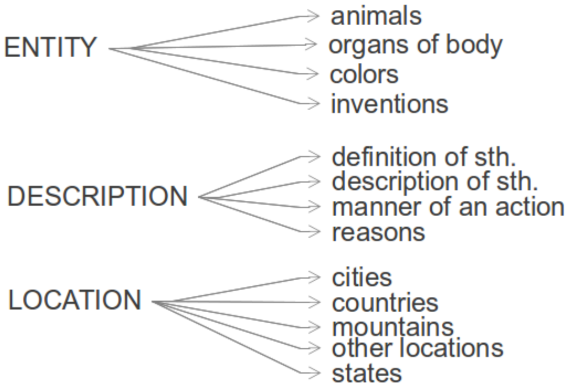

Figure 2: Some of the question categories proposed by (Li and Roth, 2002)

4 Experiments

4.1 Question classes

We used the question taxonomy proposed by (Li and Roth, 2002), which enabled us to compare our results to those obtained by (Zhang and Lee, 2003) on English. This taxonomy contains two levels: the first one contains 50 fine grained categories, the second one contains 6 coarse grained categories. Figure 2 presents a few of these categories.

4.2 Corpora

For English, we used the data from (Li and Roth, 2002), which was assembled from USC, UIUC and TREC collections, and has been manually labeled according to their taxonomy. The training set contains 5,500 labeled questions, and the testing set contains 500 questions.

For French, we gathered questions from several evaluation campaigns: QA@CLEF 2005, 2006, 2007, EQueR and Quæro 2008, 2009 and 2010. After elimination of duplicated questions, we obtained a corpus of 1,421 questions, which were divided into a training set of 728 questions, and a test set of 693 questions [1]. Some of these questions were already labeled, and we manually annotated the rest of them.

Translation was performed by Google Translate online interface, which had satisfactory performance on interrogative forms, which are not well handled by all machine translation systems [2].

[1] This distribution is due to further constraints on the system.

[2] We tested other translation systems, but Google Translate gave the best results.

Train	en	en	fr (trans.)	fr
Test	en	en (trans.)	fr	fr
Method		test-on-source	train-on-target	
50 classes	.798	.677	**.794**	.769
6 classes	.90	.735	**.828**	.84

Table 1: Question classification precision for both levels of the hierarchy (features = word n-grams, classifier = libsvm)

4.3 Classification parameters

The classifier used was LibSVM (Chang and Lin, 2011) with default parameters, which offers one-vs-one multiclass classification, and which (Zhang and Lee, 2003) showed to be most effective for this task.

We only considered surface features, and extracted bag-of-ngrams (with $n = 1..2$).

4.4 Results and discussion

Table 1 shows the results obtained with the basic configuration, for both transfer methods.

Results are given in precision, i.e. the proportion of correctly classified questions among the test questions [3].

Using word n-grams, monolingual English classification obtains .798 correct classification for the fine grained classes, and .90 for the coarse grained classes, results which are very close to those obtained by (Zhang and Lee, 2003).

On French, we obtain lower results: .769 for fine grained classes, and .84 for coarse grained classes, probably mostly due to the smallest size of the training corpus: (Zhang and Lee, 2003) had a precision of .65 for the fine grained classification with a 1,000 questions training corpus.

When translating test questions from French to English, classification precision decreases, as was expected from (Cumbreras et al., 2006). Yet, when translating the training corpus from English to French and learning the classification model on this translated corpus, precision is close to the French monolingual one for coarse grained classes and a little higher than monolingual for fine grained classification (and close to the English monolingual one): this method gives precisions of .794 for fine grained classes and .828 for coarse grained classes.

Train	en	fr (trans.)	fr
Test	en	fr	fr
Method		train-on-target	
50 classes	.822	**.798**	.807
6 classes	.92	**.841**	.872

Table 2: Question classification precision for both levels of the hierarchy (features = word n-grams with abbreviations, classifier = libsvm)

One possible explanation is that the condition when test questions are translated is very sensitive to translation errors: if one of the test questions is not correcly translated, the classifier will have a hard time categorizing it. If the training corpus is translated, translation errors can be counterbalanced by correct translations. In the following results, we do not consider the ”en to en (trans)” method since it systematically gives lower results.

As results were lower than our existing rule-based method, we added parts-of-speech as features in order to try to improve them, as well as semantic classes: the classes are lists of words related to a particular category; for example ”president” usually means that a person is expected as an answer. Table 2 shows the classification performance with this additional information.

Classification is slightly improved, but only for coarse grained classes (the difference is not significant for fine grained classes).

When analyzing the results, we noted that most confusion errors were due to the type of features given as inputs: for example, to correctly classify the question ”What is BPH?” as a question expecting an expression corresponding to an abbreviation (*ABBR:exp class* in the hierarchy), it is necessary to know that ”BPH” is an abbreviation. We thus added a specific feature to detect if a question word is an abbreviation, simply by test-

[3] We measured the significance of precision differences (Student t test, p=.05), for each level of the hierarchy between each test, and, unless indicated otherwise, comparable results are significantly different in each condition.

Train	en	fr (trans.)	fr
Test	en	fr	fr
50 classes	.804	**.837**	.828
6 classes	.904	**.869**	.900

Table 3: Question classification precision for both levels of the hierarchy (features = word n-grams with abbreviations, classifier = libsvm)

ing if it contains only upper case letters, and normalizing them. Table 3 gives the results with this additional feature (we only kept the method with translation of the training corpus since results were much higher).

Precision is improved for both levels of the hierarchy: for fine grained classes, results increase from .794 to .837, and for coarse grained classes, from .828 to .869. Remaining classification errors are much more disparate.

5 Related work

Most question answering systems include question classification, which is generally based on supervised learning. (Li and Roth, 2002) trained the SNoW hierarchical classifier for question classification, with a 50 classes fine grained hierarchy, and a coarse grained one of 6 classes. The features used are words, parts-of-speech, chunks, named entities, chunk heads and words related to a class. They obtain 98.8% correct classification of the coarse grained classes, and 95% on the fine grained one. This hierarchy was widely used by other QA systems.

(Zhang and Lee, 2003) studied the classification performance according to the classifier and training dataser size, as well as the contribution of question parse trees. Their results are 87% correct classification on coarse grained classes and 80% on fine grained classes with vectorial attributes, and 90% correct classification on coarse grained classes and 80% on fine grained classes with structured input and tree kerneks.

These question classifications were used for English only. Adapting the methods to other languages requires to annotated large corpora of questions.

In order to classify questions in different languages, (Solorio et al., 2004) proposed an internet based approach to determine the expected type. By combining this information with question words, they obtain 84% correct classification for English, 84% for Spanish and 89% for Italian, with a cross validation on a 450 question corpus for 7 question classes. One of the limitations raised by the authors is the lack of large labeled corpora for all languages.

A possibility to overcome this lack of resources is to use existing English resources. (Cumbreras et al., 2006) developed a QA system for Spanish, based on an English QA system, by translating the questions from Spanish to English. They obtain a 65% precision for Spanish question classification, while English classification are correctly classified with an 80% precision. This method thus leads to an important drop in performance.

Crosslingual QA systems, in which the question is in a different language than the documents, also usually rely on English systems, and translate answers for example (Bos and Nissim, 2006; Bowden et al., 2008).

6 Conclusion

This paper presents a comparison between two transfer modes to adapt question classification from English to French. Results show that translating the training corpus gives better results than translating the test corpus.

Part-of-speech information only was used, but since (Zhang and Lee, 2003) showed that best results are obtained with parse trees and tree kernels, it could be interesting to test this additional information; yet, parsing translated questions may prove unreliable.

Finally, as interrogative forms occur rarely is corpora, their translation is usually of a slightly lower quality. A possible future direction for this work could be to use a specific model of translation for questions in order to learn question classification on higher quality translations.

References

J. Bos and M. Nissim. 2006. Cross-lingual question answering by answer translation. In *Working Notes of the Cross Language Evaluation Forum.*

M. Bowden, M. Olteanu, P. Suriyentrakorn, T. d´Silva, and D. Moldovan. 2008. Multilingual question answering through intermediate translation: Lcc´s poweranswer at qa@clef 2007. *Advances in Mul-*

tilingual and Multimodal Information Retrieval, 5152:273–283.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

M.Á.G. Cumbreras, L. López, and F.M. Santiago. 2006. Bruja: Question classification for spanish. using machine translation and an english classifier. In *Proceedings of the Workshop on Multilingual Question Answering*, pages 39–44. Association for Computational Linguistics.

Arnaud Grappy, Brigitte Grau, Mathieu-Henri Falco, Anne-Laure Ligozat, Isabelle Robba, and Anne Vilnat. 2011. Selecting answers to questions from web documents by a robust validation process. In *IEEE/WIC/ACM International Conference on Web Intelligence*.

Bassam Jabaian, Laurent Besacier, and Fabrice Lefèvre. 2011. Combination of stochastic understanding and machine translation systems for language portability of dialogue systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5612–5615. IEEE.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

T. Solorio, M. Pérez-Coutino, et al. 2004. A language independent method for question classification. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 1374–1380. Association for Computational Linguistics.

D. Zhang and W.S. Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 26–32. ACM.