Question Classification in English-Chinese Cross-Language Question Answering: An Integrated Genetic Algorithm and Machine Learning Approach

Min-Yuh Day [a, b], Chorng-Shyong Ong [b], and Wen-Lian Hsu [a,*], *Fellow, IEEE*
[a] *Institute of Information Science, Academia Sinica, Taiwan*
[b] *Department of Information Management , National Taiwan University, Taiwan*
{myday, hsu}@iis.sinica.edu.tw; ongcs@im.ntu.edu.tw

Abstract

Question classification plays an important role in cross-language question answering (CLQA) systems, while question Informer plays a key role in enhancing question classification for factual question answering. In this paper, we propose an integrated Genetic Algorithm (GA) and Machine Learning (ML) approach for question classification in English-Chinese cross-language question answering. To enhance question informer prediction, we use a hybrid method that integrates GA and Conditional Random Fields (CRF) to optimize feature subset selection in a CRF-based question informer prediction model. The proposed approach extends cross-language question classification by using the GA-CRF question informer feature with Support Vector Machines (SVM). The results of evaluations on the NTCIR-6 CLQA question sets demonstrate the efficacy of the approach in improving the accuracy of question classification in English-Chinese cross-language question answering.

1. Introduction

Question classification plays an important role in cross-language question answering (CLQA) systems[11, 16], such as NTCIR CLQA (Cross-Language Question Answering) and QA@CLEF (Question Answering at Cross Language Evaluation Forum). The goal of question classification is to accurately classify a question in to a question type and then map it to an expected answer type (question type determination) [2]. For example, the question classification for "What is the biggest city in the United States?" (question) is "Q_LOCATION_CITY" (question type). Question types thus derived are used to extract and filter answers in order to improve the overall accuracy of a cross-language question answering system.

Question informer plays a key role in enhancing question classification for factual question answering. Krishnan et al. [5] introduced the notion of the answer type informer span of a question and showed that human-annotated informer spans substantially improve the accuracy of machine learning-based question classification. They define question informer as choosing a minimal, appropriate contiguous span of a question token, or tokens, as the informer span of question that is adequate for question classification. For example, in the question: "What is the biggest city in the United States?" the question informer is "city". Thus, "city" is the most important clue for question classification. In contrast, we define a question informer as the most important clue for question classification. Hence, in the above example "city" is actually the most important clue. Note that question informers are only useful if their informer spans can be identified automatically.

In machine learning approaches, feature selection is an optimization problem that involves choosing an appropriate feature subset. Day et al. [3] showed that a hybrid approach that integrates Genetic Algorithm (GA) and Conditional Random Fields (CRF) improves the accuracy of question informer prediction in traditional CRF models.

In this paper, we propose an Integrated Genetic Algorithm (GA) and Machine Learning (ML) approach for question classification in cross-language question answering. Specifically, we focus on a bilingual QA system for English source language queries and Chinese target document collections. To enhance the hybrid approach for cross-language question classification, we use the GA-CRF question informer feature with Support Vector Machines (SVM).

The remainder of the paper is organized as follows. Section 2 describes the background to cross language question classification and reviews related works. In Section 3, we propose an Integrated Genetic Algorithm (GA) and Machine Learning (ML) approach for cross-language question classification. Section 4 discusses the experiment and the test bed, and Section 5 details the experimental results. Finally, in Section 6 we present our conclusions and indicate future research directions.

2. Research Background

Numerous works on question classification for cross-language question answering have been reported in literature [2, 4, 10, 13, 18, 19, 21].

2.1. Cross Language Question Answering

There are three international Question Answering (QA) contests: TREC QA[20], QA@CLEF[11], and NTCIR CLQA[16]. The Text REtrieval Conference Question Answering track (TREC QA, http://trec.nist.gov/) has provided the evaluation standard for monolingual QA in English since 1999. The Question Answering track at Cross Language Evaluation Forum (QA@CLEF, http://www.clef-campaign.org/) has provided a question answering infrastructure for European languages in both non-English monolingual and cross-language contexts since 2003. The NII-NACSIS Test Collection for IR Systems Cross Language Question Answering (NTCIR CLQA, http://clqa.jpn.org/) has held an evaluation contest for Cross-Lingual Question Answering technology for Asian languages in both monolingual (i.e., Chinese) and cross-languages (i.e., English-Chinese, English-Japanese,) since 2005.

2.2. Question Classification

Approaches to question classification can be divided in two broad classes, namely, rule-based and machine learning methods. Most recent studies have been based on machine learning approaches.

Li and Roth [10] proposed 6 coarse classes and 50 fine classes for TREC factoid question answering. The UIUC QC dataset, which they developed, contains 5,500 training questions and 500 test questions, and it is now the standard dataset for question classification [3]. Li and Roth use the Sparse Network of Windows (SNoW) with over 90% accuracy.

Krishnan et al. [5] used SVM with question bi-grams, CRF question informer q-grams, and informer hypernyms. On the UIUC dataset, they derived coarse-grained categories with 93.4% accuracy and fine-grained categories with 86.2% accuracy.

Plamondon and Foster [14] proposed a method that relies on a statistical translation engine to translate keywords, as well as a set of manually written rules for analyzing French questions, so that a monolingual English question answering system can be modified to accept French questions. Using regular expressions that combine words and part-of speech tags to analyze a question, they wrote approximately 60 analysis patterns in both English and French.

Kwok and Deng [7] used heuristic rules with cue words and adjacent meta-keywords to assign a possible answer class to an English question and attained approximately 80% accuracy on the NTCIR-5 test set. In contrast, Day et al. [2] achieved 92% accuracy on the same test set by using an integrated knowledge-based and machine learning approach.

Question classification for multi-lingual queries can be performed by a single question classifier or multiple question classifiers. For instance, in English-Chinese cross-language question answering, there are two strategies for question classification: 1) Chinese Question Classification (CQC) for both English and Chinese queries. In this case, the English source language has to be translated into the Chinese target language in advance. 2) English Question Classification (EQC) for English queries and Chinese Question Classification (CQC) for Chinese queries.

In this paper, we focus on question classification in English-Chinese cross-language question answering, which is a bilingual QA system for English source language queries and Chinese target document collections. Hence, we adopt a two-question classifier strategy, namely, English Question Classification and Chinese Question Classification, for question classification in English-Chinese cross-language question answering.

3. Methods

3.1. Hybrid GA-CRF-SVM Architecture

We propose an Integrated Genetic Algorithm (GA) and Machine Learning (ML) approach for question classification in cross-language question answering. The architecture of the proposed, shown in Figure 1, comprises three phases for transforming an input question into output question type: 1) the GA feature selection phase, which uses GA for CRF feature selection to obtain a near optimal feature subset of CRF; 2) GA-CRF question informer prediction, which uses the near optimal CRF question informer prediction model to predict question informers; and 3) SVM-based question classification, which uses the GA-CRF predicted question informers as the key features for SVM-based question classification.

3.2. GA for CRF Feature Selection

We use CRF++ [6], developed by Taku Kudo, to predict question informers because it allows us to redefine feature sets and specify the feature templates in a flexible manner. We use GA to generate the best feature templates for CRF++.

The application of GA to obtain the near optimal feature subset of CRF involves the following steps.

1) Encode a feature subset of CRF with the structure of chromosomes. The value of the codes for feature subset selection is set to a one-bit digit, ‘0’ or ‘1’, where ‘0’ indicates that the corresponding feature is not selected, and ‘1’ means that it is selected. The length of each chromosome is *n* bits, where *n* is the number of features.

2) Initialization: Generate the initial population, which is initialed with random values before the search process.

3) Population: Use the initial population, which is a set of seed chromosomes, to find the optimal feature subsets.

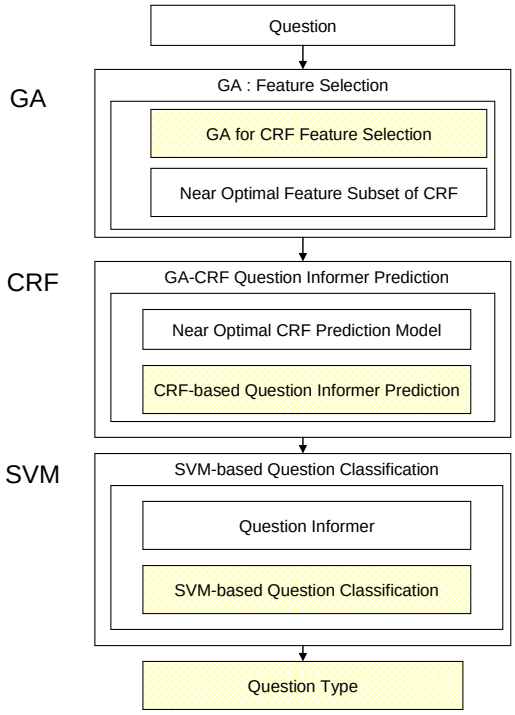

Figure 1. The architecture of the proposed integrated Genetic Algorithm (GA) and Machine Learning (ML) approach for question classification in English-Chinese cross-language question answering.

4) Evaluation: Calculate the fitness score of each chromosome.

5) CRF model 10-fold cross validation: Apply the feature subsets derived by the previous procedure to the CRF module. The fitness function is determined by the F-score of 10-fold cross validation of the CRF model. We use 10-fold cross validation on the training dataset of the CRF model as the fitness function of each chromosome to avoid over-fitting on the test dataset.

6) Stopping criteria satisfied? If the stopping criteria are satisfied the best chromosome and near optimal feature subset of CRF model is obtained; otherwise, apply GA operators and produce a new generation.

7) Apply GA operators and produce a new generation: Use three GA operators, namely, reproduction, crossover, and mutation to produce a new generation.

In summary, we can obtain a near optimal feature subset of CRF after the GA procedures for CRF feature selection.

3.3. GA-CRF Question Informer Prediction

We integrate the GA architecture with CRF to optimize feature selection for CRF-based question informer prediction. This hybrid GA-CRF approach involves two phases: the GA-CRF learning phase with a training dataset, and the CRF test phase with a test dataset. The experimental results, detailed in Section 5, demonstrate that the hybrid GA-CRF model for question informer prediction improves the accuracy of the traditional CRF model.

3.4. SVM-based Question Classification using GA-CRF Question Informer

For English question classification, we use an SVM-based machine learning approach that incorporates GA-CRF predicted question informers as important features. Because SVM consistently outperforms other machine learning techniques in several tasks, including text classification [15, 17] and question classification [21], we adopt it as the machine learning approach for question classification. To implement it, we use SVMlight [15], an implementation of Vapnik's Support Vector Machine for pattern recognition.

4. Experiment Design

4.1. Data set

4.1.1 Data set for English Question Classification

- **Training dataset**

We use Li and Roth's UIUC QC dataset [10] and the corresponding Question Informer dataset from Krishnan et al. [5] to train the classification model. There are 5,500 training questions, 500 test questions, and the corresponding question informers. Li and Roth used supervised learning for question classification of the UIUC QC dataset; this is now the standard dataset for question classification [3]. It has 6 coarse-grained and 50 fine-grained answer types in a two-level taxonomy, as well as the above training and test questions.

We derived 4,204 valid questions tagged with their question types for CLQA factoid question answering. The questions were obtained from 6,000 UIUC questions with question informers by mapping the UIUC types to the ASQA [8, 9] question types. The question type taxonomy for English question classification includes 6 coarse-grained classes and 62 fine grained classes – the same as the Chinese question classification in ASQA [2, 8, 9].

For English question classification of NTCIR-6 CLQA English questions, we use an SVM model trained from 5,288 questions (ModelQ5288E: 4,204 questions from UIUC + 500 questions from the NTCIR-5 CLQA development set + 200 questions from the NTCIR-5 CLQA test set + 384 questions from TREC2002 500 questions). Note that we use different features (including question informers) to construct the SVM model based on a total of 5,288 English questions and their labeled question types.

- **Test dataset**

For English question classification, we use NTCIR-6 CLQA's formal run of 150 English questions (CLQA2T150E) as our test dataset.

4.1.2 Data set for Chinese Question Classification

- **Training dataset**

We use the IASLQ2322C training dataset with 2,322 Chinese questions for our SVM-based CQC. The questions are derived from three sources: 500 from the NTCIR-5 CLQA development set plus 200 from the NTCIR-5 CLQA test set, 384 from a translated TREC 2002 dataset in Chinese, and 1,238 that are manually built in IASL (http://iasl.iis.sinica.edu.tw).

- **Test dataset**

We use NTCIR-6 CLQA's formal run of 150 Chinese questions (CLQA2T150C) as our test dataset for Chinese question classification.

4.2. Features for English Question Classification

The following syntactic features and semantic features are used in EQC.

1. Syntactic features
   - Word-based bi-grams of the question (WB)
   - First word of the question (F1)
   - First two words of the question (F2)
   - Wh-word of the question, i.e., 6W1H1O: who, what, when, where, which, why, how, and other (WH)
2. Semantic features
   - Question informers predicted by the GA-CRF model (QIF)
   - Question informer bi-grams predicted by the GA-CRF model (QIFB)

4.3. Features for Chinese Question Classification

The following syntactic features and semantic features are used in CQC.

1. Syntactic features

We use two syntactic features in our SVM model: bag-of words (n-grams) and part-of-speech (POS).

- Bag-of-Words

Bag-of-words features are comprised of character-based bi-grams (CB) and word-based bi-grams (WB).

- Part-of-Speech (POS)

We use AUTOTAG [1], a POS tagger developed by CKIP, Academia Sinica, to obtain the POS of the given Chinese questions, and then use the POS features for CQC.

2. Semantic Features
   - HowNet Senses

We use "HowNet 2000" to derive the semantic features of the Chinese questions. Our SVM Model uses two semantic features, namely, HowNet Main Definition (HNMD) and HowNet Definition (HND).

- TongYiCi CiLin (TYC)

To enhance the robustness of CQC, we introduced a new semantic feature called TongYiCi CiLin (TYC) [12], a Chinese synonym dictionary, for the machine learning approach of Chinese Question Classification (CQC) in CLQA2. The TongYiCi we use is an extended version of TongyiciCilin (ECilin for short), developed by the Information Retrieval Laboratory of the Harbin Institute of Technology (http://www.ir-lab.org/).

4.4. Performance Metrics

We use accuracy and the mean reciprocal rank (MRR) [13] to evaluate the performance of question classification. Given a set of questions, *M*, their *corrected* question types, and a ranked list of classification scores, the accuracy of question classification is calculated as follows:

$$Accuracy = \frac{Number\ of\ corrected\ question\ types}{Total\ number\ of\ questions} \quad (1)$$

The MRR of question classification is calculated as follows [13]:

$$MRR = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{rank_i}, \quad (2)$$

where $rank_i$ is the rank of the first *corrected* question type of the i^{th} question, and *M* is total number of questions.

5. Experimental Results and Discussion

We now present the experimental results of the proposed approach for question classification in English-Chinese cross-language question answering.

5.1 Question Informer Prediction

For question informer prediction, the experimental results show that the proposed hybrid GA-CRF model of question informer prediction outperforms the traditional CRF model. Using GA to optimize the selection of the feature subset in CRF-based question informer prediction improves the F-score from 88.9% to 93.87%, and reduces the number of features from 105 to 40. Note that the fitness function is used to evaluate the test dataset (UIUC Q500) with the training dataset (UIUC Q5500). In addition, the accuracy of our proposed GA-CRF model for the UIUC dataset is 95.58% compared to 87% for the traditional CRF model reported by Krishnan et al. Thus, the proposed hybrid GA-CRF model for question informer prediction significantly outperforms the traditional CRF model.

5.2 English Question Classification

For English question classification, the fine-grained accuracy is 82.32% for 10-fold cross validation on the training dataset (IASLEQ5288E), and approximately

88.79% for the coarse-grained accuracy. The features used for SVM-based English question classification are WB (word bi-gram), F1 (first word), F2 (first two words), QIF (question informer), QIFB (question informer bi-gram), and WH (question wh-word, 6W1H1O: who, what, when, where, which, why, how, and other).

We also conducted an experiment on the training data of IASLEQ5088E and the test data of CLQA1T200E. The results show that by using Support Vector Machines (SVM), our approach enhances the fine-grained accuracy of English Question Classification (EQC) from 68.0% (WB) to 78.5% (WB+F1+F2+WH+QIF+QIFB). Meanwhile, the coarse-grained accuracy increases from 71.0% to 83.5%.

We use the 5,288 questions mentioned in Section 4 as our training dataset and the WB+F1+F2+WH+QIF+QIFB features to train our SVM model for the test dataset, which was taken from NTCIR-6 CLQA's formal run of 150 English questions (CLQA2T150E). The experimental results are as follows.

The top-1 accuracy of fine-grained English question classification is 94% for CLQA2T150E. The results of using different features in SVM models for English question classification are shown in Figure 2. It is significant that, by integrating GA-CRF-based question informer prediction as a feature, the SVM-based English question classification model performs better than the model that uses the baseline word-based bi-grams feature.

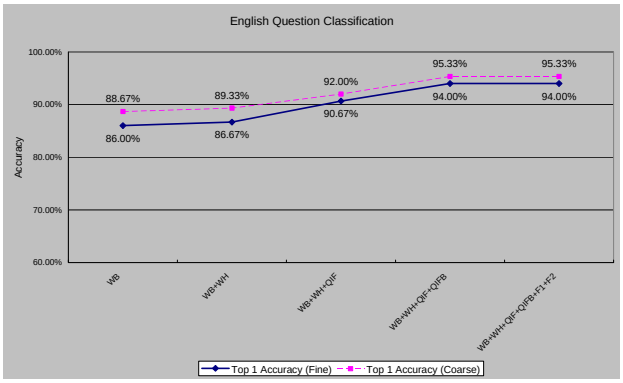

Figure 2. Experimental results for English Question Classification (EQC) using SVM

5.3 Chinese Question Classification

The features used for SVM-based Chinese question classification are 1) syntactic features: Chinese characters (C), Chinese character-based bi-grams (CB), Chinese words (W), Chinese word-based bi-grams (WB), Part-of-Speech (POS), and Part-of-Speech bi-grams (POSB); and 2) semantic features: HowNet Main Definition (HNMD), HowNet Definition (HND), TongYiCi (TYC). We use the 2,322 Chinese questions (IASLQ2322C) as our training dataset, and combinations of syntactic and semantic

Table 1. Experimental results of Chinese Question Classification (CQC) using SVM with different features

Feature Used	Top 1 Accuracy (Fine)	Top 1 Accuracy (Coarse)	Top 5 MRR (Fine)	Top 5 MRR (Coarse)
POS	53.33%	65.33%	0.5732	0.7533
POSB	60.00%	74.00%	0.6469	0.7970
HNMD	71.33%	81.33%	0.7480	0.8832
CB	74.00%	84.67%	0.7934	0.9130
HNMDB	74.00%	86.00%	0.7916	0.9117
C	74.67%	84.67%	0.7979	0.9152
TYCB	74.67%	86.00%	0.7880	0.9062
HND	74.67%	86.67%	0.7860	0.9102
W	76.00%	88.00%	0.7901	0.9208
HNDB	76.67%	88.00%	0.8000	0.9162
WB	77.33%	88.00%	0.8067	0.9162
TYC	77.33%	88.67%	0.8019	0.9240

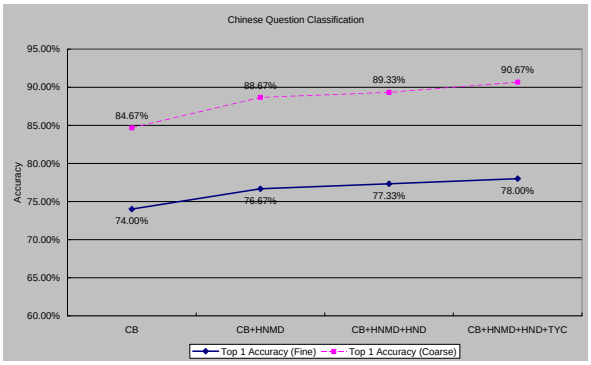

Figure 3. Experimental results of Chinese Question Classification (CQC) using SVM

features (CB+HNMD+HND+TYC) to train our SVM model for the test dataset questions, which are taken from NTCIR-6 CLQA's formal run of 150 Chinese questions (CLQA2T150C). We compare the contribution of different syntactic and semantic features to the classification performance. Table 1 shows the results of Chinese Question Classification (CQC) using SVM with different features. We observe that TYC outperforms HND and HNMD. The top-1 accuracy derived by using TYC solely is 77.33%, compared to 74.67% for HND solely, and 71.33% for HNMD solely.

However, the best Chinese question classification performance is achieved by using a combination of syntactic and semantic features (CB+NHMD+HND+TYC). Figure 3 shows the results of using SVM with different combinations of features for Chinese question classification. The top-1 accuracy of fine-grained Chinese question classification using SVM with a combination of syntactic and semantic features is

78% for CLQA2T150C, while the coarse-grained top-1 accuracy is 90.67%.

6. Conclusions

In this paper, we have proposed a hybrid genetic algorithm and machine learning approach for cross-language question classification. Our English question classifier incorporates GA-CRF based question informer as a key feature for question classification. The major contribution of this paper is that the proposed approach enhances cross-language question classification by using the GA-CRF question informer feature with Support Vector Machines (SVM). The results of experiments on NTCIR-6 CLQA question sets demonstrate the efficacy of the approach in improving the accuracy of question classification in English-Chinese cross-language question answering.

7. Acknowledgements

This research was supported in part by the thematic program of Academia Sinica under Grant AS 95ASIA02, and by the National Science Council under Grants NSC 95-2752-E-001-001-PAE and NSC 95-2416-H-002-047.

8. References

[1] CKIP, "CKIP AutoTag," 2006.

[2] M.-Y. Day, C.-W. Lee, S.-H. Wu, C.-S. Ong, and W.-L. Hsu, "An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Classification," in *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2005)* Wuhan, China, 2005, pp. 620-625.

[3] M.-Y. Day, C.-H. Lu, C.-S. Ong, S.-H. Wu, and W.-L. Hsu, "Integrating Genetic Algorithms with Conditional Random Fields to Enhance Question Informer Prediction," in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IEEE IRI 2006)* Waikoloa, Hawaii, USA, 2006, pp. 414-419.

[4] O. Feiguina and B. a. K´egl, "Learning to Classify Questions," in *CLiNE (Computational Linguistics in the North-East)*, 2005.

[5] V. Krishnan, S. Das, and S. Chakrabarti, "Enhanced Answer Type Inference from Questions using Sequential Models," in *Proceedings of HLT/EMNLP* Vancouver, British Columbia, Canada, 2005, pp. 315-322.

[6] T. Kudo, "CRF++: Yet Another CRF toolkit." vol. 2006, 2006.

[7] K.-L. Kwok and P. Deng, "Chinese Question-Answering: Comparing Monolingual with English-Chinese Cross-Lingual Results," in *Proceedings of the Asia Information Retrieval Symposium 2006 (AIRS 2006)*, 2006, pp. 244-257.

[8] C.-W. Lee, C.-W. Shih, M.-Y. Day, T.-H. Tsai, T.-J. Jiang, C.-W. Wu, C.-L. Sung, Y.-R. Chen, S.-H. Wu, and W.-L. Hsu, "ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA," in *Proceedings of NTCIR-5 Workshop* Tokyo, Japan, 2005, pp. 202-208.

[9] C.-W. Lee, M.-Y. Day, C.-L. Sung, Y.-H. Lee, T.-J. Jiang, C.-W. Wu, C.-W. Shih, Y.-R. Chen, and W.-L. Hsu, "Chinese-Chinese and English-Chinese Question Answering with ASQA at NTCIR-6 CLQA," in *Proceedings of NTCIR-6 Workshop Meeting* Tokyo, Japan, 2007, pp. 175-181.

[10] X. Li and D. Roth, "Learning Question Classifiers," in *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, 2002.

[11] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu, and R. Sutcliffe, "Overview of the CLEF 2006 Multilingual Question Answering Track," in *CLEF 2006 Working Notes*, 2006.

[12] J.-J. Mei, Y.-M. Zhu, Y.-Q. Gao, and H.-X. Yin, *TongYiCi CiLin (Chinese Synonym Forest)*: Shanghai Press of Lexicon and Books, 1983.

[13] D. Metzler and W. B. Croft, "Analysis of Statistical Question Classification for Fact-Based Questions," *Information Retrieval*, vol. 8, pp. 481-504, 2005.

[14] L. Plamondon and G. Foster, "Quantum, a French/English Cross-language Question Answering System " in *Proceedings of the Cross-Language Evaluation Forum (CLEF 2003)* Trondheim, Norway, 2003.

[15] J. D. M. Rennie and R. Rifkin, "Improving multiclass text clas-sification with the support vector machine," in *MIT Artificial Intelligence Laboratory Publications, AIM-2001-026.*, 2001.

[16] Y. Sasaki, H.-H. Chen, K.-H. Chen, and C.-J. Lin, "Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1)," in *Proceedings of NTCIR-5 Workshop Meeting* Tokyo, Japan, 2005.

[17] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1-47, 2002.

[18] T. Solorio, M. Pérez-Coutiño, M. Montes-y-Gómez, L. Villaseñor-Pineda, and A. López-López, "Question classification in Spanish and Portuguese," *Lecture Notes in Computer Science*, vol. 3406, pp. 612-619, 2005.

[19] J. Suzuki, H. Taira, Y. Sasaki, and E. Maeda, "Question Classification using HDAG Kernel," in *Workshop on Multilingual Summarization and Question Answering 2003 (post-conference workshop in conjunction with ACL-2003)*, 2003, pp. 61-68.

[20] E. M. Voorhees and H. T. Dang, "Overview of the TREC 2005 Question Answering Track," in *Prodeedings of The Fourteenth Text REtrieval Conference (TREC 2005)*, 2006.

[21] D. Zhang and W. S. Lee, "Question classification using support vector machines," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* Toronto, Canada, 2003, pp. 26-32.