An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Classification

Min-Yuh DAY [1,2], Cheng-Wei LEE [1], Shih-Hung WU [3], Chorng-Shyong ONG [2], Wen-Lian HSU [1]
[1] Institute of Information Science, Academia Sinica, Nankang, Taipei
[2] Department of Information Management, National Taiwan University, Taipei
[3] Dept. of Computer Science and Information Engineering, Chaoyang Univ. of Technology, Taichung
Email: {myday, aska, hsu}@iis.sinica.edu.tw, shwu@cyut.edu.tw, ongcs@im.ntu.edu.tw

***Abstract*-Question classification plays an important role in question-answering systems. Chinese question classification is the process that analyzes a question and labels it based on its question type and expected answer type. In this paper, we propose an integrated knowledge-based and machine learning approach for Chinese question classification that focuses on factoid question answering. We develop a Chinese question classification scheme for CLQA C-C (Cross Language Question Answering Chinese to Chinese) factoid question answering, and define a coarse-grained and fine-grained classification taxonomy for a Chinese question-answering system. We adopt INFOMAP inference engine to support the knowledge-based approach for Chinese questions, which can be formulated as templates and use SVM (Support Vector Machines) as the machine learning approach for large collections of labeled Chinese questions. Our experimental results show that the accuracy of Chinese question classification using INFOMAP alone is 88%, and 73.5% with SVM alone. In contrast, classification based on a hybrid approach that incorporates SVM and INFOMAP yields an accuracy rate of 92%.**

I. INTRODUCTION

Question classification plays an important role in automated question-answering systems, such as those created for the TREC (Text Retrieval Conference) question answering task and NTCIR CLQA (Cross Language Question Answering). In general question answering systems, 36.4% of the errors occur in the question classification module for the derivation of the expected answer type [10].

The goal of question classification is to accurately classify a question into a question type and then map it to an expected answer type (question type determination) [9]. For example, Chinese question classification for "奧運的發源地在哪裡？ (Where is the originating place of the Olympics?)" (question) is "Q_LOCATION| 地 " (question type). Question types derived from question classification can be used for answer extraction and answer filtering to improve the accuracy of the overall question answering system.

Approaches to question classification (QC) can be considered in two broad classes, namely, rule-based and statistical [1, 4, 5, 8, 9, 10, 12, 14, 15, 17]. In a rule-based approach, a domain expert produces a number of regular expressions and keywords. In a statistical (probabilistic) approach, on the other hand, the expert knowledge is replaced by a sufficiently large collection of labeled questions and a model is trained with the hope that the useful patterns for classification with the automatically capture. The statistical approach no only has the advantage of saving on expensive expert labor, but also has easier portability to other domains.

In this paper, we propose an integrated knowledge-based and machine learning approach for Chinese question classification (CQC) with the focus on factoid question-answering in Chinese. CQC is the process that analyzes a question and labels the question based on its question type and expected answer type.

We adopt INFOMAP [6, 16] as the knowledge-based approach for Chinese questions, which can be formulated as templates and use SVM (Support Vector Machines) as the machine learning approach for a large collection of labeled Chinese questions.

The remainder of this paper is organized as follows. Section 2 describes the proposed approach. Section 3 discusses the experimental results. In Section 4, we compare our approach with related works. Finally, in Section 5, we present our conclusions and the directions for future research.

II. PROPOSED APPROACH

Question classification is the process of categorizing questions into semantic question types. Here, only CLQA-style open domain factoid questions are considered.

A. Chinese Question Taxonomy

The first step in Chinese Question Classification (CQC) is to design a taxonomy of question types [5, 8]. A taxonomy can have a flat or hierarchical, and it can comprise a small (10-30) or large number of categories (above 50). After analyzing 500 questions taken from the development data set of CLQA provided by NTCIT and thousands of questions in the TREC question corpus, we developed a hierarchical taxonomy of CQC for CLQA.

We have developed a hierarchical classification for Chinese questions using a two-layer hierarchical classification scheme (Question Type; QType) that corresponds to fine-grained named entities or expected answer type (EAT). A fine-grained question type facilitates our question answering-system by extracting an answer based on the corresponding answer type of the named entity from the candidate sentences.

Having developed a hierarchical classification for Chinese questions, we define a coarse-grained and fine-grained classification taxonomy for a Chinese question-answering system. The proposed hierarchical classification taxonomy is inspired by the two-layered question taxonomy proposed in Li and Roth [8], which contains 6 coarse-grained and 50 fine-grained categories.

Our proposed taxonomy for CQC includes 6 coarse-grained classes (Q_PERSON|人, Q_LOCATION|地, Q_ORGANIZATION|組織, Q_ARTIFACT|物, Q_TIME|時間 and Q_NUMBER|數值) and 62 fine grained-class, as shown in Table 1. Each coarse-grained category contains a non-overlapping set of fine-grained categories.

B. Question Type Filter for Expected Answer Type (EAT)

We have also developed a mapping method to obtain the expected answer type (EAT) from the question type (QType) classified by CQC. Table 2 shows the partial question type filter for the expected answer type.

The default EAT is denoted with an "*" in the beginning of the QType, which indicates that EAT focuses on the category and its sub categories, while EAT without an "*" indicates the coarse category is also a candidate EAT. For example, the EAT for fine-grained QTpye "Q_LOCATION_CITY|城市" is "LOCATION_CITY|城市", which indicate that the fine grained EAT and its coarse-grained EAT "Q_LOCATION|地" are both candidate EATs.

C. INFOMAP (Knowledge-based Approach)

We adopt INFOMAP as the knowledge-based approach for CQC. INFOMAP is a knowledge representation framework that extracts important concepts from a natural language text [6, 16]. A powerful feature of INFOMAP is its capability to represent and match complicated template structures, such as hierarchical matching, regular expressions, semantic template matching, frame (non-linear relations) matching, and graph matching. Using INFOMAP, we can identify the question category from a Chinese question.

Figure 1 shows the knowledge representation for CQC in INFOMAP, which is described below. The two-layer question taxonomy is created in a hierarchical format. For example, the fine-grained category "Q_LOCATION_CITY|城市" is a sub node of the coarse-grained category "Q_LOCATION|地" represented in INFOMAP. There are two function nodes, "HAS-PART" and "Rule", in each coarse-grained and fine-grained concept node.

For example, the knowledge representation of the question "2004 年奧運在哪一個城市舉行?(In which city were the Olympics held in 2004?)" in INFOMAP can be formulated as a rule or template like "[5 Time]:[3 Organization]:[7 Q_Location]: ([9 LocaitonRelatedEvent])". We create the rule in the "Rule" function node in INFOMAP. There are four elements (denoted as "HAS-PART") in this rule. Thus, "2004 年 (Year 2004)" is an instance of "Time", "奧運 (the Olympics)" is an instance of "Organization", "在哪一個城市

TABLE I

TAXONOMY OF CHINESE QUESTION CLASSIFICATION (CQC) FOR CLQA

Coarse-grained (6)	Fine-grained (62)
Q_PERSON|人	Q_PERSON_APPELLATION|稱謂
	Q_PERSON_DISCOVERERS|發現者
	Q_PERSON_FIRSTPERSON|第一人
	Q_PERSON_INVENTORS|發明者
	Q_PERSON_OTHER|人其他類
	Q_PERSON_PERSON|人名
	Q_PERSON_POSITIONS|職位
Q_LOCATION|地	Q_LOCATION_ADDRESS|地址
	Q_LOCATION_CITY|城市
	Q_LOCATION_CONTINENT|大陸、大洲
	Q_LOCATION_COUNTRY|國家
	Q_LOCATION_ISLAND|島嶼
	Q_LOCATION_LAKE|湖泊
	Q_LOCATION_MOUNTAIN|山、山脈
	Q_LOCATION_OCEAN|大洋
	Q_LOCATION_OTHER|地其他類
	Q_LOCATION_PLANET|星球
	Q_LOCATION_PROVINCE|省
	Q_LOCATION_RIVER|河流
Q_ORGANIZATION|組織	Q_ORGANIZATION_BANK|中央銀行
	Q_ORGANIZATION_COMPANY|公司
	Q_ORGANIZATION_OTHER|組織其他類
	Q_ORGANIZATION_POLITICALSYSTEM|政治體系
	Q_ORGANIZATION_SPORTTEAM|運動隊伍
	Q_ORGANIZATION_UNIVERSITY|大學
Q_ARTIFACT|物	Q_ARTIFACT_COLOR|顏色
	Q_ARTIFACT_CURRENCY|貨幣
	Q_ARTIFACT_ENTERTAINMENT|娛樂
	Q_ARTIFACT_FOOD|食物
	Q_ARTIFACT_INSTRUMENT|工具
	Q_ARTIFACT_LANGUAGE|語言
	Q_ARTIFACT_OTHER|物其他類
	Q_ARTIFACT_PLANT|植物
	Q_ARTIFACT_PRODUCT|產品
	Q_ARTIFACT_SUBSTANCE|物質
	Q_ARTIFACT_VEHICLE|交通工具
	Q_ARTIFACT_ANIMAL|動物
	Q_ARTIFACT_AFFAIR|事件
	Q_ARTIFACT_DISEASE|疾病
	Q_ARTIFACT_PRESS|書報雜誌
	Q_ARTIFACT_RELIGION||宗教
Q_TIME|時間	Q_TIME_DATE|日期
	Q_TIME_DAY|日
	Q_TIME_MONTH|月
	Q_TIME_OTHER|時間其他類
	Q_TIME_RANGE|時間範圍
	Q_TIME_TIME|時間
	Q_TIME_YEAR|年
Q_NUMBER|數值	Q_NUMBER_AGE|年齡
	Q_NUMBER_AREA|面積
	Q_NUMBER_COUNT|數字
	Q_NUMBER_LENGTH|長度
	Q_NUMBER_FREQUENCY|頻率
	Q_NUMBER_MONEY|金額
	Q_NUMBER_ORDER|序數
	Q_NUMBER_OTHER|數值其他類
	Q_NUMBER_PERCENT|比例
	Q_NUMBER_PHONENUMBER|電話號碼、郵遞區號
	Q_NUMBER_RANGE|數字範圍
	Q_NUMBER_SPEED|速度
	Q_NUMBER_TEMPERATURE|溫度
	Q_NUMBER_WEIGHT|重量

TABLE II

PARTIAL QUESTION TYPE (QTYPE) FILTER FOR EXPECTED ANSWER TYPE (EAT)

Q_TYPE	Filter (EAT)
Q_PERSON|人	*PERSON|人
Q_LOCATION|地	"*LOCATION|地,*ORGANIZATION|組織"
Q_LOCATION_ADDRESS|地址	*LOCATION_ADDRESS|地址
Q_LOCATION_CITY|城市	LOCATION_CITY|城市
Q_LOCATION_CONTINENT| 大陸、大洲	*LOCATION_CONTINENT| 大陸、大洲
Q_LOCATION_COUNTRY|國家	*LOCATION_COUNTRY|國家
Q_LOCATION_ISLAND|島嶼	LOCATION_ISLAND|島嶼
Q_LOCATION_LAKE|湖泊	LOCATION_LAKE|湖泊
Q_LOCATION_MOUNTAIN|山、山脈	LOCATION_MOUNTAIN|山、山脈
Q_LOCATION_OCEAN|大洋	LOCATION_OCEAN|大洋
Q_LOCATION_PLANET|星球	LOCATION_PLANET|星球
Q_LOCATION_PROVINCE|省	LOCATION_PROVINCE|省
Q_LOCATION_RIVER|河流	LOCATION_RIVER|河流
Q_ORGANIZATION|組織	*ORGANIZATION|組織
Q_ORGANIZATION_BANK|中央銀行	ORGANIZATION_BANK| 中央銀行
Q_ORGANIZATION_COMPANY|公司	ORGANIZATION_COMPANY| 公司
Q_ORGANIZATION_POLITICALSYSTEM|政治體系	ORGANIZATION_POLITICALSYSTEM|政治體系
Q_ORGANIZATION_SPORTTEAM|運動隊伍	ORGANIZATION_SPORTTEAM|運動隊伍
Q_ORGANIZATION_UNIVERSITY|大學	ORGANIZATION_UNIVERSITY|大學
Q_ARTIFACT|物	ARTIFACT|物
Q_ARTIFACT_FOOD|食物	ARTIFACT_FOOD|食物
Q_ARTIFACT_INSTRUMENT|工具	ARTIFACT_INSTRUMENT|工具
Q_ARTIFACT_LANGUAGE|語言	ARTIFACT_LANGUAGE|語言
Q_ARTIFACT_PLANT|植物	ARTIFACT_PLANT|植物
Q_ARTIFACT_PRODUCT|產品	ARTIFACT_PRODUCT|產品
Q_ARTIFACT_SUBSTANCE|物質	ARTIFACT_SUBSTANCE|物質
Q_ARTIFACT_VEHICLE|交通工具	ARTIFACT_VEHICLE|交通工具
Q_ARTIFACT_ANIMAL|動物	ARTIFACT_ANIMAL|動物
Q_ARTIFACT_AFFAIR|事件	ARTIFACT_AFFAIR|事件
Q_ARTIFACT_DISEASE|疾病	ARTIFACT_DISEASE|疾病
Q_ARTIFACT_PRESS|書報雜誌	ARTIFACT_PRESS|書報雜誌
Q_ARTIFACT_RELIGION|宗教	ARTIFACT_RELIGION|宗教
Q_TIME|時間	*TIME|時間
Q_NUMBER|數值	*NUMBER|數值

(in which city)" is an instance of "Location", "舉行(is held)" is an instance of "LocationRelatedEvent".

After completing the knowledge representation of CQC in INFOMAP, we use the INFOMAP engine to process the Chinese question and determine the question type. For example, after the INFOMAP engine processes the question "2004 年奧運在哪一個城市舉行? (In which city were the Olympics held in 2004?)" it obtains the corresponding concept node fine-grained category "Q_LOCATION_CITY|

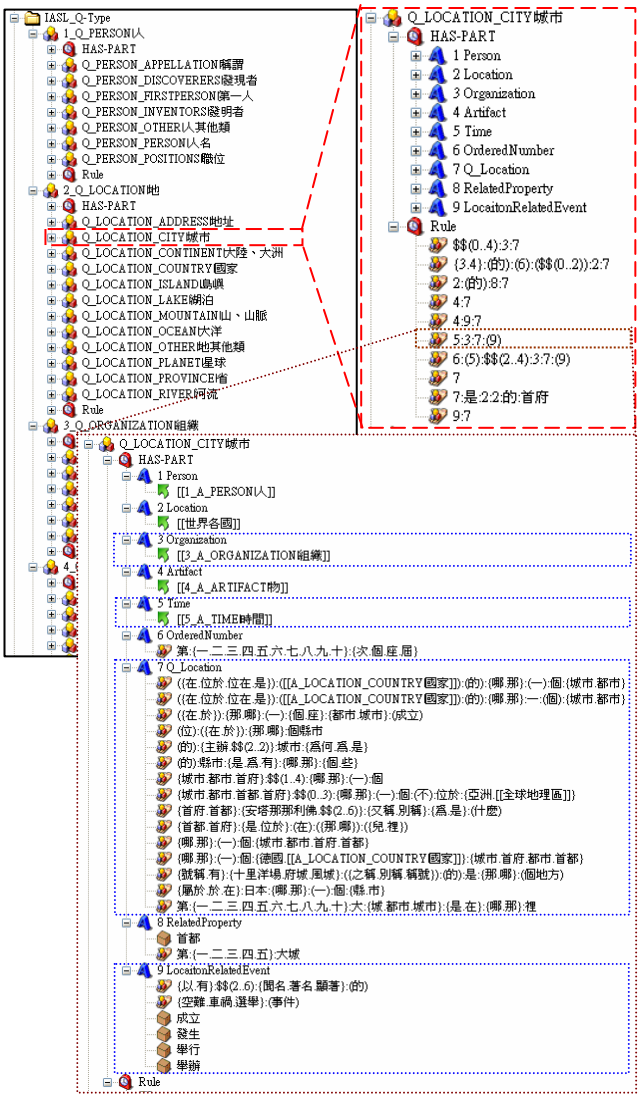

Fig.1. Knowledge representation for CQC in INFOMAP

城市", which indicates the question type. We use the mapping table presented in Table 2 to get the expected answer type from the question type classified by CQC.

D. SVM (Machine Learning Approach)

We adopt SVM as the machine learning approach for CQC, motivated by the fact that they have consistently outperformed other machine learning techniques in several tasks, including text classification [11, 13] and question classification [5, 8, 9, 15, 17].

For the implementation of the machine learning approach, we use SVMlight [7] for CQC. SVMlight is an implementation of Vapnik's Support Vector Machine for pattern recognition.

The two types of feature used in CQC are syntactic features and semantic features, which we describe below.

1. Syntactic features

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

- Bag-of-Words

In bag-of-words features, we use character-based and word-based bigram features, for example, character-based bigram (CB) and word-based bigram (WB) features.

- Part-of-Speech (POS)

We use AUTOTAG [2], a POS tagger developed by CKIP, Academia Sinica, to get the POS of given Chinese questions, and then use the POS features for CQC.

2. Semantic Features

- HowNet Senses

We use "HowNet 2000" [3] to get the semantic features of given Chinese questions. There are two semantic features used in our SVM Model, namely, HowNet Main Definition (HNMD) and HowNet Definition (HND).

E. Integration of SVM and INFOMAP (Hybrid Approach)

In our integrated CQC, each question is classified into a question type(s) by the INFOMAP and SVM module. The integrated module selects the question type with the highest confidence score from the INFOMAP or the SVM model as follows.

1. If the question matches the templates or rules represented in INFOMAP and obtains the question type, we use the question type obtain from INFOMAP first.
2. If no question type can be obtained from INFOMAP, we use the result from the SVM model.
3. If multiple question types are obtained from INFOMAP, we choose the one obtained from SVM first.
4. If one question type with a high positive score is obtained from SVM and one question type obtained from INFOMAP, which is not the same as the one from SVM, we choose the one from SVM with a high positive score.

III. EXPERIMENTAL RESULTS AND DISCUSSION

A. The Datasets

We use the CLQA's development dataset and formal run test dataset, respectively, for our training and testing of CQC. There are 300 questions for Japanese news and 200 for traditional Chinese news in development dataset. We manually build another 850 questions (518 in SVM and 332 questions in INFOMAP) for our proposed question taxonomy to train our CQC model. Note that we use different features to train the SVM model based on a total of 1350 questions and their labeled question type.

B. The Experimental Results

The experimental results of integrating knowledge-based and machine learning approach for CQC are presented below.

Table 3 shows the experimental results of the machine learning approach for CQC. In this experiment, the training and testing data is taken from the development dataset provided by CLQA. Specifically, the training data is taken from CLQAS300 (300 questions for Japanese news), and the testing data is taken from CLQAS200 (200 questions for Chinese news).

Figure 2 shows the experimental results of different features used in SVM. The experimental results show that the character-based bigram (CB) and HowNet main definition (HNMD) features and the combination of CB and HNMD achieve better performance than the other features, which are also shown in the figure.

For example, the CQC for the question "請問首位自費太空旅行的觀光客爲誰？(Who was the first self-financed space tourist?)" is "Q_PERSON_FIRSTPERSON|第一人", which is the top question type with the highest score obtained from SVM model.

We use the 1350 questions as our training dataset and the CB+HNMD features to train our SVM model for the testing dataset, which is taken from CLQA's formal run of 200 questions. The experimental results are as follows.

The accuracy of CQC using INFOMAP solely is 88% (176/200). There are 181 (90.5%) questions with answers from INFOMAP. The accuracy is 97% for the 181 questions with answer from INFOMAP. The accuracy of CQC with SVM (where the question is taken from those not answered by INFOMAP) is 42% (8/19). In addition, the accuracy of the CQC using SVM solely is 73.5% (147/200). However, it is significant that by integrating the SVM and INFOMAP (Hybrid Approach), the accuracy rate increases to 92% (184/200). The experimental results of the hybrid approach for CQC are shown in Figure 3.

C. Discussion

Our experimental results indicate that the integrated approach performs better than the individual knowledge-based or machine learning approach. In addition, we want to assess the performance of the knowledge-based and machine learning approaches for different level of questions (in terms of easy and hard questions).

Easy questions are defined as follows:

1. There are clear words that show the question type and indicate the words that are not question types. (e.g., "誰(Who)", "哪一位(Which person)", "首位(the first person)")
2. There are explicit words that identify the question type. If words are easy to identify, it means they overlap with a question type (For example, "隊伍(team)" and "運動隊伍(sports team)")
3. There are interrogative words that connect with question type words in question. (For example, "那個人(Which Person)").

Below are some examples of easy question in CQC.

- 奧運的發源地在哪裡？(Where is the originating place of the Olympics?) Q_LOCATION|地
- 羅浮宮在哪個城市？(Which country is the Louvre Museum located in?) Q_LOCATION_CITY|城市
- 請問紐西蘭的首都是？(What is the capital of New Zealand?) Q_LOCATION_CITY|城市

TABLE III

EXPERIMENTAL RESULTS OF THE MACHINE LEARNING APPROACH FOR CQC

Feature Used:	Top 1 Accuracy (Fine):	Top 1 Accuracy (Coarse):	Top 5 MRR (Fine):	Top 5 MRR (Coarse):
CB	48.00%	66.00%	0.5257	0.7430
CB_HNMD	46.50%	65.00%	0.5108	0.7386
HNMD	45.00%	59.50%	0.4874	0.6949
CB_HND	45.00%	62.50%	0.4958	0.7127
CB_HNMD_HND	44.50%	58.50%	0.4949	0.6953
HND	43.00%	54.00%	0.4748	0.6578
WB	42.00%	53.50%	0.4692	0.6496
WB_POS	41.00%	58.50%	0.4663	0.6688
POS	34.50%	50.00%	0.4193	0.6018

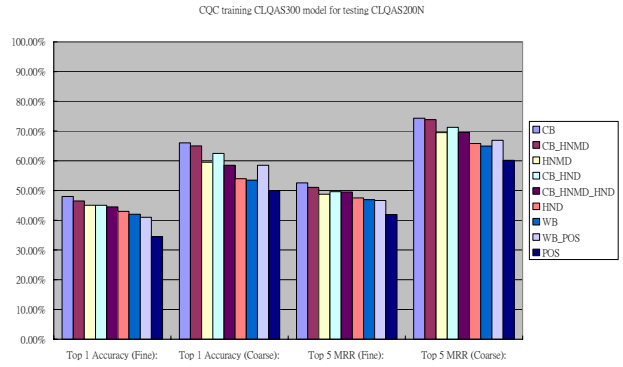

Fig.2. Experimental results of the different features used in SVM

We have found that the knowledge-based approach performs well with easy questions using the templates and rules.

We now analyze the performance of the hierarchical coarse-grained and fine-grained question taxonomy in the integrated model. A total of 36% (68/200) of the questions can be classified into fine-grained categories, and 43% can be classified into the PERSON (Q_PERSON|人) coarse-grained category. In contrast, 63% of the questions that are not in the PERSON (Q_PERSON|人) coarse-grained category can be classified into fine-grained categories.

IV. RELATED WORKS

Extensive works on question classification have been reported in the literature [1, 4, 5, 8, 9, 10, 12, 14, 15, 17]. Early works suggested various standards for question classifications. Recent studies show that semantically classifying questions achieve a better result for factoid question-answering than conceptual categories [8]. Although multi-layered taxonomies have been proposed in the literature, most question classification studies are based on machine learning approaches.

Li and Roth [8] proposed 6 coarse classes and 50 fine classes for TREC factoid question answering. They use the Sparse Network of Windows (SNoW) with over 90% accuracy.

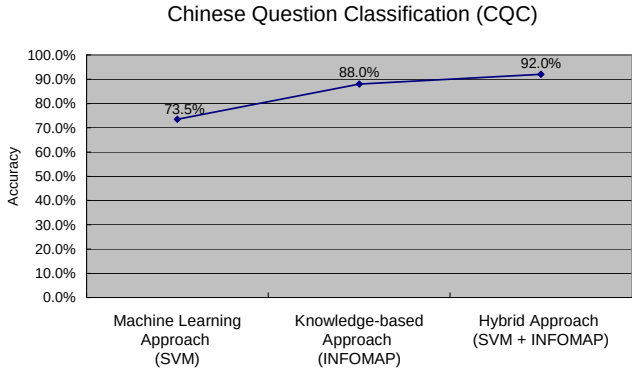

Fig.3. Experimental results of the hybrid approach for CQC

Zhang and Lee [17] used Support Vector Machines (SVMs) with only surface text features (bag-of-words and bag-of-ngrams) and derive coarse-grained categories with 86% accuracy and fine-grained categories with approximately 80% accuracy. By adding syntactic information which including sub-trees of the parse tree that has at least one terminal symbol or one production rule, they derive an accuracy of 90% for coarse-grained classes. There are no results for fine-grained classes.

Suzuki et al [15] use a hierarchical SVM to experiment on four feature sets: (1) words only; (2) words and named entities; (3) words and semantic information; and (4) words and NEs and semantic information. They measured a question type hierarchy at different depths and achieved accuracy rate ranging from 95% at depth 1 to 75% at depth 4.

In contract to our proposed approach for question classification in Chinese, the accuracy of CQC using INFOMAP solely is 88% and SVM solely is 73.5%, however, it is significant that by integrating the SVM and INFOMAP (Hybrid Approach) yields an accuracy rate of 92%.

V. CONCLUSIONS

We have proposed a hybrid approach to Chinese question classification (CQC) for CLQA factoid question-answering with hierarchical coarse-grained and fine-grained question taxonomies.

We have also developed a hierarchical classification containing 6 coarse-grained categories and 62 fine-grained categories for Chinese questions. Our proposed Chinese question type categories are designed for CLQA factoid questions. Furthermore, we propose a mapping method for question type filtering to obtain expected answer types (EAT), which can used to restrict the answers to factoid questions.

The most significant contribution of this paper is the integrated knowledge-based and machine learning approach, which achieves significantly better accuracy rate than individual approaches. Specifically, the accuracy of CQC that incorporates SVM and INFOMAP (our hybrid approach) is 92%, compared to 88% for INFOMAP solely and 73.5% for SVM solely.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science Council under GRANT NSC93-2752-E-001-001.

REFERENCES

[1] Z. Cheung, K. L. Phan, A. Mahidadia, and A. Hoffmann, "Feature Extraction for Learning to Classify Questions", Proceedings of Advances in Artificial Intelligence (AI 2004), 2004. pp. 1069-1075.

[2] CKIP, "CKIP AutoTag", http://ckip.iis.sinica.edu.tw/CKIP/, 2005

[3] Z. Dong and Q. Dong, "HowNet", http://www.keenage.com/, 2000

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

[5] K. Hacioglu and W. Ward, "Question Classification with Support Vector Machines and Error Correcting Codes", NAACL 2003, 2003. pp. 28-30.

[6] W.-L. Hsu, S.-H. Wu, and Y.-S. Chen, "Event Identification Based on the Information Map - INFOMAP", NLPKE 2001, Tucson Arizona, USA., 2001.

[7] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", Proceedings of the European Conference on Machine Learning, 1998.

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

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

[10] D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu, "Performance issues and error analysis in an open-domain question answering system", *ACM Transactions on Information Systems*, vol. 21, pp. 133-154, 2003.

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

[12] D. Roth, C. Cumby, X. Li, P. Morie, R. Nagarajan, N. Rizzolo, K. Small, and W.-T. Yih, "Question-Answering via Enhanced Understanding of Questions", TREC 2002, 2002.

[13] F. Sebastiani, "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, vol. 34, pp. 1-47, 2002.

[14] 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.

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

[16] S.-H. Wu, M.-Y. Day, and W.-L. Hsu, "FAQ-centered Organizational Memory", Proceeding of the Knowledge Management and Organizational Memory workshop on the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), 2001. pp. 112-120.

[17] D. Zhang and W. S. Lee, "Question Classification Using Support Vector Machines", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003. pp. 26-32.