A Template Alignment Algorithm for Question Classification

Cheng-Lung Sung, Min-Yuh Day, Hsu-Chun Yen and Wen-Lian Hsu, *Fellow, IEEE*

***Abstract*—Question classification (QC) plays a key role in automated question answering (QA) systems. In Chinese QC, for example, a question is analyzed and then labeled with the question type it belongs to and the expected answer type. In this paper, we propose a novel method of Chinese QC that integrates syntactic tags and semantic tags into an alignment-based approach. We adopt a template alignment (TA) algorithm to process large collections of Chinese questions and compare the classification results with those of INFOMAP, a human annotated knowledge inference engine for Chinese questions. We experimented with two approaches for the proposed system: a majority algorithm and a machine learning method that uses Support Vector Machine (SVM). The TA algorithm performs well with both approaches. The experimental results show that the accuracy achieved by TA (85.5%) is comparable to that of INFOMAP (88%). In contrast, QC based on the SVM approach, which incorporates syntactic features and TA yields an accuracy rate of 91.5%.**

***Index Terms*—Alignment, Surface Pattern, Document Classification, Text Mining**

I. Introduction

Question classification (QC) plays a key role in automated question answering (QA) systems, such as those created for the TREC (Text Retrieval Conference) QA task and NTCIR Cross Language Question Answering (CLQA)[5].

The goal of QC is to accurately classify a question as a particular question type (*qtype*) and then map it to an expected answer type (*qtype* determination). For example, Chinese QC for the question "奧運的發源地在哪裡？(Where were the Olympics first held?)" (question) is "Q_LOCATION|地" (question type). Question types derived by QC can be used for answer extraction and answer filtering to improve the overall accuracy of the QA system.

Approaches to QC can be classified into two broad categories; rule-based methods and statistical (probabilistic) methods [3, 4, 6]. In the former, a domain expert produces a number of regular expressions and keywords. In the latter, the expert knowledge is replaced by a sufficiently large collection of labeled questions and a model is trained to capture useful patterns for classification automatically. The statistical approach has the dual advantage of saving on expensive expert labor and easy portability between domains.

Surface patterns are syntactic patterns that connect answers and question keywords. Our proposed surface pattern method, Template Alignment (TA) algorithm, can efficiently help QC. TA generates useful templates from common QC patterns regardless of the question type. The proposed method, which is based on sequence alignment, takes advantage of both rule-based and statistical methods. In our method, by aligning question sentences using dynamic programming, similar (syntactic and semantic) parts of the question sentences can be extracted as templates.

II. Template Alignment (TA)

We introduce the alignment algorithm and describe the template generation process, which involves the following steps: 1) generate templates by sequence alignment; 2) select templates based on a filtering process; 3) apply a matching algorithm for determine question's category; and 4) experiment with both heuristic and machine learning methods.

A. Alignment and Template Generation

Sequence alignment is the process that finds similar sequences in a pair of sentences. Pair-wise sequence alignment (PSA) algorithms that generate templates and match them against new text have been researched extensively. Because we need surface patterns extracted from sentences that have certain morphological similarities, we employ local alignment techniques to generate surface patterns.

We incorporate a Chinese segmentation tool, AutoTag , to break a question into segments comprised of words and POS tags. We use an NER engine to tag NE words, and the remaining words are tagged "O (NE tag for others)". Thus, every segment contains a syntactic tag (POS) and a semantic tag (NE) that corresponds to a word. For example, "２０００年奧運在雪梨舉行" would be segmented as "２０００年/Nd/TIME 奧運/Nb/O 在/P/O 雪梨/Ncd/O 舉行/VC/O".

Using the proposed alignment algorithm, our TA algorithm extracts general patterns of all three types of tags. We begin by pairing all sentences based on their similarity. Closely matched

C. L. Sung is with the Electrical Engineering Department, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617 Taiwan (corresponding author; e-mail: clsung@ iis.sinica.edu.tw).

M. Y. Day is with the Institute of Information Science, Academia Sinica, Taipei 115 Taiwan (e-mail: myday@iis.sinica.edu.tw).

H. C. Yen is with the Electrical Engineering Department, National Taiwan University, Taipei, 10617 Taiwan (e-mail: yen@cc.ee.ntu.edu.tw).

W. L. Hsu is with the Institute of Information Science, Academia Sinica, Taipei 115 Taiwan (e-mail: hsu@iis.sinica.edu.tw).

ISI 2008, June 17-20, 2008, Taipei, Taiwan
1-4244-2415-3/08/$20.00 ©2008 IEEE

pairs are then aligned and a pattern that fits both pairs is created. A template is composed of ordered slots, which are chosen according to the corresponding parts of the aligned sentence pair with the following priority: word > NE tag > POS tag. If the sentences for a given slot have nothing in common, the algorithm creates a gap (“—”) in that position.

Because sentence alignment is time consuming, we only use the set of training questions, which consists of 550 NTCIR (5 and 6) CLQA questions plus 465 manually created questions. For each training question, we retrieve the top 200 most relevant passages tagged with NEs and POS tags. We applied the generation algorithm to the 1,015 training questions and generated 4,896 patterns. As we wanted patterns that could connect question keywords and the question category, we designed a filtering process to gather appropriate patterns.

B. Template Filtering

Since a question could match different templates, TA must be capable of choosing the appropriate category. The question we must consider before applying templates is how to select the appropriate templates. A template is called *noisy* if it could be matched to several *qtypes*. Thus, a method for filtering out noisy templates is required.

We use two criteria for filtering out the extracted templates. First, we test the matching confidence *Confidence(t,qtype)*, of each template t in the development set with each question type, *qtype*, using the following formula:

$$Confidence(t, qtype) = \frac{\#\text{ of matched questions in qtype}}{\#\text{ of all matched questions}}$$

For example, if the matching result of template t is {Q_LOC, 8; Q_PER, 1; Q_ORG 1}, then *Confidence(t,qtype)* will become {0.8,Q_LOC; 0.1,Q_PER; 0.1,Q_ORG}. We also define *Score(t)* and *Qtype(t)* for retrieving the score of template (i.e., highest confidence) with its corresponding *qtype*:

$$Score(t) = \underset{qtype}{Max}\big(Confidence(t, qtype)\big)$$

$$Qtype(t) = \underset{qtype}{\arg\max}\big(Confidence(t, qtype)\big)$$

For a *qtype* to be dominant, its confidence must be greater than 0.5 because a majority should larger than the sum of other confidence of *qtypes*. Our system discards templates whose score is lower than a threshold τ.

C. Template Matching algorithm

To evaluate and apply the templates generated by the filtering step, we develop a template matching algorithm for QC.

We transform our templates to regular expressions. We treat a gap “-” in a template as a wildcard. For raw words or NE tags in a template, the matched segment must have exactly the same word or NE tag, respectively. For POS matching, the POS tag in a template must be a prefix of the matched segment. Table I shows an example of matched question segments.

In the training phase, all templates are processed with every categorized question to calculate their confidence. Note that the templates are also categorized. In the development and testing phases, we first calculate the *score set* ST for uncategorized questions according to Algorithm 1. Then the *score set* ST determines each question’s category based on two different approaches: a majority algorithm and a machine learning algorithm using SVM. The majority algorithm is based on the assumption that questions of the same *qtype* have similar patterns. As a result, a *qtype* qt of a specific question qsi is selected by the algorithm when the aggregate score of the corresponding categorized templates is the largest of the *qtype* set. We describe the SVM approach latter.

Algorithm 1 Question Classification

Input: Uncategorized question set $QS = \{qs_1, \ldots, qs_l\}$,
Categorized template set $CT = \{ct_1, \ldots, ct_m\}$,
Output: Determined category Score Set $ST = \{S_1, \ldots, S_l\}$

$ST = \{\}$
for each template qs_i from qs_1 to qs_l **do**
segment and **tag** the question qs_i
  $Si = \{\}$
  for each categorized template set ct_j from ct_1 to ct_m **do**
    $s_j \leftarrow 0$
    for each template t_k from t_1 to t_n in ct_i **do**
      perform ***template matching*** on qs_i
      using template t_k, then
        if template t_k matches qs_i
          $s_j = s_j$ + Score (t_k);
    end;
    Si←Si□s_j;
  end;
  ST←ST□*Si*;
end;
return *ST*;

D. Machine Learning Approach (SVM)

Machine learning approaches applied to text categorization have produced impressive results . To complement template alignment algorithm, we adopt SVM as the machine learning approach for CQC, which has been successfully applied to several QC researches [3, 6].

We use LibSVM which is an integrated software for support vector classification, regression and distribution estimation.

The feature used in QC-SVM is the same as that used for the majority algorithm, i.e., the score set ST. We use five-fold cross validation to tune hyper parameters for our model. In our experiment, we also adopt a method that integrates two useful syntactic features: bag-of-words and part-of-speech.

III. EXPERIMENTAL RESULT

A. Datasets

We use CLQA’s development dataset and formal run test dataset, respectively, as our training and test datasets for CQC. In NTCIR CLQA task, there are 200 questions for Japanese news and 350 for traditional Chinese news in development dataset. We manually build another 465 questions for the proposed question taxonomy to train our TA model. The size of

TABLE I:
AN EXAMPLE OF QUESTION TEMPLATE MATCHING, T, R, AND Q DENOTE TEMPLATE, REGULAR EXPRESSION, AND QUESTION.

Template	V	—	N	Na	的	LOC	Na	是	誰
Regex	[^]+/V[^]+	([^]+)?	[^]+/N[^]+	[^]+/Na[^]+	的/[^]+	[^]+/LOC	[^]+/Na[^]+	是/[^]+	誰/[^]+
Question	榮獲/VJ/O	*nil*	諾貝爾 /Nb/ORG	和平獎 /Na/O	的/DE/O	南韓 /Nc/LOC	總統 /Na/OCC	是/SHI/O	誰/Nh/O

template for test dataset is 400. In the majority approach we employ ten-fold cross validation training our TA model to get the best threshold (τ=0.6). In the SVM approach we use five-fold cross validation to train the SVM model. Our question taxonomy for CQC consists of 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 classes. Each coarse-grained category contains a non-overlapping set of fine-grained categories.

B. Experimental Results

We compare the proposed method with a knowledge-based approach called INFOMAP [2] for Chinese QC (CQC), which focuses on factoid Chinese QA. The system analyzes a question and labels it based on its *qtype* and the expected answer type[1].

Fig. 1 shows the experimental results for a knowledge-based approach using INFOMAP and the TA approach for CQC, with and without the SVM model. The accuracy rates are as follows: 1) INFOMAP alone is 88%, 2) TA using a majority algorithm solely is 85%, and 3) TA using the SVM model solely is 82.5%. In contrast, the proposed integrated TA approach which combines syntactic features, bag-of-words and parts-of-speech, in-creases the accuracy to 91.5%. The integrated approach is comparable to the hybrid approach using INFOMAP and SVM [1], which has 92% accuracy.

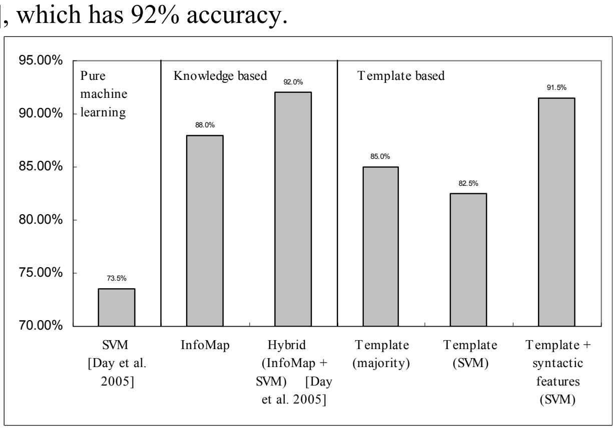

Fig. 1. Fine-grained accuracy of the proposed approaches and a knowledge-based approach.

IV. CONCLUSION

The proposed dynamic programming approach is comparable to the knowledge-based approach (INFOMAP). In our research, the best result comes from the SVM model integrated TA and syntactic features.

The best method is the one that integrates syntactic features with the proposed TA as the features of the SVM model. However, the results of TA solely are comparable to those of a human-annotated knowledge-based approach. (i.e., the classifier is defined by domain experts).

The experiment results of the majority approach and the SVM approach are similar. Specifically, the TA approach performs better than SVM because, like rule-based approaches, it uses template matching, but it may suffer long questions. Long questions are more likely to contain irrelevant parts to be matched by more short templates, which may generate noises. However, most questions can be classified by short templates, for example, "…是誰 (who is..)", "哪一年… (in which year)", and "…在哪裡 (where…)".

The most important contribution of this paper is the template alignment (TA) method, which automatically generates question classification patterns and achieves comparable accuracy rates to INFOMAP Specifically, the accuracy of the proposed approach is 85.5%, compared to 88% for INFOMAP (a knowledge-based approach) solely and 91.5% by integrating bag-of words and POS features into our SVM model.

In our future work, we will improve the scoring function by applying virtual attributes. We will also extend the use of semantic features by incorporating HowNet's senses, and enhance template generalization by taking advantage of semantic sense hierarchies.

ACKNOWLEDGMENT

This research was supported in part by National Science Council under Grant NSC 95-2752-E-001-PAE and the thematic program of Academia Sinica under grant AS95ASIA02.

REFERENCES

[1] 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 *IEEE International Conference on NLPKE*, 2005.

[2] W.-L. Hsu, S.-H. Wu, and Y.-S. Chen, "Event identification based on the information map-INFOMAP," in *IEEE International Conference on Systems, Man, and Cybernetics*, Tucson, AZ, USA, 2001, pp. 1661-1666.

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

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

[5] Y. Sasaki, C.-J. Lin, K.-h. Chen, and H.-H. Chen, "Overview of the NTCIR-6 Cross-Lingual Question Answering Task," in *NTCIR-6 Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*, Tokyo, Japan, 2007, pp. 153-163.

[6] 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.