Learning Based Approaches for Vietnamese Question Classification Using Keywords Extraction from the Web

Dang Hai Tran[1], **Cuong Xuan Chu**[1], **Son Bao Pham**[1], **Minh Le Nguyen**[2]
[1]University of Engineering and Technology, Vietnam National University
[2]Japan Advanced Institute of Science and Technology
[1]{dangth,cuongcx_54,sonpb}@vnu.edu.vn
[2]nguyenml@jaist.ac.jp

Abstract

This paper presents our research on automatic question classification for Vietnamese using machine learning approaches. We have experimented with several machine learning algorithms utilizing two kinds of feature groups: bag-of-words and keywords. Our research focuses on two most important tasks which are corpus building and features extraction by crawling data from the Web to build a keyword corpus. The performance of our approach is promising where our system's precision outperforms the state-of-the-art Tree Kernel approach (Collins and Duffy, 2001) on a Vietnamese question corpus.

Keywords

keyword collection, machine learning, Vietnamese question classification corpus.

1 Introduction

Question Classification (QC) is a task that, given a question, maps it to one of the predefined k classes, which indicates a semantic constraint on the sought-after answer (Li and Roth, 2006).

In a question answering system, before finding an answer of a question, the system has to identify which category it is asking about, and this is the obligation of question classification. Then, based on the identified category, we can narrow the space of answers that we have to find. Let us consider some examples:

- **Q:** *"Ai là người phụ nữ đầu tiên hy sinh trong chiến tranh Việt Nam?"* (*"Who was the first woman killed in the Vietnam War?"*), we expect to know that the target of this question is a **person**, thus reducing the number of possible answers significantly.
- **Q:** *Tại sao nắng màu vàng?* (*Why is the sunshine yellow?*) indicates that this question wants to get information about **reason**, thus in next steps our system just concerns about reason answers space rather than human or number categories.

The problem is that if the number of categories is more and the categories are more specific, the question answering system will spend more time classifying questions but it's performance will be better. Let consider another example, in the next two questions, they are both asked to get information about **location**:

- **Q:** *Thành phố nào ở Canada có nhiều dân nhất?* (*What Canadian city has the largest population?*)
- **Q:** *Đất nước nào trao tặng Mỹ tượng nữ thần tự do?* (*Which country gave New York the Statue of Liberty?*)

More particularly, we can see that the target of the first question is a city and the other one is a country, both city and country are locations. In this case, location is considered a coarse-grained class and city and country are fine-grained classes. Naturally, the more fine-grained classes we have, the more difficult it is to tell them apart. For hierarchical categories, we adopted a two-level learning approach: first we solve the problem of classifying questions in the coarse-grained classes then based on this predicted label we continue with the fine-grained categories.

Our main contributions are building a Vietnamese corpus, collecting and using a kind of important features, which is keyword, for Vietnamese question classification. We tackle the corpus building by translating an existing well known English question corpus to Vietnamese. To the best of our knowledge, there is no publicly available Vietnamese questions corpus. As a result of

International Joint Conference on Natural Language Processing, pages 740–746,
Nagoya, Japan, 14-18 October 2013.

this work, we will share our newly built corpus to the research community. Collecting and using keywords is another important contribution of our work, which indicates that keyword extracted from Web is the most effective feature for Vietnamese question classification task. In this paper, we are going to present how we collect and extract keyword features for Vietnamese question classification.

The paper is organized as follows: Section 2 describes the related works and section 3 presents the process we take to build the Vietnamese data corpus. In section 4 we describe our Vietnamese question classification system, especially the features extraction step which involves crawling data from the Internet to create keyword features. In section 5 we show our experiments when using different machine learning algorithms with the set of features we extracted on our data corpus to classify questions in Vietnamese. Finally, section 6 provides some conclusions.

2 Related Works

There are many different approaches to resolve the question classification problem. Zhang and Lee (Zhang and Lee, 2003) with SVM (Cortes and Vapnik, 1995) using Tree Kernel, Li and Roth (Li and Roth, 2006) with SNoW model are two state-of-the-art approaches for English question classification. In Zhang and Lee's approach, the input question for this method is parsed into a syntax tree and converted to a vector in a multi-dimensional space. They introduced a new kernel function for SVM, Tree Kernel, constructed by dynamic programming to derive the similarity between two different syntax trees. This method achieved a precision of 90% with coarse-grained classification on TREC but there isn't any published results with fine-grained classification. In Li and Roth's approach, a set of features they used not only include syntactic features such as chunking but also include semantic features by using WordNet for English and building class-specific related words. Using semantic features, this method achieved a high precision of 92.5% with coarse-grained class and 85% with fine-grained class classification on a set of data including 21500 training questions from TREC 8, 9 (Voorhees, 1999; Voorhees, 2000), USC (Hovy et al., 2001), and 1000 testing questions from TREC 10, 11 (Voorhees, 2001; Voorhees, 2002).

2.1 Question Classification in Vietnamese

Question Classification in English is a classical problem but in Vietnamese, it is still a relatively new problem. To the best of our knowledge, there is not any research which works on open-domain Vietnamese question classification using learning based approach is published. In a research about question answering system for Vietnamese (Tran et al., 2009), the authors used machine learning approach for question classification module, however, the questions are only on travelling domain. Moreover, there is another research also working on question answering in Vietnamese (Dat et al., 2009). This system also focuses on answering questions on a specific domain and the authors used rule-based approach to resolve question classification module.

Combining the strengths of many solutions applied for English and the idea of (Tran et al., 2009), we started our research and experiment in Vietnamese question classification using machine learning approaches. Our main contributions are building a question set and a class-specific related word (keywords) set for Vietnamese.

3 Corpus Building

3.1 Question Hierarchy

From 1999, to support competitive research on question answering, The Text Retrieval Conference (TREC) has launched a QA track (TREC 8). Because the TREC QA track builds a fully automatic open-domain question answering system, there are many researches using TREC as experimental data sets. Importantly, the question type taxonomies of TREC can be used for any language. Besides there is not any standard for Vietnamese question classification yet, so we decide to use the question type taxonomies of TREC in our research.

TREC defined a two-layered taxonomy, which represents a natural semantic classification for typical answers. The hierarchy contains 6 coarse-grained classes (ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION and NUMERIC VALUE) and 50 fine-grained classes. Table 1 shows the hierarchy of these classes in nearly 5500 training and 500 testing questions of TREC 10. Each coarse-grained class contains a non-overlapping set of fine-grained classes.

Coarse	Fine-Grained
ABBR	abbreviation, expansion
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual, title
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

Table 1: The TREC coarse and fine-grained question classes.

3.2 Question Translation

From the English question corpus, we translated them into Vietnamese and use them for Vietnamese question classification based on following rules:

- The content of Vietnamese question must correspond to its label.
- Some named entities can be changed without keeping the semantic meaning. For example, Washington can be changed to be Hà Nội or White House can be changed to Hồ Gươm.
- Given an English question, we can translate it into many Vietnamese questions with the same meaning but different in syntactic structures. As a result, our classifier can detect many kinds of Vietnamese questions.

In this project, we allocated 5 students who have TOEFL score > 500 for translating 6000 TREC English questions into Vietnamese in about 2 months. Every member of our group not only has to translate but also review the translated Vietnamese questions from other members to find mistakes to correct them and assure the quality of the translated data.

With our Vietnamese question classification problem, there is no publicly available Vietnamese corpus, when our corpus is made publicly available, it can be used for many works in the future.

4 Vietnamese Question Classification System (VnQCS)

4.1 Vietnamese Question Classification System Architecture (VnQCS - Architecture)

There are two main components in our system, the Feature Extractor and the Classifier (Figure 1). Source code of our system and the data corpus were made publicly available at: *https://code.google.com/p/vn-qcs/*

The Classifier contains two levels, the coarse-grained classification and the fine-grained one. In the first step, questions are classified into the coarse-grained classes, then taking result of the first step as a feature, we continue classifying questions to the fine-grained grained classes.

With a classification problem using machine learning approaches, the feature extraction is a key step. The quality of the set of features extracted directly impacts the classification precision. The Feature Extractor module consists of two components: Vietnamese Word Segmenter and Keyword Collector. Among them, Keyword Collector plays an important role in feature extraction method and it is the highlight of this research.

4.2 Feature Extraction in VnQCS

First, it is known that the linguistic characteristics of Vietnamese is different from English. Unlike English, in Vietnamese, one word may contain more than one token. For example, *mobile* (English) is translated into *điện thoại* (Vietnamese) and *mobile* is a word but *điện thoại* is a word which includes two tokens (*điện, thoại*). So, to match the characteristics of Vietnamese grammar, we will use words as a feature of algorithm in question classification.

We divide the features we extracted into three groups: bag-of-words, keywords and syntactic trees.

With **bag-of-words**, to get a set of features of Vietnamese words, we use a VnWordSegmenter tool[1] to extract them from Vietnamese training questions. With VnWordSegmenter tool, a question, for instance, "Bốn hình thức tồn tại của vàng là gì? (What four forms does gold occur in ?)", will be segmented into a sequence of words, as "Bốn hình_thức tồn_tại của vàng là_gì?".

[1]Developed tool of iTim Company, website: http://coccoc.com/about/

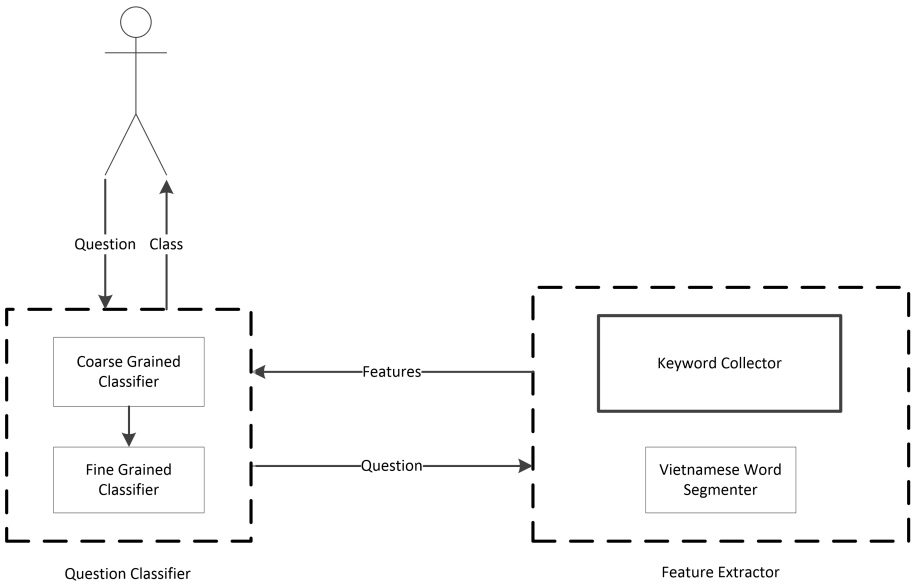

Figure 1: Architecture of VnQCS.

With **keywords**, there are many keywords for a question class. In Li and Roth (Li and Roth, 2006), they used WordNet for English and built a set of class-specific words as semantic features and achieved the high precision (see section 2). But for Vietnamese, there isn't any lexical database like WordNet, so we have to develop an algorithm which collect keywords from the Internet which are lists of class-specific words for a smaller scale and useful enough for question classification (see Section 4.2.1). Moreover, we also manually collect keywords by observing the set of training questions. For examples, in reason class, with a why-question, we usually use "tại sao (why)" to start the question, or in abbreviation class, we usually use "viết tắt (abbreviation)". Most of classes have some specific words. This really has a significant impact when we use these lists of words as semantic features in question classification using machine learning.

With **syntactic trees**, the reason we extract this kind of features is that in Zhang and Lee (Zhang and Lee, 2003), the SVM algorithm using Tree Kernel method is the state of art for English question classification. So, we intend to answer the following question: "Does this method still outperforms the other methods in Vietnamese?". To use Tree Kernel, the questions must have corresponding syntax tree forms. For this, we used a parser tool for Vietnamese, Coltech-Parser (Le et al., 2009). The Coltech-Parser requires word segmentation for each input sentence. So, before using the parser, we used the VnWordSegmenter tool to segment input questions.

4.2.1 Keyword Collection from Web in VnQCS

Keywords are important semantic features though collecting them is not an easy task. Since the Internet contains a great number of web pages, this huge resource will help us to address the above-mentioned problem. The algorithm 1 describes how we collect keywords from Web and we hope it will be the basis for building semantic lexicons for other language processing tasks in Vietnamese.

- Firstly, we manually collect websites from the Internet which focus their content to one of our classes. Note that these websites should totally contain web content about one kind of class we want. As a result, our keywords from these will correspond to that class only and good for training features.
- In the next step, we crawl all links in these websites, note that we use only internal links since some external links will lead to other websites which are not related to the class we are interested in. Besides, in some cases all links that we want are only part of a website, we have to detect the format of them to get good content for training data.
- Base on links from previous step, we segment their text content. However, the words that we get in this step may contains some

Input: Set of websites which their contents focus on specific fine-grained class input
Output: Good keywords set for fine-grained class input

$setOfPages$ = {};
foreach $website$ *in* $websitesInputSet$ **do**
    foreach $internalLink$ *in* $website$ **do**
        $setOfPages$ = $setOfPages$ + $pageOfThisInternalLink$;
    end
end

$setOfKeywords$ = {};
foreach $page$ *in* $setOfPages$ **do**
    remove all tags from $page$ to get $pageContent$;
    segment $pageContent$ to get $listOfWords$;
    foreach $word$ *in* $listOfWords$ **do**
        $setOfKeywords$ = $setOfKeywords$ + $word$;
        $wordFrequency[word]$ = $wordFrequency[word]$ + 1;
    end
end

sort all words in decreasing order of $wordFrequency$ **and save result to** $sortedWords$**;**
eliminate all words which have low IDF index from $sortedWords$**;**
choose top words from $sortedWords$ **and save result to** $keywordsResult$**;**

Algorithm 1: Collect Keywords of a Specific fine-grained Class

trivial words like "là" (is), "và" (and), "hoặc" (or)... Since these stop words have low meaning, we can threshold their IDF[2] index and eliminate small value ones. Besides, we count the frequency of each word in each class and choose top N words with biggest frequencies. Finally, reviewing these N words and removing unsuitable ones for target class are necessary works that we have to do.

5 Experiments

This section describes our experiments on the Vietnamese corpus that we built. These experiments will help us find out the most suitable combination of algorithms and features for this task on this dataset. The results of experiments indicate that the semantic features, especially keywords, are really useful.

We designed three experiments to test the precision of our classifiers on Vietnamese questions corpus which we built. The corpus consists of two data sets: a training data set which includes nearly 5500 questions and a testing data set which includes 500 questions translated from TREC 10 and all of questions are in 6 coarse-grained classes or 50 fine-grained classes. To evaluate the experimental results, we use two main measures: weight average precision and weight average recall, which are all micro-average values.

- The first experiment evaluates the individual contribution of different feature types to question classification precision. In particular, we use Weka (Hall et al., 2009) to run machine learning algorithms namely Decision Tree (DT) (Quinlan, 1986), Naive Bayes (NB) (Bayes, 1763), SVM and Voting (Parhami, 1994), which are trained from our data we built using the following feature set: bag-of-words.
- In the second experiment, both bag-of-words and keyword features will be used together on machine learning algorithms on Weka. The goal is to verify the contribution of keyword features to question classification precision.
- Finally, we experiment with syntactic tree features set on SVM-Light-TK (Moschitti, 2004). This test will show us different affection of syntactic features between English and Vietnamese to question classification precision, and the important role of semantic features in Vietnamese question classification.

[2] http://en.wikipedia.org/wiki/Tf%E2%80%93idf

5.1 Bag-of-words

	6 class		50 class	
	Precision	Recall	Precision	Recall
Decision Tree	87.3%	87%	78.2%	76.2%
Naive Bayes	84.4%	83.6%	78.2%	73%
SVM	**91.2%**	**91%**	**83.1%**	**82.4%**
Majority Voting	91%	90.6%	81.1%	79.4%

Table 2: The question classification results using different machine learning algorithms, with same kind of feature: bag-of-words.

Like results in (Zhang and Lee, 2003), table 2 shows us that in Vietnamese question classification, SVM still outperforms other methods with the same kind of features. With bag-of-words features, SVM model achieves the highest precision with 91.2% on coarse-grained class and 83.1% on fine-grained class classification.

5.2 Bag-of-words + Keywords

	6 class		50 class	
	Precision	Recall	Precision	Recall
Decision Tree	86.2%	86.2%	80.3%	77.4%
Naive Bayes	87.4%	86.2%	81.1%	78.4%
SVM	**94.1%**	**94%**	**85.4%**	**83.8%**
Majority Voting	94.1%	94%	83.5%	81.8%

Table 3: The question classification results using different machine learning algorithms, with same kind of features: bag-of-words and keywords.

If only using bag-of-words, we can not fully exploit the semantic elements of the language in Vietnamese. As we expected, with both bag-of-words and keyword features used together, although the precision of classification using DT or NB only increases slightly, it increases significantly if we use SVM. In particular, with 6 coarse-grained classes, the precision increases from 91.2% (table 2) to 94.1% (table 3), and with 50 fine-grained classes, it increases from 83.1% (table 2) to 85.4% using SVM (table 3). So, keyword features have an important role to increase the precision of question classification.

5.3 Tree Kernel

We experimented SVM-Light-TK to using SVM combined Tree Kernel for Vietnamese question classification since this state of the art method is successful for English data. However, SVM-Light-TK can only classify binary label, we use one-vs-all strategy for problems of 6 coarse-grained classes and 50 fine-grained classes. The precision of this taxonomy is 88.4% for coarse-grained classes but only 75.1% for fine-grained classes (see table 4).

	6 class		50 class	
	Precision	Recall	Precision	Recall
Tree Kernel	88.4%	88%	75.1%	67.4%
Bag-of-words	91.2%	91%	83.1%	82.4%
Bag-of-words + Keywords	**94.1%**	**94%**	**85.4%**	**83.8%**

Table 4: The question classification results using SVM algorithm with some different kinds of features.

6 Conclusion

There are two main contributions of this paper. Firstly, we created a corpus for Vietnamese question classification (section 3). All the English questions in TREC 10 were translated into Vietnamese not only for this research but also many works in the future. This corpus will be made publicly available. Secondly, we extracted several feature groups and found out that the semantic features (bag-of-words and keywords) are really helpful to Vietnamese question classification meanwhile syntactic ones (syntactic tree) don't contribute so much to taxonomy precision. So we propose a method for collecting keywords from the Internet in a large scale. There isn't any WordNet for Vietnamese but with this method, we still have enough training data features for classifying a large range of Vietnamese questions.

Though Vietnamese question classification is a new challenge and there is not any work done on this, our experimental results indicate that the Vietnamese question classification can be addressed with relatively high precision using machine learning approaches. The result of classification can achieve a high precision of 94.1% with coarse-grained class classification and 85.4% with fine-grained class classification.

Acknowledgments

We wish to thank The Vietnam National Foundation for Science and Technology Development (NAFOSTED) for financial support. Thanks also to NLP group at University of Engineering and Technology for partially supporting us in building the corpus. Finally, we thank the anonymous reviewers for helping us improve the presentation.

References

T. Bayes. 1763. An essay towards solving a problem in the doctrine of chances. In *Philosophical Transactions of the Royal Society*, volume 53, pages 370–418.

M. Collins and N. Duffy. 2001. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems (NIPS14)*.

C. Cortes and V. N. Vapnik. 1995. Support vector machines. In *Machine Learning*, pages 273–297.

Q. N Dat, Q. N. Dai, and B. P. Son. 2009. A vietnamese question answering system. In *International Conference on Knowledge and Systems Engineering, KSE*, pages 26–32.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, R. Reutemann, and I. H. Witten. 2009. The weka data mining software: an update. *ACM SIGKDD Explorations*, 11:10–18.

E. Hovy, L. Gerber, U. Hermjakob, C. Lin, and D. Ravichan-dran. 2001. Toward semantics-based answer pinpointing. In *the DARPA HLT conference*.

A. C. Le, P. T. Nguyen, Vuong H. T., M. T. Pham, and T. B. Ho. 2009. An experimental study on lexicalized statistical parsing for vietnamese. In *KSE '09 Proceedings of the 2009 International Conference on Knowledge and Systems Engineering*, pages 162–167, Washington, DC, USA.

X. Li and D. Roth. 2006. Learning question classifiers: The role of semantic information. *Nat. Lang. Eng.*, 12(3):229–249.

A. Moschitti. 2004. A study on convolution kernels for shallow semantic parsing. In *Proceedings of the 42-th Conference on Association for Computational Linguistic*, Barcelona, Spain.

B. Parhami. 1994. Voting algorithms. In *IEEE Transactions on Reliability*, volume 43, pages 617–629.

J. R. Quinlan. 1986. Induction of decision trees. In *Machine Learning 1*, pages 81–106. Kluwer Academic Publishers.

M. V. Tran, D. V. Nguyen, T. O. Tran, T. U. Pham, and Q. T. Ha. 2009. An experimental study of vietnamese question answering system. In *IALP '09 Proceedings of the 2009 International Conference on Asian Language Processing*, pages 152–155, Washington, DC, USA.

E. Voorhees. 1999. The trec-8 question answering track report. In *Proc. of 8th Text Retrieval Conference, NIST*, pages 77–82, Gaithersburg, MD.

E. Voorhees. 2000. Overview of the trec-9 question answering track. In *Proc. of 9th Text Retrieval Conference, NIST*, pages 71–80, Gaithersburg, MD.

E. Voorhees. 2001. Overview of the trec 2001 question answering track. In *Proc. of 10th Text Retrieval Conference, NIST*, pages 157–165, Gaithersburg, MD.

E. Voorhees. 2002. Overview of the trec 2002 question answering track. In *Proc. of 11th Text Retrieval Conference, NIST*, pages 115–123.

D. Zhang and W. S. Lee. 2003. Question classification using support vector machines. In *Proc. of SIGIR*, pages 26–32.