2004－NL－162　(20)
2004／7／16

SVM を用いた質問分類のための新しい着目すべき特徴について

スコブロン・マーティン　　荒木　健治
北海道大学大学院　情報科学研究科

質問分類は質問応答において非常に重要である．SVM はスパースネス，高次元の問題に対して有効であると知られている．但し，頻繁に使用される bag-of-words のアプローチでは質問に含まれている情報を完全には利用できない．この情報を獲得するために，我々は 3 つの新しい着目すべき特徴として「従属単語の分類」,「質問の焦点」,「構文意味構造」を提案する．新しい着目すべき特徴を用いた質問分類の結果は，標準的な bag-of-words 手法とその他の文献で挙げられている手法と比較してより高い精度が示された．

キーワード： 質問分類，特徴選択，SVM，機械学習，質問応答

New Features for the Question Classification Using Support Vector Machines

Marcin Skowron Kenji Araki
Graduate School of Information Science and Technology, Hokkaido University

Question classification is of crucial importance for question answering. Support Vector Machines are known to work well for sparse, high dimensional problems. However, the frequently used bag-of-words approach does not take the full advantage of information contained in a question. To capture this information we propose three new feature types: Subordinate Word Category, Question Focus and Syntactic-Semantic Structure. As the results demonstrate, the inclusion of the new features provides higher accuracy of question classification, comparing to the standard bag-of-words approach and other methods that were described in the literature so far.

Keywords: Question classification, feature selection, Support Vector Machines, Machine Learning, Question Answering

1 Introduction

Along with the rapid growth of the text available in the Internet it has become more difficult for users to find specific information. The standard approach, querying an Internet search engine often returns thousands of results, containing a ranked list of documents along with their partial content (snippets). For an average Internet user, it is often time-consuming and laborious to find requested information. Often, before accessing searched information a user has to connect to several servers and scan trough dozens of documents to locate it. We think that for a human-being the most natural and straightforward approach to such a task is to ask a question in a natural language form. The output ought to be a correct answer, resembling as close as possible, those given by human beings. The realization of this task is an active research field in the current Question Answering (QA) systems.

In order to provide a correct answer to a question from a large collection of documents, like that of the Internet, one needs to impose some constraints on the scope of the possible answers. Such a constrain frequently used in QA systems is a question category. Question classification assigns a category to a given question based on the type of answer entity the question represents [12]. The result of the question classification provides the means to decrease the number of answer candidates. Consequently, a computer system does not need to verify all candidates found in the retrieved documents to decide if it is a correct answer to a given question. Because the verification based exclusively on the expected answer type is often sufficient to find a correct answer it has been widely accepted that a question classification is of the prime importance for QA systems.

This paper describes automatic method of question classification using Support Vector Machines (SVM)[6] in a taxonomy that includes 6 coarse grained and 50 fine grained categories. We introduce 3 new feature types that help to capture additional, useful for question classification information, which is pass over in the standard, bag-of-words approach. These are Subordinate Word Category, Question Focus and Syntactic-Semantic Structure. As the results demonstrate, the inclusion of these feature types helps to achieve a higher accuracy in a question classification task, comparing to the one obtained using the bag-of-words approach. Furthermore, the accuracy obtained using the set of the introduced feature types is the highest result reported in the literature so far, for this taxonomy and dataset.

2 Question Classification

Question classification is defined as the task that, given a question, maps it to one of *k* classes, which provide a semantic constrains on the sought-after answer [13]. Question classification, typically with other constrains on the answer, is used in a downstream process that leads to selection of a correct answer, from among several candidates. As described in the literature, a QA system that is able to classify a question using more detailed taxonomy and posses appropriate tools to effectively use this information to extract and verify answer candidates, achieves higher overall accuracy [5, 15]. Additionally, in some systems question category information is used also in a question category depended query formation process [17]. As the results show, such a query retrieves less distorted set of documents, where a correct answer appears more frequently, comparing to a set retrieved with a standard keyword based query formation process.

2.1 Taxonomy

In recent years, numerous question taxonomies have been defined, but there is no one standard, used by all the systems. For example, this is the case with the systems participating in the TREC QA-Track. Most of them uses their own question taxonomy. Moreover, the used taxonomy is frequently redefined on a year to year basis. Usually, the systems use the taxonomy consisting of less than 20 question categories. However, as demonstrated by several QA systems, employing a more detailed one, consisting of fine grained category definition is beneficial in the process of positioning and verifying answer candidates.

In our work, we used hierarchical, two-layered taxonomy proposed by Li and Roth in [13] consisting of 6 coarse grained and 50 fine grained categories, which are shown in Table 1. Recently, this taxonomy was employed also in a few other QA systems, and different approaches to automatic question classification were evaluated based on it [4, 7, 12, 13, 23]. We decided to use this taxonomy because of its good overall coverage of question types that are usable by the answer candidates verification module of our QA system, as well as a freely available training dataset. Using it we could also compare the question classification results of our SVM based classifier to the other methods that used the same dataset.

2.2 Datasets

For the training and evaluation of our question classifier, we use the publicly available dataset provided by USC [8], UIUC [13] and TREC[20, 21, 22] consisting of 5,500 classified questions for the training set, and 500 more for testing. The test data is a set from the Question Answering Track of TREC 10. The training set is assembled from previous TREC questions as well as from archives of online question answering systems [Li, Roth 2002]. All the questions from these datasets have been manually labeled using the taxonomy presented in Table 1, by UIUC [13].

Table 1: The coarse and fine grained question categories

Coarse	Fine
ABBR	abbreviation, expansion
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease, 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

3 Approaches to the Question Classification Task

The approaches to question classification can be discriminate into the following, three main groups; rule-based, language modeling and machine learning based[1].

3.1 Rule Based Classification

In the rule based approach, hand-written grammar rules and a set of regular expression are employed to parse a question and to determine the answer type [Van Durme, 2003]. With this approach the researches have faced a number of limitations:

- Hand-writing the rules and preparing the efficient regular expressions is a difficult and time consuming process
- Hand-written rules have a limited coverage and it is fairly complicated to broaden the scope of answer categories to include more detailed ones
- In case of adopting a new taxonomy, many previously prepared rules have to be modified or completely rewritten

Considering these limitations, the majority of systems that use hand-written rules are bound to use a limited number of question type categories. Consequently, question category information is usable only in a limited extend, which as previously described, influences the performance of the whole QA system [5, 15].

3.2 Machine Learning Based Classification

In the machine learning approach, the expert knowledge is replaced by a sufficiently large set of labeled questions. Using this collection, a classifier is trained in a supervised manner. Possible choice of classifiers include but

[1] We do not include the explanation on language modeling approach to question classification, since for a detailed taxonomy, similar to one used in our work, no successful implementation of classifier has been found. For the description and results of language modeling based approaches to question classification see [4, 12].

is not limited to: Neural Network, Naive Bayes, Decision Tree and Support Vector Machines. The machine learning approach addresses many limitations of the rule-based method, which were presented above. The advantages include:

- Short creation time
- No need for expert knowledge (automatic creation of a classifier)
- Broader coverage, can be obtained by providing new training examples
- If needed, the classifier can be flexibly reconstructed (retrained) to fit to a new taxonomy

At present, the results achieved using the machine learning constitute a state of the art in the question classification. The different machine learning methods presented below utilized the same taxonomy and dataset, described in **2.1** and **2.2**.

4 State of the Art in Question Classification

Currently, the primary machine learning algorithm used for question classification is Support Vector Machines (SVM) [7, 18, 23]. Researches are usually motivated to apply the SVM to question classification task by the fact that SVM constantly outperformes other machine learning techniques in several applications including the text classification, which to some extend is similar to the question classification [9, 16, 19]. However, as the results presented in the literature demonstrate, the highest accuracy was obtained using the SNoW learning architecture based classifier.

The research of Zhang and Lee [23] presented work on question classification using the SVM and compared its results to these obtained by other machine learning approaches like Nearest-Neighbors (a simplified version of well-known kNN algorithm), Naive-Bayes, Decision Tree and Sparse Network of Winnows (SNoW). All the classifiers were trained using the same dataset. The SVM classifier achieved the highest results comparing to other machine learning based classifiers, both in the bag-of-words and the bag-of-ngrams approaches. The advantage of the SVM was especially significant under the fine grained category definition [2]. The research proposed also a specific kernel function called the tree kernel, to enable the SVM to take advantage of the syntactic structures of question. Unfortunately, its application to the classifier under the fine grained category definition did not bring improvements. The highest accuracy reported in this work for the first classification, under the fine grained category definition was achieved using the bag-of-words (BOW) features. This and other results of the state of the art systems, obtained using the same dataset, for the first classification(P1) under the fine grained category definition [13] are presented in Table 2.

[2]For the details of the evaluation of several machine learning approaches in the question classification task see [23].

Table 2: The question classification accuracy for the fine grained categories obtained by the state of the art systems

	SVM (BOW) [23]	SVM (BSH) [7]	SNoW [13]
P1	80.2%	82.0%	84.2%

Similar result was reported in the later work that used the SVM classifier with the bag-of-words features [7]. The authors performed the experiments after the dimensionality reduction by computing the term space transformation using singular value decomposition (SVD) and applying BCH codes to convert a multi-class classification problem into a number of two-class problems. The accuracy improvement to 82.0%, was reported in a bag-of-bigrams approach, after the inclusion of the name entity based features, for the seven selected Named Entity categories [2].

The work of Li and Roth [13] described the system that obtained the highest question classification accuracy described up to date for the presented taxonomy and dataset, using the classifier based on the SNoW (Sparse Network of Winnows) learning architecture. The classifier was trained using a rich selection of the features including: part-of-speech (POS) tags, non-overlapping phrases (chunks), named entities (NEs), head chunks, semantically related words, conjunctive (n-grams) and relational features. The total number of used features is about 200,000; for each question, up to a couple of hundreds of them are active.

As presented in Table 2, despite the fact that SVM was found to outperform other machine learning approaches in several applications, the highest result obtained so far for the question classification task, was achieved using the SNoW learning architecture. We think that the high performance of SNoW classifier is the result of the sensible selection and effective application of a rich set of features, especially these based on the semantic analysis. Up to date, no SVM based classifier was able, to successfully employ a similar number of features, to provide such detailed representation of questions, helpful in the classification task. Our work addresses the problem of the feature selection for the SVM based classifier. Below we introduce three new feature types and present their impact on the accuracy of the question classification.

5 Question Classification with Support Vector Machines

5.1 Binary Classifier for the Multi-class Problems

The task of our experiment is to classify a given question to one of the 50 possible categories. Although the SVM is inherently binary classifier, it is possible to extend its use to a multi-class problems like that of question classification. This is performed by reducing the multi-class problem to multiple binary classifications [1]. There are two popular alternatives: one-against-all and all-pairs. We used the former approach, constructing 50 separate classifiers trained on data where the questions from one question category formed one class and all the remaining questions

from other categories created the second one. The SVM Light [10] implementation of SVM is used in the following experiments.

6 Feature Selection

As demonstrated in the previous works, the feature selection is of the crucial importance in a wide spectrum of classification task that uses machine learning [13, 18, 19]. Question classification to some extend is similar to the text categorization. The goal in the later one is to assign a given text to a previously defined class. In the question classification, a given text is usually a few words long question sentence. As showed in [13], question classification requires much more complicated features than text categorization. However, up to date there was no similarly effective application of such a rich set of features for SVM based classier.

The feature selection requires to find a balance between the need to provide a sufficient information to the classifier and the danger of providing them in exceed. In the first situation, because of lack of sufficient information the classifier is not able to effectively discriminate the test questions based on the learned model. On the other hand, providing to many features leads to overfitting during training process with sparse data, introduces noise in the feature space and inflicts higher computational complexity. A frequently used solution of this problem is the dimensionality reduction. In this process, a great care has to be taken to minimize the loss of features that are useful for the classification.

6.1 New Feature Types for the Question Classification

The bag-of-words approach is frequently used in a number of classification tasks including the question classification. However, in our opinion, with this approach the classifier is not able to take the full advantage of information contained in a question, which is useful for classification. In the bag-of-words approach, a word can be used only directly, by checking whether it exists in a feature space or not. Similarly, in the training process, the model is created without utilizing the semantic information contained in question words. A position of a word in a sentence is another pass over information in this approach, similar to information on syntactic-semantic structures. To address these limitations we introduce three new feature types for a question classification task. These are: Subordinate Word Category, Question Focus and Syntactic-Semantic Structure.

6.1.1 Subordinate Word Category

In the bag-of-words and similar approaches (eg. bag-of-ngrams), information contained in a word can be used only directly. In the training process of a classifier, as well as during the classification of test questions, other types of information existing on different layers (eg. semantic), are not utilized. In a consequence, without providing a representation of a given word in a higher, more general level, the words that less frequently occur in a dataset are used only to a very limited extend, or not used at all. We think that these words posses valuable semantic information, which is useful for the question classification. In several cases, the remaining words exist at the same time, in several question categories, and as such do not provide sufficient information to the classifier, to correctly assign a question category. For example in the test question "What is the proper name for a female walrus ?" the words "What", "is", "proper", "for" or "female" can be found in several categories, while the word "walrus" did not appear in training data. In this situation, the word "walrus", the only one that could potentially provide really useful information to a classifier, can not be used in the bag-of-words approach, thus it is difficult to correctly discriminate such questions.

To capture semantic information contained in a word on a higher level of representation, we propose a new feature type, the Subordinate Word Category. This feature type is realized by assigning a WordNet [14] hyponym to a common nouns found in a given question. The list of selected hyponyms includes 25 categories like: animal, plant, vehicle, quantitative relation, length, body part, land, water, people, etc. If found, these hyponyms are assigned for all common nouns found in a question and add as a new entry to a feature space. Additionally, a common category "YEAR" is assigned for cardinal numbers consisting of four digits and is used to substitute the original word; similarly to the category "NUMBER" used for all the remaining cardinal numbers.

6.1.2 Question Focus

In the bag-of-words approach all words are treated equally, without considering their position in a question. Question focus word, which is often a good indication of question category is another type of information that cannot be used in this approach. To capture this additional, useful for classification information we introduce the Question Focus feature type.

Question focus word is recognized using a set of a regular expression applied to a POS tagged question. For example, one of the regular expression searches for the first common noun appearing after the word "What". For instance, in the question: "What county is Modesto California in?" the word "county" is recognized to be the question focus word. After applying this feature, a few questions from the "LOC::other" category, both in training and test data, gain one more additional feature. Similarly, if discovered, the question focus words are assigned as a common features for the remaining questions from this category, as well as for the questions contained in the other categories from the dataset. As the results demonstrate the inclusion of this feature type leads to the improvement in the accuracy of question classification.

6.1.3 Syntactic-Semantic Structure

Our analysis of the dataset revealed that some syntactic-semantic patterns that frequently exist in questions from one category, do not appear in the others. In our opinion, the ability to capture these patterns provides a valuable information for a classifier that is not used in the standard bag-of-words approach. To construct the highly distinguishable structures, the syntactic-semantic patterns need to be general enough to allow variation of different questions that belongs to one category, and at the same time, strict enough to capture the differences between questions from one category and the others. Based on the training dataset, the patterns were automatically generated using the following processing:

- Using the set value of TFIDF, select and later preserve in the original form the collection of "the categories important nouns"
- Substitute the remaining nouns with the tokens that respect the surface feature of a given word
- Substitute the cardinal numbers with one, common token

If such a pattern is found to exist at least twice in one and only one question category, it is stored and used to assign an additional feature to questions that share a similar syntactic-semantic pattern.

7 Results

As explained in [13] the authors were aware that using their taxonomy, the classification of some questions may be ambiguous between few question categories. In their works, the classifier is permitted to assign a multiple labels to one question in case if the classifier confidence level is low. Although this approach can be beneficial in practical application to a QA system, for the sake of achieving a strict measure of classification accuracy, we decided to count the precision of correctly classified questions using only the first answer category assigned by the classifier.

Table 3 shows the accuracy of question classification for the fine grained categories, achieved using the standard bag-of-words approach (BOW), as well as the results obtained after extending the BOW with the new feature types (SWC - Subordinate Word Category, QF - Question Focus, SSS - Syntactic-Semantic Structure). As the results demonstrate, the inclusion of each of the proposed feature type contributed to a higher accuracy, compared to the bag-of-words approach. The biggest improvement of 3.0%, was achieved after the inclusion of the Subordinate Word Category feature type.

The results obtained after adding various sets of the feature types are presented in Table 4. The highest accuracy was achieved in the run using all the proposed feature types (SWC QF SSS). This result, obtained by the SVM based classifier, is higher than those reported in the previous researches [4, 7, 12, 13, 23], for the same training and test data collection.

Table 3: The question classification accuracy for the first classification under the fine grained categories using different feature types

		New Feature Types		
	BOW	SWC	QF	SSS
P1	80.2%	83.2%	82.6%	81.4%

Table 4: The question classification accuracy for the first classification under the fine grained categories using different set of feature types

		Set of Feature Types	
	BOW	SWC QF	SWC SSS
P1	80.2%	84.4%	84.2%
	BOW	QF SSS	SWC QF SSS
P1	80.2%	82.6%	**84.6%**

8 Discussion

The research confirmed that the high-performance question classification requires to employ much richer set of features than this available on the word level. The introduction of the new feature types supplied additional information to the SVM based classifier that could not be captured and used in the standard bag-of-words approach. Using the whole set of the presented feature types the classifier, achieved the result of 84.6%, for the first classification under the fine-grained categories definition. This result demonstrates that semantic and structural information contained in a question can provide highly discriminative features that help to classify a given question to a correct category. All the presented feature types are based on the freely available tools, like POS tagger [3] and WordNet [14], and are constructed automatically, which is not always a case in the other methods (eg. the good performance of the SNoW based classifier, depends heavily on the feature called "RelWords" (related words), which are constructed semi-automatically).

9 Conclusions and Future Work

This paper presented a machine learning approach to question classification task using the Support Vector Machines. We proposed three new feature types, that address the limitations of the bag-of-words and similar approaches (eg. bag-of-ngrams), frequently used in several classification tasks. The experimental results demonstrated that the inclusion of the new features types: Subordinate Word Category, Question Focus and Syntactic-Semantic Structure was useful for improving the performance of the classifier over the bag-of-words approach. Using the set of these three feature types, the result of 84.6% was achieved, bringing the error reduction of 22% comparing to the bag-of-word approach. A comparison with the state of the art systems has shown that using these features, the classifier was able to achieve better accuracy than any other machine based classifier before, including the SVM and

SNoW learning architecture based ones. The additional advantage of this approach is the fact that, the creation of the new feature types was performed fully automatically, using only the freely available tools like POS Tagger and WordNet.
Our future work includes, further evaluation of the introduced feature types, especially the Syntactic-Semantic Structure, which in our opinion, posses the potential to provide a higher coverage of various question categories. We intend also to perform a detailed evaluation of different kernel functions applied to the classification tasks with the Support Vector Machines.

Bibliography

[1] Allwein E., Schapire E., Singer Y. (2000) Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113-141

[2] Bikel D.M., Schwartz R.L., Weischedel R.M. (1999) An algorithm that learns what‘s in a name. Machine Learning, vol. 34, no. 1-3, 211-231

[3] Brill E. (1995) Transformation-Based Error Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics 21(4), 543-566

[4] Brown J., (2003) Entity-Tagged Language Models for Question Classification in a QA System, Report on: http://www-2.cs.cmu.edu/˜jonbrown/IRLab/Brown-IRLab.pdf

[5] Cardie C., Ng V., Pierce D., Buckley C., (2000) Examining the Role of Statistical Knowledge Sources in a General-Knowledge Question-Answering Systems. In Proceedings of the 6th Applied Natural Language Processing Conference, 180-187

[6] Cortes C., Vapnik V. (1995) Support-Vector Network, Machine Learning 20, 1-25

[7] Hacioglu K., Ward W., Question Classification with Support Vector Machines and Error Correcting Coded (2003) In the Proceedings of HLT-NACCL 2003, 28-30

[8] Hovy E., Gerber L., Hermajakob U., Lin C., Ravichandran D., (1999) Towards Semantics-based Answer Pinpointing. In Proceedings of the DARPA Human Language Technology conference (HLT)

[9] Joachims T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In Proceedings of European Conference on Machine Learning (ECML98), 137-142

[10] Joachims T. (1999) Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, MIT-Press

[11] Joachims T. (1999) Advances in Kernel Methods: Support Vector Learning, chapter Making large-Scale SVM Learning Practical. Chap. 11. MIT-Press

[12] Li W., (2002) Question Classification Using Language Modeling. CIIR Technical Report

[13] Li X., Roth D. (2002) Learning Question Classifiers. In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), 556-562

[14] Miller G. (1995) WordNet: a lexical database for English. Communications of the ACM 38(11), 39-41

[15] Pasca M.A., Harabagiu S.M., (2001) High Performance Question/Answering. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 366-374

[16] Rennie J.D.M., Rifkin R., (2001) Improving multiclass text classification with the support vector machines, MIT Artificial Intelligence Laboratory Publications, AIM-2001-026

[17] Skowron M., Araki K. (2004) What Can Be Learned from Previously Answered Questions? A Corpus-Based Approach to Question Answering. Intelligent Information Systems 2004. New Trends in Intelligent Information. Proceedings of the International IIS: IIPWM04 Conference, 379-387

[18] Suzuki J., Taira H., Sasaki Y., Maeda E., (2002) Question Classification using HDAG Kernel, Workshop on Multilingual Summarization and Question Answering 2003, post-conference workshop in conjunction with ACL-2003, 61-68

[19] Taira H., Haruno M., (1999) Feature Selection in SVM Text Categorization. In Proceedings of the 16th Conference of the American Association for Artificial Intelligence (AAAI99), 480-486

[20] Voorhees E., (1999) The TREC-8 Question Answering Track Report. In Proceedings of the 8th Text Retrieval Conference (TREC8), 77-82

[21] Voorhees E., (2000) Overview of the TREC-9 Question Answering Track Report. In Proceedings of the 9th Text Retrieval Conference (TREC9), 71-80

[22] Voorhees E., (2001) Overview of the TREC 2001 Question Answering Track Report. In Proceedings of the 10th Text Retrieval Conference (TREC10), 157-165

[23] Zhang D., Lee W.S. (2003) Question Classification using Support Vector Machines. Proceedings of the 26th ACM SIGIR, 26-32