Effectiveness of Combined Features for Machine Learning Based Question Classification

Marcin Skowron† and Kenji Araki†

Question classification is of crucial importance for question answering. In question classification, the accuracy of ML algorithms was found to significantly outperform other approaches. The two key issues in classification with a ML-based approach are classifier design and feature selection. Support Vector Machines is known to work well for sparse, high dimensional problems. However, the frequently used Bag-of-Words approach does not take full advantage of information contained in a question. To exploit this information we introduce 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 compared to the standard Bag-of-Words approach and other ML based methods such as SVM with the Tree Kernel, SVM with Error Correcting Codes and SNoW. A classification accuracy of 85.6 % obtained using the three introduced feature types is, as of yet the highest reported in the literature, bringing error reduction of 27 % compared to the Bag-of-Words approach.

KeyWords: *Question Classification, Feature Selection, SVM, Machine Learning, Question Answering*

1 Introduction

With the rapid growth of text available on the Internet, it has become more difficult for users to find specific information. The standard approach of 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, in order to find it, a user has to connect to several servers and scan through dozens of documents. 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 much 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 possible answers.

† Language Media Laboratory, Graduate School of Information Science and Technology, Hokkaido University

A constraint 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 (Li 2002). The outcome of question classification serves 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 a verification based exclusively on the expected-answer type is often sufficient to find a correct answer, question classification is of prime importance for QA systems.

The Bag-of-Words (BOW) approach is frequently used in a number of classification tasks, including question and text classification, where it obtains a state of the art performance. However, in our opinion in the case of question classification, where a given text is a few words long sentence, a classifier achieves higher precision by combining many features and learning "dense" concepts. In this paper we describe the automatic method of question classification using Support Vector Machines (SVM) (Cortes and Vapnik 1995) (Vapnik 1995) in a taxonomy that includes 6 coarse-grained and 50 fine-grained categories and evaluate 3 new feature types (Subordinate Word Category, Question Focus and Syntactic-Semantic Structure) that help to exploit additional information that is useful for question classification. As the results demonstrate, the inclusion of these feature types provides higher accuracy in the question classification task, compared to that obtained using the BOW approach. Furthermore, the accuracy achieved 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 of crucial importance for QA Systems. Question classification is defined as the task that, given a question, maps it to one of k classes, which provide a semantic constraint on the sought-after answer (Li and Roth 2002). This information, typically with other constraints 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 with more detailed taxonomy and use this information to extract and verify answer candidates, achieves higher overall accuracy (Cardie, Ng, Pierce and Buckley 2000) (Pasca, Harabagiu 2001). Additionally, in some systems question category information is also used in a question category dependent query formation process (Skowron, Araki 2004). As the results show, such a query retrieves a less distorted set of documents, where a correct answer appears more frequently, compared to a set retrieved with a query formed in the standard keyword-based approach.

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

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 of the systems participating in the TREC QA-Track. Most of them implement their own question taxonomy. Moreover, the used taxonomy is frequently redefined on a year-to-year basis. Usually the systems use a taxonomy consisting of less than 20 question categories. However, as demonstrated by several QA systems, employing a more detailed one consisting of a 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 (Li et al. 2002) consisting of 6 coarse-grained and 50 fine-grained categories [Table 1]. Recently, this taxonomy was employed also in other QA systems, and different approaches to automatic question classification were evaluated based on it (Brown 2003; Hacioglu and Ward 2003; Li 2002; Li et al. 2002; Zhang and Lee 2003). We decided to use this taxonomy because of its effective overall coverage of question types that are usable by the answer candidate verification module of our QA system, and a freely available training dataset. Using it, we could also compare the question classification results of our SVM based classifier to other methods that used the same dataset.

For the training and evaluation of our question classifier, we use the publicly available dataset provided by USC (Hovy, Gerber, Hermajakob, Lin and Ravichandran 1999), UIUC (Li et al. 2002) and TREC(Voorhees 1999, 2000, 2001), which consists 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. All the questions from these datasets have been manually labeled using the taxonomy presented in Table 1, by UIUC (Li et al. 2002).

3 Previous Method

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

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 (Durme, Huang, Kupsc and Nyberg 2003). With this approach the researches have faced a number of limitations:

- Hand-writing classification rules is a difficult and time-consuming process.
- Hand-written rules have limited coverage and is fairly complicated to broaden the scope of answer categories to include more detailed ones.
- In order to adopt 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 limited to its use, which as previously described, influences the performance of the whole QA system (Cardie et al. 2000; Pasca et al. 2001). In the machine learning approach, expert knowledge is replaced by a sufficiently large set of labeled questions. Using this collection, a classifier is trained in a supervised manner. Possible choices of classifiers include but are 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.
- Classifier is created automatically; no classification rules need to be provided by hand.
- Broader coverage; can be obtained by providing new training examples.
- If required, the classifier can be flexibly reconstructed (retrained) to fit to a new taxonomy.

At present, the results achieved using the machine learning approach represents a state of the art in question classification. The different machine learning methods presented below utilized the same taxonomy and dataset as described in Section 2.

Currently, the primary machine learning algorithm used for question classification is Support Vector Machines (SVM) (Hacioglu et al. 2003; Suzuki, Taira, Sasaki, Maeda 2002; Zhang et al. 2003). Researchers apply SVM to question classification because it constantly outperforms other machine learning techniques in several applications including text classification,

1 We do not include an explanation of the language modeling approach, due to its low performance using a detailed taxonomy. For more information refer to (Brown 2003; Li 2002).

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

	SVM (BOW) (Zhang et al. 2003)	SVM (BSH) (Hacioglu et al. 2003)	SNoW (Li et al. 2002)
P1	80.2 %	82.0 %	84.2 %

which is similar to question classification (Joachims 1998; Rennie, Rifkin 2001; Taira, Haruno 1999). However, as the results presented in the literature demonstrate, the highest question classification accuracy was obtained using the SNoW learning architecture-based (Carlson, Cumby, Rosen and Roth 1999) classifier.

The research of Zhang and Lee (Zhang et al. 2003) presented work on question classification using Support Vector Machines, 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, presented in Section 2. The SVM classifier achieved the highest results compared to other machine learning based classifiers, both in the Bag-of-Words and the Bag-of-Bigrams approaches[2]. The advantage of the SVM was especially significant under the fine-grained category definition [3]. 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 (Li et al. 2002) are presented in Table 2.

Similar results was reported in later work that used the SVM classifier with the BOW features (Hacioglu et al. 2003). The authors performed the experiments after 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 (Bikel, Schwartz, Weischedel 1999).

The work of Li and Roth (Li et al. 2002) described the system that obtained the highest question classification accuracy achieved up to date for the presented taxonomy and dataset,

2 We confirmed this finding in the set of test performed using the collection of the machine learning algorithms implemented in the WEKA Data Mining Software (Witten and Frank 2000).

3 For the details of the evaluation of several machine learning approaches in the question classification task refer to (Zhang et al. 2003).

using the classifier based on the SNoW (Sparse Network of Winnows, (Roth 1998)) learning architecture. The classifier was trained using a rich selection of 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 features used is approximately 200,000; for each question, up to a couple hundred 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 those 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. Support Vector Machines (Cortes et al. 1995; Vapnik 1995) is based on the Structural Risk Minimization principle from Computational Learning Theory (Vapnik 1995). The SVM in the basic form learns the linear hyperplane that separates a set of positive examples from a set of negative examples with maximum margin (the margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples) (Marquez 2000). By using appropriate kernel functions, SVM can be extended to learn polonymical classifiers, radial basic function (RBF) network, and three-layer sigmoid neural nets.

The selection of this classifier was based on the following observations, concerning the properties of the SVM and the requirements of the question classification with the proposed method[4]:

- **High dimensional input space**.
  In the experiments to be discussed later, the number of used features is close to 9900.
- **Dense concepts and sparse instances**.
  As the previous results demonstrated, the effective SVM based classifier should combine many features (learn a "dense" concepts). The feature types introduced in this work, provide such an additional density to a used questions representation.

The objective of our experiment is to classify a given question to one of 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 (Allwein, Schapire, Singer 2000). There are two popular alternatives: one-against-all and all-pairs. We used the former approach,

4 Similar reasons were presented in (Joachims 1999) for the justification of the SVM application to the text categorization task.

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 (Joachims 1999) implementation of SVM is used in the following experiments.

4 Proposed Method

The feature selection is required to find a balance between the need to provide sufficient information to the classifier and the danger of providing them in excess. In the former, because of a 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 too many features leads to overfitting during a training process with sparse data, introduces noise in the feature space, and inflicts higher computational complexity. A frequently used solution is dimensionality reduction. Here, care has to be taken to minimize the loss of features that are useful for the classification.

As demonstrated in the previous works, the feature selection is of crucial importance for a wide spectrum of classification task, that use machine learning (Li et al. 2002; Suzuki et al. 2002; Taira et al. 1999). Question classification to some extent is similar to text categorization. The goal in the latter is to assign a given text to a previously defined class. In question classification, a given text is usually a few words long sentence. As shown in (Li et al. 2002), question classification requires more complicated features than text categorization. However, in spite of SVM robustness to handle large feature sets, as of yet there are no similarly effective applications of such a rich set of features for the SVM based classifier. Motivated by this, we decided to introduce new feature types for the SVM based classifier and to evaluate their impact on the accuracy of question classification.

The Bag-of-Words approach is frequently used in a number of classification tasks including question classification. However, in our opinion, with this approach the classifier is not able to take full advantage of information contained in a question. In the BOW 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 word position in a sentence is another overlooked information, similar to information on syntactic-semantic structures. To address these limitations we introduce three new feature types for the question classification task. These are: Subordinate Word Category, Question Focus and Syntactic-Semantic Structure.

4.1 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 classification of test questions, other types of information existing on different layers (eg. semantic) are not utilized. Consequently, 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 extent, if used at all. We think that these words possess valuable semantic information, which is useful for 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 BOW 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 (Miller 1995) hypernym to common nouns found in a given question. For a given noun a WordNet hypernym provides several generic words at different abstract level, starting from the most specific to the ones that convey more general meaning of a noun. While seeking more universal representation of a given noun, the procedure used to discover Subordinate Word Category aims also at ensuring that assigned hypernym closely preserves the original meaning of an initial noun. The procedure includes the following steps:

- Extract common nouns from a question.
- If exists, find a set of hypernyms for each of a noun.
- Preserving the order of assigned hypernyms reflecting hypernyms abstraction level, find the first one that matches a hypernym from the ordered list, as presented in Table 3.
- Assign a discovered hypernym as a new feature for a given question and in case of the training data adding as a new entry to a feature space.

The list of used hypernyms includes 20 categories, presented in Table 3 along with the examples of the corresponding nouns. Additionally, a common category "YEAR" is assigned for cardinal numbers consisting of four digits, and used to substitute the original word. Similarly, the category "NUMBER" is used for all the remaining cardinal numbers.

Table 3 List of the hypernyms used to assign Subordinate Word Category feature, along with the examples of nouns for which hypernyms were found.

#	**Hypernym**: example noun	#	**Hypernym**: example noun
1	**animal**: kangaroo	11	**magnitude**: depth
2	**plant**: flower	12	**sport**:golf
3	**vehicle**: plane	13	**show**: movie
4	**quantitative relation**: percentage	14	**structure**: monastery
5	**length**: diameter	15	**location**: province
6	**charge**: tax	16	**measure**: ton
7	**land**: continent	17	**substance**: silver
8	**water**: river	18	**time period**: summer
9	**series**: streak	19	**area**: center
10	**people**: population	20	**equipment**: camera

4.2 Question Focus

For the purpose of the question classification task, the question focus can be defined as a word in the given question that disambiguates it and emphasizes the expected answer type. 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 valuable indication of question category is another type of information that cannot be used in this approach. To exploit this additional, useful for classification information we introduce the Question Focus feature type.

In the experiments that follow a question focus word is recognized using a set of the regular expressions applied to a POS tagged question (Brill 1995). The set of expressions used to discover a question focus word, was developed based on human-knowledge, supported by the analysis of questions from TREC-8, TREC-9, TREC-10, TREC-2002, TREC-2003 [5]. The procedure used to assign the Question Focus feature includes the following steps:

- From the list of regular expressions as presented in Table 4, find one applicable to a given POS-tagged question. If for a given question, such regular expression does not exit the Question Focus feature is not assigned.
- Apply a discovered regular expression, to find a question focus word.
- Assign a discovered question focus word as a new feature for a given question and in case of the training data add as a new entry to a feature space.

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 Chicago 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 the additional common feature. Similarly, if discovered the question focus words are assigned for the remain-

5 The set of questions available at http://trec.nist.gov/data/qa.html.

Table 4 Regular expressions used to discover a Question Focus word.

Regular Expression	Example question with the QF word
What\/WP[^\/V]+?([a-z]+)\/NN	What state has the most Indians?
How\/WRB\s([a-z]+)\/[A-Z]+?	How far away is the moon?
What\/WP\stype\/NN\sof\/IN\s(\w+)\/NN	What type of bridge is the Golden Gate Bridge?
what\/WP\s(\w+\/NN\s)*(\w+)\/NN	CNN began broadcasting in what year?
What\/WP\s(is|was).+?\sname\/NN.+?\s([a-z]+?)\/NN\b	What is the name of the firm that makes Spumante ?
Who\/WP\s(is\/VBZ|was\/VBD)\s[A-Z][a-z]+?\/NNP\s	Who was Galileo ?
What\/WP\s.*?\b(\w+)\/[A-Z]+?\ss\/PRP	What city' s newspaper is called "The Enquirer" ?
What\/WP\swas\/VBD.+? \s(\w+)\/NN\b	What was the first Sam Spade novel?
What\/WP.*?\b(\w+)\/NN\s(is|was)	What soft drink is most heavily caffeinated?
What\/WP\sis\/VBZ.+?\s(\w+\/NN \s)*(\w+)\/NN\b	What is the fastest commercial automobile that can be bought in the US ?

ing questions from this category, as well as for the questions contained in the other categories from the dataset. Other regular expressions used to discover the question focus words include the one that assigns the word "speed" to be the focus word in the question "How fast is the speed of light?"(NUM:speed), word "flower" in the sentence "What is Australia's national flower?"(ENTY:plant) or "language" to be the question focus word in "What is the most frequently spoken language in the Netherlands?"(ENTY:language). Table 4 presents the regular expressions used to discover a question focus word.

4.3 Syntactic-Semantic Structure

Our analysis of the dataset revealed that some syntactic-semantic structures that frequently exist in questions from one category do not appear in the others. In our opinion, the ability to exploit these structures provides a valuable information for a classifier that is overlooked in the standard Bag-of-Words approach. To construct highly distinguishable patterns, the syntactic-semantic structures 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. In this work the structures were automatically generated based on the training dataset, with the following processing:

- Using the TFIDF (Term Frequency / Inverted Document Frequency (Salton, Buckley 1988)) value, select and later preserve in the original form the collection of "categories important nouns". The TFIDF value is obtained based on the training questions, where each question is treated as a single document.
- Substitute the remaining nouns with the tokens that respect the surface feature of a

given word(first letter capitalization, capitalization of the following letters, existence of the non-letter characters and numbers)[6].

- Substitute the cardinal numbers with one, common token.

If such a structure is found to exist at least twice in one and only one question category it is stored and assigned as an additional, common feature to questions that share it. The examples include the structure "What are <noun composed of only small letters>" frequently found in the DESC:def category, like in the question "What are sunspots", structure "Where is the <noun composed of a capitalized letter followed by small letters> <noun composed of a capitalized letter followed by small letters>" characteristic for the LOC:other category, like in the sentence "Where is the Euphrates River" or "What is <noun composed of only small letters> made of", that exists in the ENTY:substance category, like in the question "What is pastrami made of". Using the described method, 147 structures were found, providing an additional feature for various questions from 25 different categories.

5 Experiments

As explained in (Li et al. 2002) 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.

Our experiments, as well as the results presented in (Zhang et al. 2003) demonstrated, that under the fine-grained category definition the SVM based classifier achieves the highest accuracy with the linear kernel, using the Bag-of-Words, compared to ones obtained with other kernels, and using Bag-of-Bigrams approach. Hence, in the experiments that follow, the results obtained using the linear kernel with the BOW features are considered as a base-line for the results comparison. Additional experiments, which results are presented in Table 5 show that the usage of different number of features (set obtained after excluding the words that appeared more than: 1000 times (F1), 700 times (F2) and 1200 times (F3)), using uppercased letters (UC), the POS tagged words (POS), and Bag-of-Ngrams approach (BON) did not bring improvement.

6 For example, the common token is assigned for the nouns like Galileo and Beatles; another one for the nouns like atom and flower, different from the token assigned for the nouns like Coca-Cola and Rolls-Royce.

Table 5 The question classification accuracy for the first classification under the fine-grained categories using base-line approach with different number and type of features.

	BOW F1	BOW F2	BOW F3	UC F1	POS F1	BON
P1	**80.2%**	79.8%	79.4%	79.4%	79.6%	75.2%

Table 6 The question classification accuracy for the first classification under the fine-grained categories using different feature types and number of training examples (1000-5500).

		New Feature Types		
	BOW	SWC	QF	SSS
1000	66.8%	69.6%	69.4%	69.4%
2000	71.4%	75.2%	75.4%	73.4%
3000	75.0%	76.8%	78.0%	76.2%
4000	77.8%	78.0%	80.2%	79.2%
5500	**80.2%**	**83.2%**	**83.8%**	**81.4%**

The classifier was trained on 5 different size training datasets and tested on the set of 500 questions from TREC10. Table 6 shows the accuracy of question classification for the fine-grained categories, achieved using the 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). In case of the SSS feature, the TFIDF value was experimentally set to 0.2. The classification accuracy is measured as the proportion of the correctly classified questions among all test questions. As the results demonstrate, the inclusion of each of the proposed feature type contributed to a higher accuracy compared to the baseline approach. The biggest improvement of 3.6% was achieved after the inclusion of the Question Focus feature type. As the results show, the SVM handles large set of features without overfitting; the accuracy grows evenly along with the larger training set provided.

The results obtained after adding various sets of the feature types are presented in Table 7. The highest accuracy of 85.6% was achieved in the run using all the proposed feature types (SWC QF SSS), bringing approximately 27% error reduction compared to the Bag-of-Words features only. This result, obtained by the SVM based classifier, is higher than those reported in the previous researches (Brown 2003; Hacioglu et al. 2003; Li 2002; Li et al. 2002; Zhang et al. 2003), for the same training and test data collection. A complete list of test questions misclassified using the proposed method is presented in Appendix in Tables 10, 11 and 12.

Table 8 shows the examples of questions misclassified in a BOW approach and correctly classified using the set of new feature types (SWC QF SSS) along with the assigned categories.

Table 7 The question classification accuracy for the first classification under the fine-grained categories using different set of feature types and number of training examples (1000-5500).

		Set of the New Feature Types			
	BOW	SWC QF	SWC SSS	QF SSS	SWC QF SSS
1000	66.8 %	71.0 %	70.6 %	70.4 %	71.6 %
2000	71.4 %	78.8 %	76.8 %	76.6 %	79.8 %
3000	75.0 %	81.0 %	78.2 %	79.6 %	82.4 %
4000	77.8 %	81.4 %	78.8 %	80.6 %	82.4 %
5500	**80.2 %**	84.6 %	84.2 %	83.8 %	**85.6 %**

Table 8 Examples of questions misclassified in a BOW and correctly classified using the set of new feature types (SWC QF SSS).

Question	**Correct Class.**	**BOW**	**SWC QF SSS**
What hemisphere is the Philippines in?	LOC:other	DESC:def	LOC:other
Material called linen is made from what plant?	ENTY:plant	ENTY:termeq	ENTY:plant
What is the speed hummingbirds fly?	NUM:speed	DESC:def	NUM:speed
Who was Abraham Lincoln?	HUM:desc	HUM:ind	HUM:desc
What is the Ohio state bird?	ENTY:animal	LOC:state	ENTY:animal

Table 9 Results of McNemar's Test.

New Feature Type/ Set of Feature Types	p-value
SSS	0.21
SWC	0.0963
QF SSS	0.000912
SWC QF	0.00068
QF	0.000277
SWC SSS	0.000192
SWC QF SSS	3.47e-06

To verify how significant was the improvement obtained with the new feature types the set of McNemar's tests (Dietterich 1997) comparing the baseline approach to the BOW extended with the different sets of the new feature types was performed. The test results proofed that there was an extremely statistically significant difference (the p-value of 3.47e-06) between the proposed method (full set of the feature types SWC QF SSS) and the baseline-approach. Table 9 presents the results of McNemar's test for the different feature types and set of feature types.

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 exploited in the standard Bag-of-Words approach. Additionally, using these features, the classifier could "learn faster", from a smaller set of training data; a similar accuracy to this obtained in the BOW approach using 5500 training questions, was achieved using a set of 2000 questions. Using the whole set of the presented feature types the classifier, achieved the result of 85.6 %, 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 (Brill 1995) and WordNet (Miller 1995).

The closer analyze of the misclassified questions revealed that some of them are the result of inconsequent labeling of questions in the dataset. For example, the questions " Where is Amsterdam?" from the training data and similar question " Where is Milan?" from the test data have different labels, "LOC:other" and "LOC:city ", respectively. Similarly, the test questions "What is the life expectancy for crickets?" and "What is the life expectancy for dollar bill?" are labeled "NUM:other " while the training question "What is the life expectancy of an elephant?" is labeled "NUM:period". In the test set, twelve questions with the inconsistent labels were discovered. In the run that used the corrected labels for these questions, the accuracy improvement of 1.2 % was achieved for the set all of feature types, bringing the accuracy rate to 86.8 %.

6 Conclusion

Question classification is of prime importance for question answering, and as the previous works demonstrated a system that is able to correctly classify a question with a detailed taxonomy and use this information to extract and verify answer candidates obtains higher overall accuracy. This paper presented a machine learning based approach to question classification task using 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 demonstrate 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 three feature types, a result of 85.6 % was achieved, bringing error reduction of 27 % compared to the BOW approach. A comparison with the state of the art systems has shown that using these features, the classifier was able to achieve higher accuracy than any other machine learning-based classifier. The additional advantage of this approach is

the fact that the new feature types were created using only the freely available tools like POS Tagger and WordNet and as such can be easily adapted to other question answering systems. Our future work includes further tests and refining of the introduced feature types, especially the Syntactic-Semantic Structure, which in our opinion, posses the potential to provide higher coverage of various question categories.

Acknowledgement

This work was partially supported by Grant from Ministry of Internal Affairs and Communications Strategic Information and Communications R&D Promotion Programme (SCOPE).

Reference

Allwein, E., Schapire, E. and Singer Y. (2000). " Reducing multiclass to binary: A unifying approach for margin classifiers." *Journal of Machine Learning Research*, pp. 1:113–141.

Bikel, D.M., Schwartz, R.L. and Weischedel, R.M. (1999). " An algorithm that learns what's in a name.." *Machine Learning*, *34* (1)–(3), pp. 211–231.

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

Brown, J.(2003). " Entity-Tagged Language Models for Question Classification in a QA System." http://www-2.cs.cmu.edu/~jonbrown/IRLab/Brown-IRLab.pdf.

Cardie, C., Ng, V., Pierce, D. and 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*, pp. 180–187.

Carlson, A., Cumby, C., Rosen, J. and Roth, D.(1999). " SNOW User's Guide." UIUC Tech report UIUC-DCS-R-99-210.

Cortes, C. and Vapnik, V. (1995). " Support-Vector Network." *Machine Learning*, *20* ,pp. 1–25.

Dietterich, T.(1997). " Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms." *Neural Computation*, *10* (7), pp. 1895–1924.

Durme, V., Huang, Y., Kupsc, A. and Nyberg, E.(2003). " Toward Light Semantic Processing for Question Answring." *HTL/NAACL Workshop on Text Meaning 2003.*

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

Hovy, E., Gerber, L., Hermajakob, U., Lin C. and Ravichandran D.(2003). " Towards Semantics-based Answer Pinpointing." *In Proceedings of the DARPA HLT conference.*

Joachims, T. (1998). " Text Categorization with Support Vector Machines: Learning with Many Relevant Features." *In Proceedings of European Conference on Machine Learning*, pp. 137–142.

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

Li, W. (2002). " Question Classification Using Language Modeling." *CIIR Technical Report.*

Li, X. and Roth, D. (2002). " Learning Question Classifiers." *In Proceedings of the 19th International Conference on Computational Linguistics*, pp. 556–562.

Marquez, L. (2000). "Seminar: Industrias de la lengua / La ingeniera Linguistica en la sociedad de la informacion." *Machine Learning and Natural Language Processing.*

Miller, G. (1995). " WordNet: a lexical database for English." *Communications of the ACM*, *38* (11), pp. 39–41.

Pasca, M.A. and 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*, pp. 366–374.

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

Roth, D. (1998). " Learning to Resolve Natural Language Ambiquities. A Unified Approach." *In Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98)and of the 10th Conference on Innovative Applications in Artificial Inteligence (IAAI-98)*, AAAI Press, pp. 806–813.

Salton, G. and Buckley C.(1988). Term Weighting Approaches in Automatic Text Retrieval." *Information Processing and Management*, *24* , (5), pp. 513–523.

Skowron, M. and 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*, pp. 379–387.

Suzuki, J., Taira, H., Sasaki, Y. and Maeda E., (2004). Question Classification using HDAG Kernel." *Workshop on Multilingual Summarization and Question Answering 2003, post-conference workshop in conjunction with ACL-2003*, pp. 61–68.

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

Witten, I. and Frank, E. (2000). " Data Mining: Practical machine leaning tools with Java implementations." Morgan Kaufmann.

Vapnik, V. (1995). " The Nature of Statistical Learning Theory." Springer.

Voorhees, E. (1999). " The TREC-8 Question Answering Track Report." *In Proceedings of the 8th Text Retrieval Conference (TREC8)*, pp. 77–82.

Voorhees, E. (2000). " Overview of the TREC-9 Question Answering Track Report." *In Proceedings of the 9th Text Retrieval Conference (TREC9)*, pp. 71–80.

Voorhees, E. (2001). " Overview of the TREC 2001 Question Answering Track Report." *In Proceedings of the 10th Text Retrieval Conference (TREC10)*, pp. 157–165.

Zhang, D. and Lee, W.S. (2003). " Question Classification using Support Vector Machines." *Proceedings of the 26th ACM SIGIR*, pp. 26–32.

Appendix

Tables 10, 11 and 12 present a complete list of the misclassified test questions along with the original and assigned category using the proposed method (Bag-of-Words extended with the whole set of the proposed feature types: Subordinate Word Category, Question Focus and Syntactic-Semantic Structure).

Table 10 List of misclassified test questions along with the original and assigned category using the proposed method (1-24).

#	Test Question	Original Category	Assigned Category
1	What county is Modesto California in?	LOC:city	LOC:other
2	What is the life expectancy for crickets?	NUM:other	NUM:period
3	What metal has the highest melting point?	ENTY:substance	LOC:mount
4	What is Valentine's Day?	DESC:def	DESC:desc
5	Where are the Rocky Mountains?	LOC:mount	LOC:other
6	What birthstone is turquoise?	ENTY:substance	DESC:def
7	What is a group of turkeys called?	ENTY:animal	HUM:gr
8	What is done with worn or outdated flags?	DESC:desc	ENTY:cremat
9	Where is Milan?	LOC:city	LOC:other
10	What is the scale?	DESC:def	ENTY:other
11	What is the melting point of copper?	NUM:other	DESC:def
12	What is the electrical output in Madrid Spain?	ENTY:other	DESC:def
13	What county is Phoenix in?	LOC:city	LOC:other
14	What soviet seaport is on the Black Sea?	LOC:other	LOC:city
15	What is Hawaii's state flower?	ENTY:plant	LOC:state
16	What mineral helps prevent osteoporosis?	ENTY:substance	DESC:def
17	What is the diameter of a golf ball?	NUM:dist	ENTY:other
18	What is the earth's diameter?	NUM:dist	DESC:def
19	How wide is the Milky Way galaxy?	NUM:dist	DESC:manner
20	What was the first satellite to go into space?	ENTY:product	LOC:other
21	What position did Willie Davis play in baseball?	HUM:title	HUM:gr
22	What is the name of Roy Roger's dog?	ENTY:animal	HUM:ind
23	In the late 1700's British convicts were used to populate which colony?	LOC:other	ENTY:other
24	What is natural gas composed of?	ENTY:substance	LOC:other

Table 11 List of misclassified test questions along with the original and assigned category using the proposed method (25-48).

#	Test Question	Original Category	Assigned Category
25	What French ruler was defeated at the battle of Waterloo?	HUM:ind	HUM:gr
26	What is the birthstone for June?	ENTY:substance	DESC:def
27	What is the sales tax in Minnesota?	ENTY:other	DESC:def
28	What is the distance in miles from the earth to the sun?	NUM:dist	ENTY:termeq
29	What was the most popular toy in 1957?	ENTY:product	HUM:gr
30	What is the name of the satellite that the Soviet Union sent into space in 1957?	ENTY:product	HUM:gr
31	How much does the human adult female brain weigh?	NUM:weight	NUM:count
32	What is the longest major league baseball-winning streak?	ENTY:other	HUM:gr
33	What are the houses of the Legislative branch?	ENTY:other	DESC:def
34	What imaginary line is halfway between the North and South Poles?	LOC:other	ENTY:other
35	What is the criterion for being legally blind?	ENTY:other	DESC:def
36	What is the depth of the Nile river?	NUM:dist	LOC:other
37	Mexican pesos are worth what in US dollars?	NUM:money	LOC:state
38	What is strep throat?	DESC:def	LOC:other
39	What is the life expectancy of a dollar bill?	NUM:other	NUM:period
40	What are the types of twins?	ENTY:other	NUM:perc
41	What does the technical term mean?	ABBR:exp	DESC:def
42	What is the conversion rate between dollars and pounds?	NUM:money	NUM:other
43	What is a group of frogs called?	ENTY:animal	HUM:gr
44	What is the name of William Penn's ship?	ENTY:veh	HUM:ind
45	What is the melting point of gold?	NUM:other	DESC:def
46	What was President Lyndon Johnson's reform program called?	ENTY:event	HUM:ind
47	What is the murder rate in Windsor Ontario?	NUM:perc	NUM:other
48	Name a stimulant.	ENTY:dismed	HUM:ind

Table 12 List of misclassified test questions along with the original and assigned category using the proposed method (49-72).

#	Test Question	Original Category	Assigned Category
49	What is the Illinois state flower?	ENTY:plant	LOC:state
50	What is the width of a football field?	NUM:dist	HUM:gr
51	What is the only artery that carries blue blood from the heart to the lungs?	ENTY:body	DESC:desc
52	How often does Old Faithful erupt at Yellowstone National Park?	NUM:other	DESC:manner
53	What is the elevation of St Louis?	NUM:dist	DESC:desc
54	What is the length of the coastline of the state of Alaska?	NUM:dist	LOC:state
55	What is mad cow disease?	DESC:def	ENTY:dismed
56	What is the proper name for a female walrus?	ENTY:animal	HUM:ind
57	How long is the Columbia River in miles?	NUM:dist	NUM:period
58	What is the state flower of Michigan?	ENTY:plant	LOC:state
59	What are Canada's territories?	LOC:other	ENTY:animal
60	What monastery was raided by Vikings in the late eighth century?	ENTY:other	NUM:date
61	What is the name given to the Tiger at Louisiana State University?	ENTY:animal	HUM:ind
62	What gasses are in the troposphere?	ENTY:substance	DESC:def
63	What is the active ingredient in baking soda?	ENTY:food	ENTY:substance
64	How cold should a refrigerator be?	NUM:temp	DESC:manner
65	What did Jesse Jackson organize?	HUM:gr	ENTY:other
66	What is the National Park in Utah?	LOC:other	DESC:def
67	What type of polymer is used for bulletproof vests?	ENTY:other	ENTY:food
68	What was the name of the first US satellite sent into space?	ENTY:product	HUM:ind
69	What precious stone is a form of pure carbon?	ENTY:substance	DESC:def
70	What kind of gas is in a fluorescent bulb?	ENTY:substance	ENTY:animal
71	What is the source of natural gas?	ENTY:other	LOC:other
72	What is the birthstone of October?	ENTY:substance	DESC:def

Marcin Skowron: Marcin Skowron was born in 1976 in Gdynia, Poland. He received his MS degree from the University of Gdansk. He was a research student at Otaru University of Commerce, and since 2003 he is studying towards his Ph.D. degree at Graduate School of Information Science and Technology, Hokkaido University, Japan. His research interest includes natural language processing, question-answering, information retrieval, and knowledge acquisition. He is a member of the IEEE and AAAI.

Kenji Araki: Kenji Araki was born in 1959 in Otaru, Japan. He received B.E., M.E. and Ph.D. degrees in electronics engineering from Hokkaido University, Sapporo, Japan in 1982, 1985 and 1988, respectively. In April 1988, he joined Hokkai-Gakuen University, Sapporo, Japan. He was a professor of Hokkai-Gakuen University. He joined Hokkaido University in 1998 as an associate professor of the Division of Electronics and Information Engineering. He was a professor of the Division of of Electronics and Information Engineering of Hokkaido University from 2002. Now he is a professor of the Division of Media and Network Technologies of Hokkaido University. His interest is natural language processing, spoken dialogue processing, machine translation and language acquisition. He is a member of the IEICE, the IPSJ, the JSAI, the JCSS, the ACL, the IEEE and the AAAI.

(Received November 30, 2004)
(Revised February 26, 2005)
(Rerevised February 26, 2005)
(Accepted July 10, 2005)