A Re-examination of Question Classification

Håkan Sundblad
Linköpings universitet
581 83 Linköping
Sweden
hakjo@ida.liu.se

Abstract

This paper presents a re-examiniation of previous work on machine learning techniques for questions classification, as well as results from new experiments. The results suggest that some of the work done in the field have yielded biased results. The results also suggest that Naïve Bayes, Decision Trees and Support Vector Machines perform on par with each other when faced with actual users' questions.

1 Introduction

One of the most important factors for a question answering system to succeed is the ability to correctly identify the expected answer's semantic type (Moldovan et al., 2002).

This paper presents results from an evaluation of five different machine learning approaches to question classification (Naïve Bayes, k Nearest Neighbours, Decision Tree Learning, Sparse Network of Winnows, and Support Vector Machines). The paper also presents a review of earlier work on question classification as well as results from experiments using slightly different data than used in previous work. The reason for re-examining the results from previous work is that only performance in terms of accuracy has been reported in the literature. No significance testing has been made to see if there really is a difference in results between learners. Furthermore, the data used in many of the experiments have been submitted to manual selection, and also the test data is slightly different from the training data.

2 The Question Classification Task

Question classification can loosely be defined as the task of given a question (represented by a set of features), assign the question to a single or a set of categories (answer types). Adopting the formal definition of text categorization (Sebastiani, 2002) to the problem of question classification, the task can be defined as follows: Question classification is the task of assigning a boolean value to each pair $\langle q_j, c_i \rangle \in \mathcal{Q} \times \mathcal{C}$, where $\mathcal{Q}$ is the domain of questions and $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ is a set of predefined categories. The task therefore requires a taxonomy of answer types according to which questions should be categorized on the on hand, and a means for actually making this classification on the other.

3 Previous Work

Radev et al. (2002) experiment with machine learning for question classification using decision rule learning with set-valued features. This is a standard decision tree/rule approach that has been augmented in that instead of being restricted to features with single values, the values can also be a set of values. The answer type taxonomy consists of 17 types, and the training data is TREC-8 and TREC-9 data. Testing data is TREC-10. In the experiment, questions are represented by 13 features, 9 of which are semantic features based on WordNet.

Li and Roth (2002) use a Sparse Network of Winnows (SNoW) to classify questions with respect to their expected answer type. The taxonomy consists of 6 coarse and 50 fine semantic classes. The training corpus used consists of 5,500 questions. Some

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 394–397

of these are manually constructed, while other stems from the TREC-8 and TREC-9 conferences. The test corpus comprise 500 questions from the TREC-10 conference. The input to the classifiers is a list of features. The features used were words, part-of-speech tags, chunks, named entities, head chunks (e.g. the first noun chunk in a sentence), and semantically related words (words that often occur with a specific question class). Apart from these primitive features, a set of operators were used to compose more complex features.

Zhang and Lee (2003) used the same taxonomy as Li and Roth (2002), as well as the same training and testing data. In an initial experiment they compared different machine learning approaches with regards to the question classification problem: Nearest Neighbors (NN), Naïve Bayes (NB), Decision Trees (DT), SNoW, and Support Vector Machines. The feature extracted and used as input to the machine learning algorithms in the initial experiment was bag-of-words and bag-of-*n*grams (all continuous word sequences in the question). Questions were represented as binary feature vectors. In a second experiment the linear kernel of the SVM was replaced with a tree kernel developed by the authors.

Suzuki et al. (2003b) used a SVM with a hierarchical directed acyclic graph kernel (Suzuki et al., 2003a) for the question classification problem. The answer type taxonomy used consists of 150 different types. The corpus used was in Japanese and consisted of 1011 questions from NTCIR-QAC, 2000 questions of CRL-QA data, and 2000 other questions reported to be of TREC-style (Suzuki et al., 2002). After removing answer types with too few (less than 10) examples, a total of 68 answer types were actually used.

Hacioglu and Ward (2003) used a SVM with error correcting codes to convert the multi-class classification problem into a number of binary ones. In essence each class is assigned a codeword of 1's and -1's of length m, where m equals or is greater than the number of classes. This splits the multi-class data into m binary class data. Therefore, m SVM classifiers can be designed and their output combined. The SVM:s also used linear kernels. The same taxonomy, training and testing data was used as in Li and Roth (2002)

4 Method

In order to compare the five algorithms (Naïve Bayes (NB), *k* Nearest Neighbours (*k*NN), Decision Tree Learning (DT), Sparse Network of Winnows (SNoW), and Support Vector Machines (SVM)) significance testing have been used. Significance scores can not be found in any previous work on question classification and hence it is difficult to draw any real conclusions from this work. For present purposes the micro and macro sign tests established by Yang and Liu (1999) have been used. Thses were originally developed for the text categorization task, but as question classification bears many resemblances and can be seen as a special case of text categorization.

The taxonomy used is the taxonomy proposed by Li and Roth (2002). This taxonomy has been chosen since it is the most frequently used one in earlier work in the field (Li and Roth, 2002; Zhang and Lee, 2003; Hacioglu and Ward, 2003). The corpora used is both the corpus constructed and tagged by Li and Roth (2002), as well as a newly tagged corpus extracted from the AnswerBus logs. AnswerBus is a question answering system that has been online and logged real users questions. The AnswerBus corpus consists of 25,000 questions. For present purposes 2,000 questions have been selected and tagged according to the aforementioned taxonomy. Questions are in all experiments treated as a bag-of-words and represented as binary feature vectors.

The results will be reported in terms of micro- and macro-averaged precision, recall and *F*-score. Micro-averaged precision and recall are dominated by the large categories, whereas macro-averaged precision and recall illustrate how well a classifier performs across all categories. Micro-averaged precision is denoted as π^{μ}, macro-averaged precision as π^{M}, micro-averaged recall as ρ^{μ}, and macro-averaged recall as ρ^{M}. Combined measures for micro-averaged results is denoted as F_1^{μ} while the corresponding macro-averaged measure is denoted as F_1^{M}.

Performance is for the purpose of this paper seen as solely related to accuracy in terms of precision and recall. The learning and classification speed of the algorithms are ignored.

5 Experiment 1

The first experiment is intended to be a straightforward re-examination of previous work to establish what differences in performance there really are between machine learners. This experiment has been done under two different settings. First, we have used the corpus originally developed by (Li and Roth, 2002), but since the test corpus used consists of questions solely from TREC-10 and the TREC conferences have a specific agenda the test corpus might be slightly different from the training data. Therefore, a second setting was used where the questions from the training and test corpora were pooled together and a randomized test corpus was extracted. This will be refered to as the repartitioned corpus. The performance of the different learners on setting 1 can be found in table 1, setting 2 in table 2 while significance testing between the learners is shown in table 3

Classifier	π^{μ}	ρ^{μ}	F_1^{μ}	π^{M}	ρ^{M}	F_1^{M}
kNN	.6720	.6720	.6720	.6002	.5028	.5472
NB	.7162	.7120	.7141	.5979	.5775	.5875
SNoW	.7642	.7535	.7588	.7080	.6413	.6730
DT	.7780	.7780	.7780	.7460	.6819	.7125
SVM	.8149	.8100	.8124	.7574	.6655	.7085

Table 1: Performance of classifiers on original TREC data.

Classifier	π^{μ}	ρ^{μ}	F_1^{μ}	π^{M}	ρ^{M}	F_1^{M}
kNN	.6285	.6260	.6273	.6196	.5557	.5859
NB	.6713	.6700	.6707	.5970	.6498	.6223
SNoW	.6633	.6593	.6613	.6511	.4999	.5656
DT	.7194	.7180	.7187	.6381	.6202	.6290
SVM	.7820	.7820	.7820	.8122	.7112	.7584

Table 2: Performance of classifers on repartitioned TREC data.

As can be seen in table 1 and 2 the performance of the different learning algorithms with regards to micro-averaged precision and recall is at best equal to and in most cases worse on the repartitioned data than on the original data. When it comes to macro-averaged precision and recall the results are more varied.

In table 3 we can find differences when comparing the algorithms with regards to significant differences in performance. In the table, "$<$" means a

		Original		Repartitioned	
sysA	sysB	s-test	S-test	s-test	S-test
kNN	NB	$<$	-	$<$	$\ll$
kNN	SNoW	$\ll$	$<$	-	-
kNN	DT	$\ll$	$\ll$	$\ll$	$\ll$
kNN	SVM	$\ll$	$\ll$	$\ll$	$\ll$
NB	SNoW	$<$	-	-	-
NB	DT	$\ll$	$\ll$	$\ll$	-
NB	SVM	$\ll$	$\ll$	$\ll$	$\ll$
SNoW	DT	$<$	-	$\ll$	-
SNoW	SVM	$\ll$	-	$\ll$	$\ll$
DT	SVM	$\ll$	-	$\ll$	$\ll$

Table 3: Sigificance testing of classifiers on both original and repartitioned TREC data.

significantly on the .05 level, "$\ll$" and "$\gg$" means a difference on the .01 level. NB $<$ SNoW should be read as NB performs significantly worse than SNoW on the .05 level. The column "s-test" means micro sign test, and "S-test" means macro sign test. It is interesting to note that where there were no significant differences in performance on the original corpus there now are to some extent differences on the repartitioned corpus and also the other way around to a smaller extent. This might be an indication that the training and test corpora in fact are not balanced in the original setting, and some of the results reported in previous work is somewhat biased.

6 Experiment 2

To further investigate the performance of different machine learners in the face of a corpus consisting of actual users' questions a second experiment was conducted. As mentioned earlier, in this setting 2,000 questions from the AnswerBus logs are used, but everything else remains the same as in experiment 1. Results in terms of performance is found in table 4 and significance testing is found in table 5.

Classifier	π^{μ}	ρ^{μ}	F_1^{μ}	π^{M}	ρ^{M}	F_1^{M}
kNN	.7159	.7076	.7117	.6088	.5252	.5639
NB	.8000	.7953	.7976	.6959	.6605	.6778
SNoW	.6913	.6588	.6746	.6138	.7702	.6831
DT	.8143	.7953	.8047	.6871	.6583	.6724
SVM	.8176	.8128	.8152	.7319	.6499	.6885

Table 4: Performance of classifiers on AnswerBus data.

As can be seen in table 4 the performance in terms of micro-averaged precision and recall is higher on the AnswerBus corpus than on any of the TREC

corpora. When it comes to macro-averaged performanve the results are more varied and it is hard to draw any clear conclusions.

		Original	
sysA	sysB	s-test	S-test
kNN	NB	$\ll$	$<$
kNN	SNoW	-	-
kNN	DT	$\ll$	$\ll$
kNN	SVM	$\ll$	$\ll$
NB	SNoW	$\gg$	-
NB	DT	-	-
NB	SVM	-	-
SNoW	DT	$\ll$	-
SNoW	SVM	$\ll$	-
DT	SVM	-	-

Table 5: Sigificance testing of classifiers on AnswerBus data.

In terms of significant differences between classifiers, the results from the AnswerBus corpus deviates from what could have been expected given the results on the TREC corpora. It seems than Naïve Bayes, Decision Trees and Support Vector Machines are on par with each other, while k Nearest Neighbours and Sparse Network of Winnows are sigificantly worse in terms of performance.

7 Conclusions

The results in this paper indicate that some of the results found in previous work (Li and Roth, 2002; Zhang and Lee, 2003; Hacioglu and Ward, 2003) on question classification might be incorrect due to an unbiased training and test corpus. This bias stems from the fact that the training corpus is derived exclusively from TREC-10 data, while the training data stems from other sources. Since the TREC conferences have an explicit agenda that shifts from year to year this is perhaps no surprise. In relation to this, TREC material is maybe not the best source of information if one is interested in how different machine learners might perform on actual user data.

8 Future Work

The results from experiment 2 in this paper stems from a corpus of 2,000 questions. We will go on to categorize 3,000 more questions from the AnswerBus logs and run the learners on this data in order to get even more accurate results. This work is well on the way.

We will also go on to make a deeper analysis of exactly which questions that pose problems for learning algorithms. Such work has not been reported in the literature thus far.

References

K. Hacioglu and W. Ward. 2003. Question classification with support vector machines and error correcting codes. In *Proceedings of HLT-NACCL 2003*.

X. Li and D. Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 556–562.

D. Moldovan, M. Paşca, S. Harabagiu, and M. Surdeanu. 2002. Performance issues and error analysis in an open-domain question answering system. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Philadelphia.

D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. 2002. Probabilistic question answering on the web. In *Proceedings of the eleventh international conference on World Wide Web (WWW2002)*, Hawaii.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March.

J. Suzuki, Y. Sasaki, and E. Maeda. 2002. SVM answer selection for open-domain question answering. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*.

J. Suzuki, T. Hirao, Y. Sasaki, and E. Maeda. 2003a. Hierarchical directed acyclic graph kernel: Methods for natural language data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*, pages 32–29.

J. Suzuki, H. Taira, Y. Sasaki, and E. Maeda. 2003b. Question classification using HDAG kernel. In *The ACL 2003 Workshop on Multilingual Summarization and Question Answering*.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49, Berkeley, CA.

D. Zhang and W. S. Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 26–32.