Research of influence of lemmatisation in questions classification in question answering systems

Fedor Vityugin

National Research University Higher School of Economics,
Nizhny Novgorod, Russia

Abstract. This work is a report on an attempt to research on influence of lemmatization in questions classification . The result of the research is answer of question "How influence lemmatisation on question classification machine learning method which can be used in future works for classify texts?"

1 Goal

The purpose of research is to find the best method for classify questions in Russian from community question answering service; can it used methods of machine learning which have a good results in English; how influence lemmatization on question classification machine learning method which can be used in future works for classify texts, decide which method will be at the base of future question-answering system and .

2 Assignment

Assignment was to research machine learning methods and tested it for classify questions in natural language from Russian community question answering system and compare methods with using lemmatization. Research process exploited the following machine learning methods: Naive Bayes and Maximum Entropy.

3 Evaluation and results

Texts were collected from among CQA-service "Questions and Answers" (otvety.google.ru). All questions have been classified into six major categories listed in previous table. This collection was used for training the classifier machine learning methods. For the test set was randomly selected 150 questions that were not used for training the classifier. Details of training and the test described in next table.

For comparison with the questions in English was taken marked question collection from TREC. Collection was preprocessed. The size and composition

Table 1. Russian set

Collection	ABBR	ENTY	DESC	HUM	LOC	NUM	All
Training set	500	500	500	500	500	500	3000
Test set	5	20	30	37	29	29	150

Table 2. English set

Collection	ABBR	ENTY	DESC	HUM	LOC	NUM	All
Training set	500	500	500	500	500	500	3000
Test set	23	52	18	5	5	47	150

of the resulting sample for study in English corresponds to the sample in Russian. For the test set were also randomly selected 150 questions that were not used for training the classifier. Details of training and the tests described in next table.

Results of studies using unigrams feature vector and lemmatisation described in next table.

Table 3. Results

Features	English		Russian		Russian + lemmatisation	
	NB	MaxEnt	NB	MaxEnt	NB	MaxEnt
Unigram	56.67	58.00	27.33	28.00	30.67	30.67

4 Future work

Machine learning techniques perform well for classifying questions. I believe the accuracy of the system could be still improved. Below is a list of ideas could help the classification.

Bigger dataset: the training dataset in the order of millions will cover a better range of twitter words and hence better unigram feature vector resulting in an overall improved model. This would vastly improve upon the existing classifier results.

Another machine learning methods: there are many another machine learning methods which have a better results in related works (for example SVM).

Weighted unigram: in this approach, some words must have been sense to weight the one of classes more than other words while trying to classify the class of a question.

This work will be part of research in the "Adapting language material RNC for the electronic textbook "Russian as a foreign language" carried out within The National Research University Higher School of Economics Academic Fund Program in 2013, grant No 13-05-0031.

References

1. Lehnert, W., (1986). A conceptual theory of question answering. In B. J. Grosz, K. Sparck Jones, and B. L. Webber, editors, Natural Language Processing, pages 651. Kaufmann, Los Altos, CA.
2. Li, X. and Roth, D. (2002), Learning Question Classifiers, In Proceedings of COLING 2002.
3. Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R, Goodrum, R, Rus, V., The Structure and Performance of an Open-Domain Question Answering System, (2000), In Proceedings of the Conference of the Association for Computational Linguistics (ACL-2000), p 563
4. Manning C. D. and Schutze H. Foundations of statistical natural language processing. MIT Press, 1999.
5. Nigam K. , Lafferty J., and Mccallum A. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 6167, 1999.