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 Related works

In order to find a correct answer to a users question, we need to first know what to look for in our large collection of documents. The type of answer required is related to the form of the question, so knowing the type of a question can provide constraints on what constitutes relevant data, which helps other modules to correctly locate and verify an answer.

The question type classification component is therefore a useful, if not essential component in a QA system, as it provides significant guidance about the nature of the required answer.

Many researchers have proposed various different taxonomies for question classification. Wendy Lehnert, for example, proposed a conceptual taxonomy with 13 conceptual classes [1] back in 1986. More recently, Li and Roth propose a multi- layered taxonomy [2], which has 6 coarse classes (ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION and NUMERIC VALUE), and 50 fine classes. Moldovan et al. [4] provide another set of question classes and subclasses along with corresponding answer types, based on the 200 question used in TREC 8.

Table 1. Question classes

Lowlevel	*Highlevel*
Abbreviation (ABBR)	Abbreviation, expansion
Entyty (ENTY)	Animal, body, color, creation, currency, disease/medical, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
Description (DESC)	Definition, description, manner, reason
Human (HUM)	Description, group, individual, title
Location (LOC)	City, country, mountain, other, state
Numeric value (NUM)	Code, count, data, distance, money, order, other, percent, perion, speed, temperature, size, weight

The focus of a question has been defined by Moldovan et al [3] to be a word or sequence of words which indicate what information is being asked for in the question.

A statistical approach might again make use of n-grams to identify likely focus words of questions. Such an approach would require a training corpus of questions with known question foci to be developed, which may be prohibitively expensive in terms of time and effort.

The process of extracting keywords could be performed with the aid of standard techniques such as named entity recognition, stop-word lists, and part-of-speech taggers, along with a set of ordered heuristics, such as those described in [4]. Based on the work in [3], all words satisfying any of the following 8 heuristics would be chosen as keywords:

1. For each quoted expression in a question, all non-stop words in the quotation;
2. Words recognized as proper nouns (using named-entity recognition);
3. Complex nominals and their adjective modifiers;

4. All other complex nominals;
5. All nouns and their adjectival modifiers;
6. All other nouns;
7. All verbs;
8. The question focus.

The set of question keywords is sorted by priority, so if too many keywords are extracted from the question, only the first N words are passed onto the next module. N would be a configurable value that could be tuned, based on an evaluation of performance with different numbers of keywords for information retrieval.

4 Feature Reduction

Questions in natural language has many unique properties. Following properties was taken for reduce the feature space.

URLs: users very often include links in their questions. An equivalence class is used for all URLs. Links convert in URL like "http://goo.gl/FTFnJ" to the token "URL".

Stop-words: there are a lot of stop words or filler words that used in a questions which does not indicate any class and hence all of these are filtered out. The complete list of stop words can be found at http://goo.gl/3pH24.

Repeated letters: community question-answering services contain very casual language. For example, if you search "hungry" with an arbitrary number of us in the middle (e.g. huuuungry, huuuuuuungry, huuuuuuuuuungry) on service, there will most likely be a nonempty result set. In this work used preprocessing so that any letter occurring more than two times in a row is replaced with two occurrences. In the samples above, these words would be converted into the token "huungry".

Feature vector After preprocessing the training set data which consists of 500 questions of each class, was computed the feature vector:

- **Unigrams** List of features was formed at the end of preprocessing where each of the features has equal weights.
- **Lemmatisation** Russian questions processed by morphological analyzer pymorphy2 for getting lemmas of unigrams. unigrams. List of features was formed at the end of preprocessing where each of the features has equal weights.

5 Data

Questions in community question-answering systems have many unique attributes, differentiates this work from previous research.

Length: The maximum length of questions in CQA-service otvety.google.ru is 120 characters. This research focused on classifying all-size questions.

Language model: CQA-service users post messages from many different media, including their cell phones. The frequency of misspellings and slang in this question is much higher than in other domains.

Domain: CQA-service users post messages about a variety of topics unlike other sites which are tailored to a specific topic. Question collection for this research includes questions from different topics, it means that it was as questions from closed-domain like medicine, history, gadgets as from cooking, relationships, etc.

Language: in this work researched questions in Russian, this differs from a large percentage of past research, which focused on questions in another languages.

6 Machine learning methods

In work was tested classifiers namely Naive Bayes and Maximum Entropy.

Naive Bayes. Naive Bayes is a simple model which works well on text categorization [4]. In work was used a multinomial Naive Bayes model. Class c is assigned to question d, where

$$c^* = argmax_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c) sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

In this formula, f represents a feature and $n_i(d)$ represents the count of feature f_i found in question d. There are a total of m features.

Parameters P(c) and $P(f|c)$ are obtained through maximum likelihood estimates, and add-1 smoothing is utilized for unseen features. In project was used the Python based Natural Language Toolkit library to train and classify using the Nave Bayes method.

Maximum Entropy. The idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint [5]. MaxEnt models are feature-based models. In a two class scenario, it is the same as using logistic regression to find a distribution over the classes. MaxEnt makes no independence assumptions for its features, unlike Naive Bayes. The model is represented by the following:

$$P_{ME}(c|d,\lambda) = \frac{exp[E_i \lambda_i f_i(c|d)]}{E'_c exp[E_i \lambda_i f_i(c|d)}$$

In this formula, c is the class, d is the question, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the λ'_is so as to maximize the conditional probability.

The Python NLTK library was used to train and classify using the Maximum Entropy method. Conjugate gradient ascent was used for training the weights. Theoretically, MaxEnt performs better than Naive Bayes because it handles feature overlap better. However, in practice, Naive Bayes can still perform well on a variety of problems [5].

7 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.

Table 2. 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

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

Table 4. 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

8 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.