Language Model Based Query Classification

Andreas Merkel and Dietrich Klakow

Spoken Language Systems
Saarland University
D-66123 Saarbrücken, Germany

Abstract. In this paper we propose a new way of using language models in query classification for question answering systems. We used a Bayes classifier as classification paradigm. Experimental results show that our approach outperforms current classification methods like Naive Bayes and SVM.

1 Introduction

Unlike most current information retrieval systems, which just return documents to keyword queries, a question answering (QA) system tries to return an accurate answers to natural language questions. So, the step of retrieving relevant documents is just a part of a complex QA system. In order to simply find the one correct answer such a system has to "understand" the meaning of a question. For example, if the question is "*In what country is Luxor?*", the answer should be a country name and not a date. That means the question has to be analyzed in a separate step. There, the question is classified into several semantic categories. This categorization not only helps to find and verify the answer by reducing the search space but also may determine the search strategies for further QA modules ([1]).

Normally, most QA systems just use no more than 20 coarse classes for classification, but in this paper we decided to use the taxonomy proposed by [1] which takes 6 coarse and 50 fine grained classes into account We used the fine grained classification in our experiments because they showed that they are more useful to locate and verify answers. Based on this categorization we used a Bayes classifier with language models as classification paradigm. We show that this approach outperforms systems in current literature.

2 Methods

2.1 Language Models for Classification

Next we have to introduce a suitable classification framework. Ponte and Croft suggested in [3] language models to information retrieval from text collections and showed that they can outperform more traditional methods. We want to propose to use this technique also in the classification of questions.

As already mentioned we used the Bayes classifier as categorization paradigm. The Bayes classifier is defined by

$$\hat{c} = \text{argmax}_c P(Q|c)P(c) \tag{1}$$

which provides minimum error rate if all probabilities are exactly known. Here, $P(Q|c)$ is the probability of a question Q and a given semantic class c. $P(c)$ is the prior for that class. The probability of $P(Q|c)$ is a language models trained on the class c. In case of unigram language models $P(Q|c)$ is calculated as the product of $P(w|c)$ for all w in Q. The major advantage of the language modeling (LM) approach is that a huge amount of techniques are available to estimate and smooth probabilities even if there is just little training data available. On average, there are only about 100 training questions per question type.

Unlike in previous work ([5]) we do not assume a uniform prior. $P(c)$ can be considered as a unigram language model on semantic classes. As all classes are seen at least four times and therefore sufficiently often, there is no need for smoothing at all and relative frequencies can be used to estimate the language model.

Next we will describe how to estimate the probability $P(w|c)$. It is essential to avoid zero probabilities because that would exclude specific terms from the classification. Hence language model smoothing methods come into play.

2.2 Absolute Discounting

Absolute discounting and its variants are the most popular smoothing techniques in speech recognition. It is defined by

$$P_\delta(w|c) = \frac{\max\left(N(w,c) - \delta, 0\right)}{\sum_w N(w,c)} + \frac{\delta B}{\sum_w N(w,c)} P_{BG}(w|c) \tag{2}$$

where $N(w,c)$ is the frequency of observations of the term w together with class c. $P_{BG}(w|c)$ is a background model used for smoothing, δ is the smoothing parameter and B denotes how often $N(w,c)$ is larger than δ.

2.3 Dirichlet Prior

Using a Dirichlet prior results in

$$P_\mu(w|c) = \frac{N(w,c) + \mu P_{BG}(w|c)}{\sum_w N(w,c) + \mu} \tag{3}$$

where μ is a smoothing parameter to be determined on the development data.

2.4 Linear Interpolation

Linear interpolation was first introduced by Jelinek and Mercer [2] and hence some people refer to it also as Jelinek-Mercer smoothing. It is defined by

$$P_\lambda(w|c) = (1-\lambda)\frac{N(w,c)}{\sum_w N(w,c)} + \lambda P_{BG}(w|c) \quad (4)$$

where $N(w,c)$ are frequencies on the training data, λ is a smoothing parameter to be tuned on the development data.

3 Experiments

3.1 Data

For classification we used 6 coarse and 50 fine grained classes as defined in [1]. So, for example, the coarse class *LOCATION* contains the fine classes *city, country, mountain, other and state* whereas *HUMAN* consists of *group, individual, title and description* and so on. As training data for our experiments we used the 5,500 questions provided by the Cognitive Computing Group at University of Illinois at Urbana Champaign[1]. For the evaluation task we used the TREC 10 [2] dataset consisting of 500 questions. Both training and test sets are labeled with the corresponding coarse and fine classes.

3.2 Results

As some examples for our experiments Fig. 1 shows the results for the classification with linear interpolation and absolute discounting as smoothing methods. The x-axis shows the interpolation weight and on the y-axis the accuracy is printed.

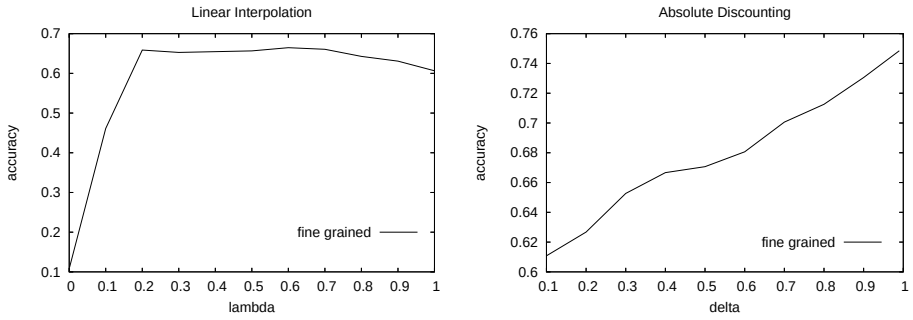

Fig. 1. Accuracy vs. interpolation weight for linear interpolation (a) and absolute discounting (b)

The plot on the left hand side (a) shows the linear interpolation smoothing method. It has a maximum near $\lambda = 0.2$ and is relatively independent to the

[1] http://l2r.cs.uiuc.edu/~cogcomp/Data/QA/QC/
[2] http://trec.nist.gov/

interpolation weight. The graph on the right hand side shows the absolute discounting method. In contrast to the linear interpolation it strongly depends on the discounting parameter and has it maximum at $\delta = 1$.

Table 1 compares results from [4] with the proposed LM approach for various machine learning algorithms. They used two different feature sets (bag-of-words and bigram features), so we always print the better result. Our LM approach utilizes optimized bigram features. In particular, we used a log-linear interpolation between a bigram and a unigram distribution.

Table 1. Comparison of various algorithms (Naive Bayes, ... SVM) investigated in [4] with the proposed language model based approach, denoted by LM in the table.

Algorithm	Accuracy	Error Rate
Naive Bayes	67.8%	±2.5%
Neural Network	68.8%	±2.5%
Decision Tree	77.0%	±2.1%
SVM	80.2%	±2.0%
LM	80.8%	±2.0%

The table shows that our approach is much better than Naive Bayes. This difference is probably due to the used smoothing techniques. The SVM is the best algorithm from [4] however it is outperformed by the LM approach by a small margin. But in terms of error rate it is not significantly better than SVM. The exact error rates are shown in the last column of the table.

4 Conclusion

In this paper we showed a language modeling approach for query classification based on a Bayes classifier. We experimented with different smoothing methods and various unigram and bigram models. As result we showed that our proposed approach outperforms current categorization methods. So it is significantly better than a classification with Decision Trees and as good as SVM.

References

1. Xin Li and Dan Roth. Learning Question Classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics (2002).*
2. Hermann Ney, Ute Essen and Reinhard Kneser On Structuring Probabilistic Dependencies in Stochastic Language Modeling. In *Computer Speech and Language 8 (1994) 1-38.*
3. Jay M. Ponte and Bruce Croft A Language Modeling Approach to Information Retrieval. In *Proceedings SIGIR (1998).*
4. Dell Zhang and Wee Sun Lee Question Classification using Support Vector Machines. In *Proceedings SIGIR (2003).*
5. Chengxiang Zhai and John Lafferty A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings SIGIR (2001).*