Question Classification with Log-Linear Models

Phil Blunsom
Department of Computer Science and Software Engineering
University of Melbourne
Victora 3010, Australia
pcbl@cs.mu.oz.au

Krystle Kocik, James R. Curran
School of Information Technologies, University of Sydney
NSW 2006, Australia
{kkocik,james}@it.usyd.edu.au

ABSTRACT

Question classification has become a crucial step in modern question answering systems. Previous work has demonstrated the effectiveness of statistical machine learning approaches to this problem. This paper presents a new approach to building a question classifier using log-linear models. Evidence from a rich and diverse set of syntactic and semantic features is evaluated, as well as approaches which exploit the hierarchical structure of the question classes.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**].

General Terms: Algorithms, Experimentation

Keywords: Maximum entropy, Question Classification, Question Answering, Machine Learning

1. INTRODUCTION

Research in Question Answering (QA) seeks to move beyond the existing keyword-based Information Retrieval (IR) approaches by providing one or more *exact answers to a question* from a large document collection. The syntactic and semantic interpretation of a question is crucial in a QA system. The most common approach to semantic interpretation is to classify the question into a closed set of *question types* (*qtype*) which describe the expected semantic category of the answer to the question.

Maximum Entropy (ME) or log-linear models [5] have been successfully applied to many Natural Language Processing (NLP) problems which require complex and overlapping features. Here we make use of this ability to incorporate syntactic and semantic information extracted from the questions. The result is a question classifier which significantly outperforms the state-of-the-art systems on the standard question classification test set [4].

2. LOG-LINEAR MODELS

Conditional log-linear models, also known as Maximum Entropy models, produce a probability distribution over multiple classes and have the advantage of handling large numbers of complex overlapping features. These models have the following form:

$$p(y|x,\lambda) = \frac{1}{Z(x|\lambda)} \exp\left(\sum_{k=1}^{n} \lambda_k f_k(x,y)\right) \quad (1)$$

where the f_k are feature functions of the observation x and the class label y. λ_k are the model parameters, or feature weights, and $Z(x|\lambda)$ is the normalisation function.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

FEATURE	DESCRIPTION
UNIGRAMS	all words in Q
BIGRAMS	all bigrams in Q
TRIGRAMS	all trigrams in Q
FBG	bigram of first 2 words in Q
FTG	trigram of first 3 words in Q
LENGTH	the length of the Q (in groups of 4)
POS	all POS tags in Q
CHUNK	all chunk tags in Q
SUPERTAGS	all CCG supertags in Q
NE	NE types in Q (by type)
T-WORD	target word
T-POS	target POS
T-CHUNK	target chunk tag
T-NE	target NE
T-SC	target supertag
T-CASE	target is lower, upper or titlecase
T-WORDNET	target in a WordNet lexfile
T-SEM	target in semantically related words
T-GAZ	target in gazetteer
FBGTGT	bigram of target and 1st word
FTGTGT	trigram of 1st 2 words and target
FBGWN	bigram of 1st word and target lexfile
FTGWN	trigram of 1st 2 words and target lexfile
PWTGT	target and previous word bigram
QUOTES	a (double) quoted string in Q
T-QUOTED	target within a quoted expression

Table 1: Extracted feature types.

In order to train the model we employ the common practice of defining a prior distribution over the model parameters and derive a maximum *a posteriori* (MAP) estimate from the training observations.

3. FEATURES

Features were derived from both lexical and syntactic information. Each question was parsed using the C&C CCG parser [1] with a model specifically created for parsing questions. This involved annotating questions from previous TREC competitions with their correct lexical categories and retraining the supertagging model.

The *target word*, also called the *question focus*, was found by traversing the CCG dependency graph produced by the C&C CCG parser. Kocik [3] developed and evaluated the dependency finding algorithm using 1000 Li and Roth training set questions which she annotated with their correct target word.

4. EXPERIMENTS

There are few data sets available for training machine learning approaches to question classification. Li and Roth [4] created the most frequently used data set. Their classification scheme, or *question ontology*, consists of 6 coarse-grained categories which are di-

FINE	P_1	P_2	P_3	Coarse P_1
ALL	86.6	91.8	94.4	92.0
NGRAMS	83.4	88.2	90.0	88.4
NO SEMANTIC	85.2	89.8	91.4	91.0
NO TARGET	83.4	89.6	91.2	92.0

Table 2: Evaluation of feature groups.

COARSE	P_1	P_2	P_3	P_4	P_5
Li & Roth	91.0	-	-	-	98.8
coarse	91.8	97.4	**99.2**	**99.8**	**100.0**
hierarchy	91.4	95.8	99.0	**99.8**	**100.0**
two-stage	**92.6**	**97.8**	98.8	99.2	99.6
flat	92.0	97.2	99.2	**99.8**	99.8

Table 3: Evaluation on coarse-grained labels.

vided unevenly into 50 fine-grained categories. The data set[1] consists of approximately 5,500 annotated questions for training and 500 annotated questions from TREC 10 for testing. The training questions were collected from four sources: 4,500 English questions collected by Hovy et al. [2], plus 500 manually created questions for rare qtypes and 894 questions from TREC 8 and TREC 9. We use the data in exactly the same manner as Li and Roth [4] in their original experiments.

We conducted two sets of experiments to investigate different aspects of the QC task. The first experiments aim to evaluate the contribution of each of our proposed feature types using a standard log-linear classification model, while the second experiments investigate whether the incorporation of hierarchical label information can assist the classification. Table 1 lists the feature types used by our classifier.

In evaluating our experiments we have used precision over the top n labels returned from the classifier. In this case P_1 refers to the true precision of the classifier when it is only allowed to predict one qtype for each test instance. P_n refers to the precision when the classifier is allowed to return the n most probable qtypes for each instance and if the correct qtype is in these n qtypes it is counted as a correct prediction.

Table 2 shows the fine-grained results for including all features, as well as the contribution of particular groups of features: *NGRAM* is just the UNIGRAM, BIGRAM and TRIGRAMS, *NO SEMANTIC* is all the features except those that have a semantic content (any that use WordNet, named entities and the gazetteer), and *NO TARGET* is all the features except those that refer to the target.

From these results we can see that, in addition to the ngram features being important for fine classification, the target features also contribute significantly to the end results, while the semantic features have a more marginal impact.

4.1 Hierarchical Classifier

As the labels employed in the current QC scheme actually encode a semantic hierarchy over answer types it makes sense to attempt to use this additional information in our classifiers. Here we propose two hierarchical classification schemes: the first is an integrated approach using feature functions defined over the coarse labels, while the second is a two-stage approach employing an initial coarse classifier to feed a distribution over coarse labels to a second classifier.

The integrated hierarchical classifier builds upon the standard log-linear model described in Section 2 by adding feature functions that are conditioned on only the coarse component of a label.

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

FINE	P_1	P_2	P_3	P_4	P_5
Li & Roth	84.2	-	-	-	95.0
feature-hierarchy	85.6	91.0	94.4	**96.0**	**97.0**
two-stage	86.0	**92.0**	**95.2**	95.8	96.4
flat	**86.6**	91.8	94.4	95.4	95.8

Table 4: Evaluation on fine-grained labels.

The two-stage model first trains a classifier on the training observations using only their coarse labels. This classifier is then used to derive a distribution over coarse labels for the training and test data. Unlike the existing binary features of the model, this distribution is then encoded in real valued feature functions for a second classifier that performs a full labelling.

In order to evaluate our proposed hierarchical classifiers we compare it to a number of other classifiers: *Li & Roth* are the results from [4], *coarse* is the classifier trained only on coarse qtypes, and *flat* is the baseline classifier that treats all the classes independently (no hierarchical information about classes is used).

Tables 3 and 4 show the results of these classifiers for labelling coarse and fine qtypes. The coarse results for the flat, two-stage and feature-hierarchy classifiers are obtained by summing over the probabilities of the child class.

Neither of the hierarchical classifiers can match the flat classifier on the P_1 evaluation, although all three of our classifiers outperform the Li and Roth standard. It is of note however that the hierarchical classifiers do produce a significantly better probability distribution over labels, as evidenced by the P_5 results. In addition, the two-stage classifier outperforms the base coarse classifier. These results suggest that exploiting hierarchical structure could be of benefit for practical QA systems.

5. CONCLUSION

In this paper we have developed a number of log-linear models for question classification. We have systematically explored a wide variety of syntactic and semantic features for this task. We have demonstrated that our novel target word based features can lead to a significant improvement in classifier accuracy. The contribution of this work are new features for question classification which, in combination with a log-linear model, obtain state-of-the-art results. This will immediately result in an improvement in the accuracy and efficiency of question answering systems.

6. REFERENCES

[1] S. Clark and J. Curran. Parsing the WSJ using CCG and log-linear models. In *Proceedings of the 42nd Meeting of the ACL*, pages 103–110, Barcelona, Spain, 2004.

[2] E. Hovy, L. Gerber, U. H. M. Junk, and C. Lin. Question answering in webclopedia. In *Proceedings of the Ninth Text REtrieval Conference (*TREC-*9)*, page 655, 2001.

[3] K. Kocik. Question classification using maximum entropy models. Honours thesis, University of Sydney, 2004.

[4] X. Li and D. Roth. Learning question classifiers. In *In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, 2002.

[5] A. Ratnaparkhi. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, 1996.