Topic Term Identification for Context Question Answering

Matt Gerber
Department of Computer Science
Michigan State University
gerberm2@msu.edu

Joyce Chai
Department of Computer Science
Michigan State University
jchai@cse.msu.edu

Abstract

In context question answering, questions in a session usually center around a specific information goal; in other words, the session as a whole contains an implicit topic that is explored by questions in the session. As shown in previous work, this session topic information is important in processing individual questions in the session. One question is how to automatically identify topic terms from a session of questions. This paper presents our investigation of this problem. Our experiments have shown that given the coherent nature of context questions, shallow processing of questions can achieve reasonable results.

1 Introduction

Recent research in automated question answering has moved towards contextual question answering scenarios (Voorhees, 2001). An example session from the 2005 TREC dataset is the following:

- (Topic) Russian submarine Kursk sinks
- (Q_1) When did the submarine sink?
- (Q_2) Who was the on-board commander of the submarine?
- (Q_3) The submarine was part of which Russian fleet?

As shown in this example, a topic for the session is provided explicitly by TREC. The topic is essential to the interpretation of the questions that follow and can be used to expand the set of query terms for each question. Note that a session topic differs from a question topic; the former is defined with respect to the entire question session and the latter relates to the specific focus of a single question. To differentiate the two, we call the topic of a question the question focus. For example, the question focus in (Q_2) is the name of the commander of the submarine.

Session topic and question focus play important roles in the formation of a coherent query for each question. In fact, a majority of the questions in the TREC collection can be effectively expanded by simply adding terms present in the topic header. However, in practice users might not always define the topic explicitly when issuing a sequence of questions. Therefore, it is important for QA systems to automatically identify the topic terms given a contextual QA session.

For example, in one study conducted by (Sun and Chai, 2006), participants provided sets of questions to satisfy an information need in one of four topic areas: Tom Cruise, 2004 Presidential Debates, Hawaii, and Pompeii. An excerpt is given below:

- (Q_1) Who are the main candidates in the 2004 presidential debate?
- (Q_2) What did the first debate cover?
- (Q_3) Who won?

The topic of the above session is *the presidential debates of 2004*, and the foci of the first question are the names of the *main candidates*. This set of questions differs qualitatively from the TREC data because the topic of the session is implied

by the first question instead of being stated explicitly in a stand-alone field. In addition, the topic implied in the first question is accompanied by other non-topic terms such as *main* and *candidates*. One question is how to differentiate topic terms from other terms and use topic terms to facilitate answer retrieval.

This paper presents results from our current investigation of this question. We developed a probabilistic approach based on logistic regression that computes the probability of a word indicating the topic of a session. Logistic regression provides a distribution of words that can potentially be topic terms. This distribution can be used with backend retrieval components to expand queries and improve retrieved results. We further investigate the complexity of this problem by comparing the probabilistic approach with two baseline approaches. Our results indicate that, given the coherent nature of context questions, shallow processing can achieve good performance in automatically identifying topic terms.

2 Related Work

There has been extensive work done in question classification using both machine learning ((Li and Roth, 2002), (Li et al., 2004), (Zuckerman and Horovitz, 2001), (Hacioglu and Ward, 2003)) and handwritten rule-based systems (Hermjakob, 2001). Machine learning approaches such as those mentioned have been shown to perform very well in practice and mitigate the expense incurred by the construction of hand-written rule sets.

Question focus, as defined above, is related to the question categorization task in that the category assigned to a question characterizes some aspect of its focus. For example, question Q_1 from (Sun and Chai, 2006) might be categorized as *person*, where *person* is a predicate that the answer to Q_1 should satisfy. The primary difference between question categorization and our proposed task is that the former results in expected answer types which are predicates over the potential answer strings while the latter results in the classification of words directly from the questions.

There has also been work done in the area of topic identification for text summarization ((Lin, 1995), (Lin and Hovy, 1997)). Different from summarization where topic identification is based on a discourse of text, our work focuses on topic identification given a coherent discourse of questions.

3 Topic Term Classification

In the regression model, we are interested in computing the value $p(topic|w_i)$, i.e., the probability that word w_i is indicative of the session topic. We model this probability distribution with the Bayesian Logistic Regression toolkit developed by (Genkin et al., 2005). Logistic regression as implemented by (Genkin et al., 2005) provides a simple and efficient solution to our classification problem.

The general form of the logistic regression model is as follows:

$$log\left(\frac{p(x=+)}{p(x=-)}\right) = \vec{w}\cdot\vec{x}+c \quad (1)$$

$$p(x=+) = \frac{e^{\vec{w}\cdot\vec{x}+c}}{1+e^{\vec{w}\cdot\vec{x}+c}} \quad (2)$$

Here, x is a data instance represented as a vector of features, $\vec{w}$ is a vector of weights associated with features from x, and c is a constant. $\vec{w}$ and c are learned from labeled training data. Equation 2 is used to predict the class label for instance x.

3.1 Features

To build the regression model, we use a set of features both from the question session and from the retrieved results. In the TREC QA task, each question in the session is processed sequentially. Here we consider the question set as a whole. The entire discourse can be processed for features of a word w_i. In our view, context QA resembles batch QA since a set of questions is formed by a user without regard for retrieved results. This view is also supported by the observation that the performance of context QA depends very little on the system's ability to track the context of a session from one question

Table 1: Annotated corpus statistics

	(Sun and Chai, 2006)	TREC 2005
% Terms in first question that are topic	53.10	64.12
% Topic terms that are in first question	82.76	96.03
% Percentage of terms in NPs that are topic terms	9.52	7.02
% Percentage of topic terms that are in NPs	42.53	21.81
Topic term distribution over POS		
Noun	79.02%	89.79%
Verb	7.37%	3.39%
Adjective	6.90%	2.55%
Number	6.61%	3.12%
Foreign word	0.0%	1.13%

to the next (Voorhees, 2001). Specifically, for a content word w_i we use the following features from the question session:

Part of Speech Part of speech of w_i. Our intuition is that topic terms are most frequently indicated by nouns.

Frequency of Occurrence Percentage of queries in the session that contain w_i. Terms appearing in multiple questions are more likely to be topic terms.

Presence in First Query Whether w_i appears in the first question of the current session. It is common in the TREC and (Sun and Chai, 2006) data for the session topic to be defined implicitly in the first question.

Noun Phrase (first) Whether w_i is (part of) a noun phrase in the first question. This feature accounts for a common case wherein topic terms modify nouns (e.g., the noun phrase "*2004 presidential* debates" where the italicized words modify the head word).

Noun Phrase (current) Whether w_i is (part of) a noun phrase in the current question

From the retrieved results for the preceding question we obtain the following features:

TF Term frequency of w_i in the top-ranking documents. Terms that are frequent within answering documents to other questions often denote the topics of those documents.

DF Percentage of top-ranking documents that contain w_i. Terms that are common across answer documents often indicate the topic of those documents.

We use Brill's tagger to obtain part of speech information and the LT Chunk tool to do the noun phrase parsing. We have done preliminary experiments with this feature set, and give an analysis of each feature's contribution to the classifier in section 4.

3.2 Data and Annotation

Training the regression model requires labeled data. TREC 2004 and 2005 data is a good source of training data because topics are provided for all sessions. In this study, we are interested in cases where the user does not explicitly define the topic. To make TREC data useful for our purpose, we augmented the first question in the TREC data with terms from the topic header. Our goal in doing so is to remove the topic header and force the system to infer the topic, which is no longer stated explicitly. For example, query Q_1 in the TREC excerpt above would become the following:

- (Q_1') When did the Russian submarine Kursk sink?

Our modfications result in the first question of each session being a stand-alone question with no reliance on context for interpretation. This type of session structure is present throughout

the user study data of (Sun and Chai, 2006). Our annotated corpus contains data from the TREC 2005 question set as well as the (Sun and Chai, 2006) dataset. Table 1 lists statistics from the two data sources. In the table, "NP" stands for "noun phrase" and "POS" stands for "part of speech". As is shown, topic terms are most often indicated by noun phrases in the first question. Our baseline methods and classifier features were designed to reflect this tendency.

We have annotated the above corpora at the word level (ignoring stopwords), as we wish to build a classifier applicable to any candidate expansion term. Each annotated instance (word) is assigned to one of two classes, the positive class if it indicates the topic of the session and the negative class otherwise. Label judgement is based on our assumption that the topic of a query session should be the information that is general across the entire session; therefore, we require positive instances to be semantically relevant to each question in the session, though it is not necessary for positive instances to appear in every question. An example annotated TREC session is shown below, with positive instances shown in boldface:

Q_1: What was the official name of the **Boston Big Dig**?

Q_2: When did the **Big Dig** begin?

Q_3: What was the original estimated cost of the **Big Dig**?

Q_4: What was the expected completion date?

A similar session from the (Sun and Chai, 2006) data is shown below:

Q_1: When were the **2004 presidential debates**?

Q_2: What the topic of the **debates**?

Q_3: What Bush opinion on Tax?

Q_4: Which **debate** has the largest impact?

Table 2 summarizes the amount of annotated data we have produced. Stopwords are not annotated and so are not included in our statistics.

Table 2: Annotation statistics

	Positive	Negative
(Sun and Chai, 2006)	348	1059
TREC 2005	353	629
Total	701	1688

Table 3: Annotations versus reference label set

	Ref pos	Ref neg
Annotated pos	337	16
Annotated neg	1	628

The purpose of annotation is to mark as positive only those terms that indicate the topic of the query session. To assess how accurate our annotation procedure is, we first assume the topic header in the TREC data contains correct labels for positive instances. Next, we compare our set of annotations to this reference set of labels. Table 3 shows the results of this comparison. The annotation was performed while blind to the TREC topic header. The Kappa-measure of agreement between our annotation and the TREC topic header is 0.965, given the probability of agreement by chance is 0.5[1]. The results of this comparison indicate that our annotation process can reliably select topic terms.

4 Evaluation

In all of our evaluations, the logistic regression classifier was trained using ten-fold cross-validation. We also consider two baseline methods. The first method labels any term occuring in the first question a positive instance and all other terms negative instances. This is a reasonable baseline because the first question in all sessions is complete with respect to the topic of the session. The second baseline method labels any term that participates in a noun phrase from the first question positive and all other terms negative. Better performance of the second baseline method would indicate that topic terms are more commonly provided within noun phrases,

[1]The probability of raters r_1 and r_2 agreeing by chance on the value of a binary variable is:
$p(r_1 = 1, r_2 = 1) + p(r_1 = -1, r_2 = -1) = 0.5$

Table 4: Average classifier training accuracy

	Train P	Train R	Train F	Train Acc.
POS	67.56	54.58	60.38	77.55
Session frequency	81.84	60.68	69.69	83.45
Occurs in first	71.02	89.77	79.30	85.32
Term frequency	72.55	39.86	51.45	76.43
Document frequency	100.00	0.00	0.00	69.65
NE current	100.00	0.00	0.00	69.65
NE first	78.60	86.36	82.29	88.36
All features	89.06	84.74	86.86	91.95

and from a cursory examination of the data this appears to be the case.

Table 4 presents the average training accuracy achieved by different logistic regression models trained on all of the annotated data. The first seven rows describe the performance of classifiers trained on single attributes only. The final row of the table indicates the performance of the classifier trained on all features. In the table, P, R, and F stand for *precision*, *recall*, and *f-measure*, respectively. Table 5 presents evaluation results for testing on a held-out portion of the annotated TREC 2005 dataset. Rows following "Logistic Regression" describe logistic regression models trained on the single features listed. The final row of the table indicates the performance of the classifier trained on all features.

Both baseline methods performed quite well on the held-out data. As we expected, the baseline method using noun phrase identification performs better than the baseline that matches words in the first question. This supports the hypothesis that topic terms are commonly provided in noun phrases. Precision is low for both baselines because of their liberal classifying of topic terms. The main improvement of the logistic regression model is in the areas of precision, f-measure, and overall accuracy, where we observed an increase of 13.24%, 8.37%, and 5.96%, respectively, over the best baseline. Table 6 presents results from evaluation with the (Sun and Chai, 2006) dataset. For this dataset, we observed 14.64%, 8.17%, and 4.17% increases

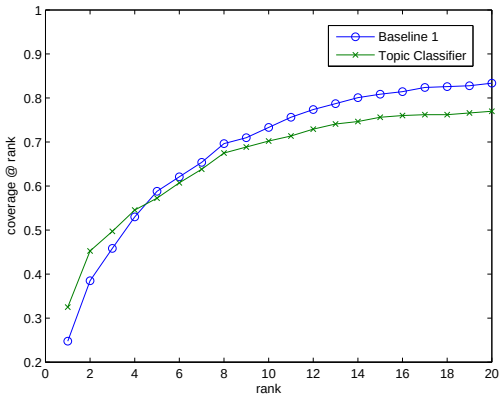

Figure 1: (Sun and Chai, 2006) coverage curve

in precision, f-measure, and accuracy over the best baseline.

5 Classifier Impact on Document Retrieval

We have also conducted preliminary experiments using our logistic regression classifier in a document retrieval system. Our document retrieval system is based on the KL-divergence measure between query and document models. The query and document models are both probability distributions of words, and a particular document's relevance to the query being modeled is the inverse of the KL-divergence between the two models. The query expansion process adds terms to the query model weighted by their classification scores as determined by the topic classifier. The baseline procedure for document retrieval adds all terms from the first question

Table 5: TREC 2005 evaluation results

	P	R	F	Acc.
Baseline 1	70.48	93.67	80.43	85.71
Baseline 2	79.35	92.41	85.38	90.08
Logistic Regression				
POS	70.41	87.34	77.97	84.53
Session frequency	62.69	53.17	57.53	75.40
Occurs in first	70.48	93.67	80.44	85.72
Term frequency	86.36	24.05	37.62	75.00
Document frequency	100.00	0.00	0.00	69.65
NP Current	100.00	0.00	0.00	69.65
NP First	79.35	92.41	85.38	91.08
All features	92.59	94.94	93.75	96.04

Table 6: (Sun and Chai, 2006) evaluation results

	P	R	F	Acc.
Baseline 1	56.48	93.85	70.50	85.83
Baseline 2	67.05	90.77	77.12	90.28
Logistic Regression				
POS	45.67	89.23	60.42	78.89
Session frequency	58.07	27.69	37.50	83.34
Occurs in first	56.48	93.85	70.52	85.84
Term frequency	41.67	38.46	40.00	79.17
Document frequency	100.00	0.00	0.00	81.94
NP Current	100.00	0.00	0.00	81.94
NP First	67.05	90.77	77.12	90.28
All features	81.69	89.23	85.29	94.45

Table 7: Mean reciprocal rank for document retrieval

	(Sun and Chai, 2006)	TREC 2005
Baseline	0.399	0.599
Topic Classifier	0.450	0.640

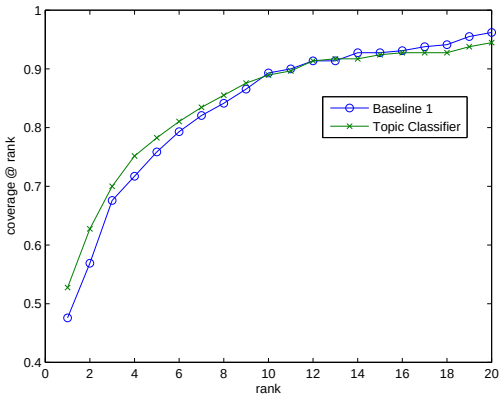

Figure 2: TREC 2005 coverage curve

to the current question, weighted by their frequency of occurrence. This corresponds to the first baseline classifier - all terms in the first question are considered topic (expansion) terms.

Table 7 gives the mean reciprocal rank (MRR) performance for the two methods in document retrieval. MRR rewards systems that improve the ranking of the first answering document and is defined in equation 3. Figures 1 and 2 show the coverage curves for the two methods on the two datasets. Coverage rewards systems that introduce the correct answer above a given rank and is defined in equation 4 (Gaizauskas et al., 2004).

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{rank_q} \quad (3)$$

where Q is the question set and $rank_q$ is the rank of the first retrieved document that answers question q.

$$coverage(n) = \frac{|q \in Q : A_{q,n} \neq \emptyset|}{|Q|} \quad (4)$$

where $A_{q,n}$ is the set of answer documents for question q above rank n. As can be seen in the MRR and coverage graphs, applying the topic classifier to the query model has a positive impact on document rank and answer coverage.

6 Conclusions and Future Work

Topical information is essential to the interpretation of each question in a session and must be identified automatically when not provided by the user. To automatically identify the session topic, we have constructed and evaluated a probabilistic model relying on a small number of informative features. The model computes, for any given word, the probability of that word indicating the topic of the session. We have compared our probabilistic model with two baseline approaches and have shown that our model consistently outperforms each. It should be noted, however, that shallow methods like our baselines perform quite well on the task we looked at. Depending on the application, the expense incurred by training a classifier might not always be justified by the performance gain.

We have also assessed the impact of our classifier on the process of document retrieval and found that document rankings and question coverage scores are both improved by weighting expansion terms in the query model by their classification scores.

Primary future work on this project will investigate how our topic classifier can be integrated into a document retrieval system. The current query model weighting scheme improves retrieval results but is fairly simple - expansion term weights are the normalized product of the term's frequency and classification probability. We believe more sophisticated weighting schemes have the potential to further improve retrieved results.

7 Acknowledgments

This work was supported by an IGERT training grant from the National Science Foundation. We would like to thank Ming Wu for help in developing the KL-divergence document retrieval model.

References

R. Gaizauskas, M. Greenwood, M. Hepple, I. Roberts, and H. Saggion. 2004. The University of Sheffields TREC 2004 QA experiments. In *The Thirteenth Text Retrieval Conference.*

A. Genkin, D. Lewis, and D. Madigan. 2005. Large-scale bayesian logistic regression for text categorization. In preparation.

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

U. Hermjakob. 2001. Parsing and question classification for question answering. In *ACL Workshop on Open-Domain Question Answering*.

X. Li and D. Roth. 2002. Learning question classifiers. In *19th International Conference on Compuatational Linguistics*.

X. Li, K. Small, and D. Roth. 2004. The role of semantic information in learning question classifiers. In *First Joint International Conference on Natural Language Processing*.

C. Lin and E.H. Hovy. 1997. Identifying topics by position. In *5th Conference on Applied Natural Language Processing*.

C. Lin. 1995. Knowledge based automated topic identification. In *33rd Annual Meeting of the Association for Computational Linguistics*.

M. Sun and J. Chai. 2006. Towards intelligent QA interfaces: Linguistically-driven discourse processing for answer retrieval. In *Proceedings of the International Conference on Intelligent User Interfaces*.

E. Voorhees. 2001. Overview of trec 2001 question answering track. In *Proceedings of the Text REtrieval Conference (TREC)*.

I. Zuckerman and E. Horovitz. 2001. Using machine learning techniques to interpret wh-questions. In *Proceedings of the Conference of the Association for Computational Linguistics*.