AUTOMATIC CLASSIFICATION OF QUESTION TURNS IN SPONTANEOUS SPEECH USING LEXICAL AND PROSODIC EVIDENCE

Sankaranarayanan Ananthakrishnan, Prasanta Ghosh, and Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
Department of Electrical Engineering
Viterbi School of Engineering
University of Southern California
Los Angeles, CA 90089
{ananthak,prasantg}@usc.edu, shri@sipi.usc.edu

ABSTRACT

The ability to identify speech acts reliably is desirable in any spoken language system that interacts with humans. Minimally, such a system should be capable of distinguishing between question-bearing turns and other types of utterances. However, this is a non-trivial task, since spontaneous speech tends to have incomplete syntactic, and even ungrammatical, structure and is characterized by disfluencies, repairs and other non-linguistic vocalizations that make simple rule based pattern learning difficult. In this paper, we present a system for identifying question-bearing turns in spontaneous multi-party speech (ICSI Meeting Corpus) using lexical and prosodic evidence. On a balanced test set, our system achieves an accuracy of 71.9% for the binary question vs. non-question classification task. Further, we investigate the robustness of our proposed technique to uncertainty in the lexical feature stream (e.g. caused by speech recognition errors). Our experiments indicate that classification accuracy of the proposed method is robust to errors in the text stream, dropping only about 0.8% for every 10% increase in word error rate (WER).

Index Terms— question turn, speech act, dialog, prosody, spontaneous speech

1. INTRODUCTION

Spontaneous interaction between humans is characterized by various types of speech acts, including but not limited to questions, statements and exclamatory phrases. Knowledge of speech act categories can be useful for automated dialog systems that interact with humans using spoken language. For instance, if the system senses that the user has posed a question, it can use this knowledge in conjunction with information extracted from other modalities (e.g. ASR transcription) to generate a suitable response. Automatic dialog act classification has previously been performed with the aid of prosodic [1, 2], lexical [3, 4] and syntactic [5] cues.

While the vocabulary of dialog acts is usually abstract and domain-specific (the Switchboard-DAMSL corpus [6], for instance, defines 42 types of dialog acts), we focus on a more universal subset of the speech act categorization problem - that of distinguishing question-bearing turns from other types of utterances in spontaneous speech. The fragmented, ungrammatical structure of spontaneous speech makes this a difficult problem. This problem is also characterized by an inherent bias in favor of non-question turns, which significantly outnumber question turns. Previous work in this area has been presented in Jackson et al. [7], where the authors automatically identify question turns in student-tutor interactions in the ITSPOKE [8] database using prosodic and lexical information. Shriberg et al. [2] investigate the use of prosodic cues for identifying dialog acts, including question turns.

In this paper, we present a system that uses prosodic and lexical evidence to detect question turns in multi-party spontaneous speech using two different techniques: 1) maximum-likelihood classification and 2) boosting decision stumps on a bag-of-words representation. Since most dialog systems operate on ASR output, it is desirable that the classifier be robust to ASR errors and degrade gracefully as the error rate increases. To evaluate the behaviour of this aspect of our system, we investigate the variation of classification accuracy as a function of the word error rate (WER) by randomly introducing insertion, substitution and deletion errors to approximate the behavior of an ASR. The rest of this paper is organized as follows: Section 2 gives a detailed description of the database used in our experiments. Section 3 gives an overview of the acoustic-prosodic features and classifiers for question turn classification. Section 4 describes the two lexical classifiers and describes how they are integrated with acoustic evidence to improve performance. Section 5 summarizes our experimental results; Section 6 concludes with an overview of this work and outlines future directions for research.

2. DATA CORPUS

We used the ICSI Meeting Corpus to train and evaluate our question turn classifiers. This corpus consists of spontaneous multi-party conversations from 75 meetings collected at ICSI, UC Berkeley, during the years 2000-2002. The database contains both male and female speakers (native and non-native), though not in equal proportion. We chose this corpus to test the efficacy of our question-turn classification scheme in a real-world scenario with spontaneous utterances, most of which are fragmented, have compound structures, or are otherwise grammatically incorrect. Out of 75 meetings, we obtained our training and test data from 27 randomly selected meetings (ca. 25 hours of speech). Each meeting had between 9 and 15 participants.

We obtained a total of 22,511 turns, of which 2,223 were question bearing turns and the remaining 20,288 were non-questions. Given this bias in favor of non-question bearing turns, we created a balanced test set consisting of 500 samples of each category so as to be able to provide a meaningful evaluation of our system. We used the remaining 21,511 turns (19,788 non-questions and 1,723 questions) for training.

3. ACOUSTIC-PROSODIC CLASSIFIER

We investigate acoustic features associated with pitch, loudness and zero-crossing that discriminate between question and non-question turns. We specifically focus on pitch behaviour at the end of the utterance, as it has been shown to be an indicator of question turns [7].

3.1. Acoustic Features

We estimated fundamental frequency (F0) using an algorithm similar to that presented in [9]. F0 values, short-time energy and zero-crossing rate (ZCR) were computed every 10ms. Each feature was normalized by its mean value in the utterance. Since rising intonation is a characteristic of some types of questions, we extracted a total of 12 prosodic features based on the above parameters from the terminal 200ms of the voiced portion of each utterance. The information gain criterion [10], computed using the Weka toolkit [11], was used to rank the features in order of their importance for classification. Table 1 lists them in order of decreasing information gain. According to this criterion, F0 range within the terminal window is the most informative feature that distinguishes questions from other utterances.

Table 1. Acoustic-prosodic features in order of decreasing information gain for question turn classification

Feature	Description
rng_val	F0 range
min_val	minimum F0
avg_val	average F0
max_val	maximum F0
a_1	F0 slope
zcr_a_2	2nd order term of ZCR polynomial fit
eng_a_1	slope of short-time energy
sd_val	F0 standard deviaton
perc_diff	% difference between terminal avg. F0 to overall avg. F0
eng_a_2	2nd order term of short-time energy polynomial fit
zcr_a_1	slope of ZCR
a_2	2nd order term of F0 polynomial fit

3.2. Acoustic Classifiers

We compared Gaussian mixture model (GMM) and multilayer perceptron (MLP) classifiers for identifying question turns based on the acoustic features. Since our test set was balanced and did not reflect the bias in the training set, we implemented maximum-likelihood (ML) as opposed to traditional maximum *a-posteriori* (MAP) classification. We trained 5-mixture, diagonal covariance GMMs for question and non-question turns using the EM algorithm while discarding the class priors. For classification, the likelihood of acoustic features derived from each test utterance was evaluated using both GMMs, and the utterance was labeled with the class corresponding to the GMM that better fit the observations.

The MLP was trained with 20 hidden nodes and 2 output nodes with softmax activation that provided class posterior probabilities. In order to alleviate the effect of prior bias, we performed post-scaling of the MLP outputs by the appropriate class priors and converted them to pseudo-likelihood scores, which we used for classification.

4. LEXICAL CLASSIFIERS

Although F0-related prosodic features are useful for question turn classification, many types of questions do not exhibit a rising intonation. For instance, interrogatives, also known as *wh*-questions because they often contain the words *what, why, who, which*, etc. are usually characterized by a falling F0 contour. These can easily be confused with declarative statements, which also exhibit a falling intonation pattern. In this case, it is the lexical evidence that distinguishes questions from non-questions. In this section, we describe two types of question-turn classifiers trained on lexical evidence.

4.1. Language model classifier

This is a generative classifier that models short-range context typical of question-bearing turns. It is particularly useful for capturing words and phrases that are commonly found

Table 2. Discriminating words

1-grams	1+2-grams
yeah	what
what	yeah
you	you
mmhmm	do;you
do	do
how	how
is	mmhmm
or	are;we
the	is;it
are	is

in questions. We built two trigram LMs, one for each class, from the training data using the SRILM toolkit [12]. We interpolated these models with a large, spontaneous speech background model (consisting of Switchboard and data mined from the WWW) in order to obtain smooth probability estimates and reduce the effect of out-of-vocabulary (OOV) terms in the test set. For each test utterance, we computed the log probability of the text given the two LMs, and assigned the class label whose model better fit the input text.

4.2. Bag-of-words classifier

In a bag-of-words (BOW) representation, each utterance is described by a feature vector that contains counts of each vocabulary item that occurs in it. Such a representation is quite popular for document matching and classification. We used the CMU BOW [13] toolkit to obtain a BOW representation for the two classes. The information gain criterion was used to determine which words were important for discriminating between question and non-question turns. Table 2 shows the 10 most important unigrams and bigrams ranked in order of decreasing information gain. For classification, we used discrete AdaBoost over *decision stumps*, which are simple rules (single node decision trees) that classify a test utterance based on a threshold of the count of a word in the BOW representation. The BoosTexter tool [14] was used to implement this classifier.

4.3. Combined acoustic and lexical classifier

For ML classification, we combined likelihood scores provided by the MLP with log probability scores provided by the language model. For the BOW representation, classification was performed using boosted decision stumps (AdaBoost) that combined the acoustic and lexical features (counts of unigrams and bigrams).

Table 3. Question classification performance

Method	Accuracy
Chance	50.0%
Acoustic (GMM)	55.4%
Acoustic (MLP)	61.0%
Lexical (LM)	69.9%
MLP + LM	71.2%
Lexical (BOW)	71.3%
BOW + Acoustic	**71.9%**

5. EXPERIMENTAL RESULTS

We divided the entire corpus into 10 random training and testing partitions, creating balanced test sets containing a total of 1,000 turns, with 500 samples each of question and non-question turns. Since the test set was balanced, the chance level was 50%. The remaining data was used for training. First, we used clean transcriptions from the corpus to train the lexical models. Table 3 summarizes the performance of various individual and compound classification techniques. The use of prosodic features with the MLP improved performance over chance by 11% absolute, whereas the language model performed much better with a gain of almost 20% absolute. Integrating the acoustic features with the language model classifier provided an additional boost of 1.3% absolute. The BOW classifier bettered the chance level by over 21% absolute. When combined with the acoustic evidence, performance of the BOW-based classifier improved by nearly 22% over chance.

We then studied the effect of errors in the text transcription on classification performance. Instead of performing full-blown speech recognition, we used a script to corrupt the test text to the desired word error rate (WER). This allowed us to easily manipulate the error rate and also its individual constituents - insertions, substitutions and deletions - and perform experiments in this controlled environment. The downside of this method is that the errors introduced are random and not characteristic of ASR (i.e. not based on acoustic or language model confusability). Figure 1 illustrates the variation in question turn classification performance as a function of the WER for different configurations. The LM classifier exhibits a 9.1% degradation in performance as the WER increases from 0% to 50%. On the other hand, the BOW classifier is more robust to word errors, showing just a 4.9% degradation over this range. When combined with acoustic features, this classifier exhibits just a 3.8% reduction in classification accuracy over the same range of WER.

6. DISCUSSION AND FUTURE WORK

Question turn classification is one aspect of the speech act identification problem that we have addressed in this paper. For ML classification, we used GMM and MLP classifiers

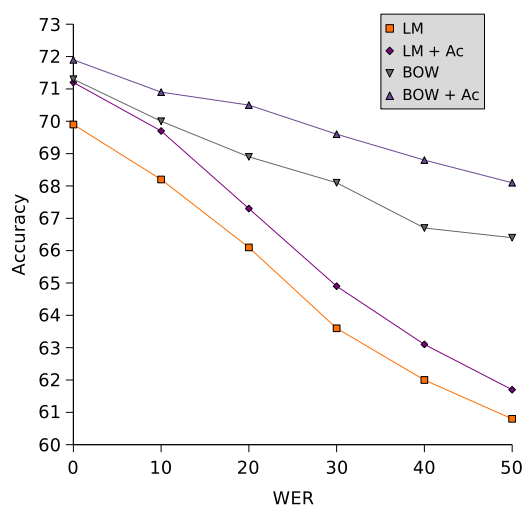

Fig. 1. Variation of question classification accuracy vs. WER

for acoustic features and a n-gram language model for lexical features. For boosted decision stump classification, we used a BOW representation for the lexical features. We note that, while the prosodic features are useful and perform better than chance, it is the lexical features that provide more predictive power for distinguishing questions from non-questions. Combining the two results in a small improvement over the lexical-only classifier.

We also examined the effect of errors in the text transcription on classification performance. The LM classifier suffers significant performance degradation as the WER increases. This is because introduction of errors not only affects the keywords that discriminate between the two classes, but also the surrounding context, causing a progressively larger mismatch between the noisy text and its corresponding LM. On the other hand, the BOW classifier is more robust to these errors, with a much smaller reduction in performance over the same range of WER. This is due to the fact that the BOW classifier does not utilize local context. This is a useful result for systems that work with the output of ASR. Another interesting observation about the BOW-based system is that, while the acoustic features do not provide much gain in combination with clean text (0.6%), the performance gap widens as the WER increases - at 50% WER, the difference is 1.7%.

One limitation of this work is that the question turn detector works in a context-free fashion. However, context is likely to provide important cues for identifying question turns. Along with other information such as trigger words and speaker change, contextual constraints can be combined with the context-free classifier to track the flow of conversation and improve question-turn identification. Interesting issues in this regard include selection of contextual features and the development of a suitable framework for modeling the progression of these events.

7. REFERENCES

[1] M. Mast, R. Kompe, S. Harbeck, A. Kiessling, and V. Warnke, "Dialog act classification with the help of prosody," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, Philadelphia, PA, 1996, pp. 1732–1735.

[2] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. V. Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, pp. 439–487, 1998.

[3] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic and syntactic cues for dialog acts," in *Proceedings of the ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada, August 1998, pp. 114–120.

[4] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialog act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, September 2000.

[5] S. Bangalore, G. D. Fabbrizio, and A. Stent, "Learning the structure of task-driven human-human dialogs," in *Proceedings of ACL*, Sydney, July 2006, pp. 201–208.

[6] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. V. Ess-Dykema, "Switchboard discourse language modeling project final report," Johns Hopkins University, Baltimore, MD, Tech. Rep. Research Note No. 30, 1998.

[7] J. Liscombe, J. J. Venditti, and J. Hirschberg, "Detecting question-bearing turns in spoken tutorial dialogues," in *Proceedings of the International Conference of Spoken Language Processing*. Pittsburgh, PA, 2006.

[8] D. Litman and S. Silliman, "ITSPOKE: An intelligent tutoring spoken dialogue system," in *Proceedings of the 4th Meeting of HLT/NAACL (Companion Proceedings)*. Boston, MA, May 2004.

[9] B. Secrest and G. Doddington, "An integrated pitch tracking algorithm for speech systems," in *Proceedings of the International Conference on Acoustics and Speech and Signal Processing*, 1983, pp. 1352–1355.

[10] J. R. Quinlan, "Induction of decision trees," in *Machine Learning*. Kluwer Academic Press, 1986, vol. 1, pp. 81–106.

[11] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, 2005.

[12] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference of Spoken Language Processing*, 2002.

[13] A. K. McCallum, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering," 1996, http://www.cs.cmu.edu/ mccallum/bow.

[14] R. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39(2/3), pp. 135–168, 2000.