Any Questions? Automatic Question Detection in Meetings

Kofi Boakye, Benoit Favre, Dilek Hakkini-Tür

International Computer Science Institute
1947 Center St., Suite 600, Berkeley, CA, USA
{kaboakye,favre,dilek}@icsi.berkeley.edu

***Abstract*—In this paper, we describe our efforts toward the automatic detection of English questions in meetings. We analyze the utility of various features for this task, originating from three distinct classes: lexico-syntactic, turn-related, and pitch-related. Of particular interest is the use of parse tree information in classification, an approach as yet unexplored. Results from experiments on the ICSI MRDA Corpus demonstrate that lexico-syntactic features are most useful for this task, with turn- and pitch-related features providing complementary information in combination. In addition, experiments using reference parse trees on the Broadcast Conversation portion of the OntoNotes release 2.9 data set illustrate the potential of parse trees to outperform word lexical features.**

I. Introduction

Identifying questions in human dialogs is an important first step to automatically processing and understanding natural speech. In the case of human-computer dialog systems, it can be critical for the conversational agent to know that a user asked it a question so that the dialog can be directed accordingly. In more passive systems, such as those utilized in multiparty meetings, question detection is useful for meeting indexing and summarization. Information about the presence of questions can be used to make for more coherent summaries, detect action items, and generally improve offline meeting browsing. This is especially the case if question/answer pairs are identified, as in [1].

In this work we sought to analyze the utility of various features for automatic detection of English questions in the meetings domain. For this task we used pre-segmented utterances and attempted to classify the utterances as questions or statements. The choice to pre-segment was made so as not to conflate the challenges of identifying these dialog act units with those of classifying them.

Though related studies exist, they have generally focused on other domains [2], other languages [3], [4], or have been performed in the broader context of dialog act classification [5], [6], [7], [8]. Based on these studies, common features related to words, part-of-speech, speaker turns, and pitch, were examined. In addition, initial results on a related data set (discussed in Section VI) indicated that lexical and syntactic information from parse trees, as yet unexplored, would be of particular use, so this feature was used as well.

The paper is organized as follows. Section II identifies and discusses the primary question types and Section III the corpus with which our experiments were performed. We describe the features analyzed in Section IV and various individual feature and combination experiments in Section V. Results are discussed in Section VI while the conclusion and future work is presented in Section VII.

II. Question Types

The detection of English questions is complicated by the fact that this dialog act consists of subclasses, each with different attributes and canonical forms. Though other variations exist, there are three main types of questions. An example of each question type can be found in Table 1. **Yes-No**-questions attempt to elicit a limited range of responses ("Yes", "No", "I don't know", etc.) from the recipient and are characterized by the presence of an auxiliary verb ("do", "have", "be", etc.) and subject-verb order inversion with this verb. **Wh**-questions contain a *wh*-word (i.e. an interrogative such as "who", "what", "where", etc.) substituted for the subject or object and moved to the beginning of the sentence, a process sometimes referred to as "*wh*-fronting". The final type, **declarative** questions, do not differ in syntax from statements, but are understood to be questions by the recipient using other cues. One cue often discussed in the literature is intonation, in particular a rising pitch at the end of the utterance [2], [3]. Some research has shown [9], however, that declarative questions are often intonationally equivalent to proper declaratives. In such cases it is believed that speakers use lexical and contextual information to identify the utterance as a question. Liscombe et al. in [2], for example, note that the presence of second-person pronouns (e.g, "you" and "your") are more likely to indicate a question than first-person pronouns (e.g., "I" and "my") because a speaker presumably knows his or her cognitive state, but not necessarily that of the person to whom he or she is speaking.

TABLE I
Question type examples.

Question Type	Example
Yes-No	*Have you looked at that?*
Wh	*What was the nature of the email?*
Declarative	*You're editing your slide?*

III. Corpus

For our experiments, we used the ICSI Meeting Recorder Dialog Act (MRDA) Corpus [10], a derivative of the ICSI

Meeting Corpus. This corpus consists of 75 meetings averaging about an hour in length with audio data obtained from both nearfield (lapel and individual headset) and farfield (tabletop) microphones. For this work, a subset of 73 meetings was used, divided into sets of 51 training, 11 development, and 11 test meetings, as in previous work involving this data [10], [11].

Automatic speech recognition output (both word hypothesis and word timing information) was obtained using the SRI CTS Recognizer [12] trained on conversational telephone speech. Recognition was performed on the nearfield audio data with a word error rate (WER) of 38.2. Human reference transcripts were also utilized, with word timing information produced using forced alignment.

Utterances were pre-segmented using reference dialog act information contained in the corpus. A question was defined to be a sequence of words in which the final word was followed by a question dialog act label/sentence boundary class. A statement was similarly defined using the statement dialog act label/sentence boundary class. Segment start and end times were determined using the word timing information obtained as described above. The 73 meetings collectively contain approximately 63,000 utterances of which approximately 10% are questions. Summary statistics of the data can be found in Table III.

TABLE II
Summary statistics for the 73-meeting subset of the ICSI MRDA Corpus.

Number of utterances	63,514
Number of questions	6,318
Number of statements	57,196
Average utterance length (words)	8
Average meeting duration (minutes)	56
Number of unique unigrams	11,890
Number of unique bigrams	138,244

IV. Features for question detection

Several features were examined for the task. These can roughly be divided into three groups: Features related to words and syntax; features related to the turn-taking nature of conversational speech; and acoustic features related to pitch/intonation. Descriptions of these features follow.

A. Lexico-syntactic features

Word n-grams

Lexical cues serve as the primary source for identifying an utterance as a question or statement. As previously mentioned, the presence of an auxiliary verb, an utterance-initial wh-word, the second-person pronoun "you", and word order inversion are all indicators of an interrogative utterance. Each of these can, to some extent, be encoded using word unigrams and bigrams ("do you", "what is", "can we", etc.). These n-grams were included as features for the classifier. To provide some word-position information, each word sequence was enclosed by BEGIN and END tags.

Part-of-speech (POS) tag n-grams

A simple representation of syntax can be obtained using part-of-speech (POS) n-grams, and as such this is a common feature for question detection [2], [13]. POS tags were obtained using the tagger PoST [14], [15]. This is an HMM based tagger which performs discriminative reranking of N-best hypotheses using features derived from n-grams. Training data consisted of a combination of data from the Penn, Fisher, and Switchboard treebanks. The Penn data was "speechified" by normalizing case and converting digits to their written form to provide a better match to the other two training data sources (both consisting of conversational speech) as well as the Meeting Corpus data used in the experiments. Features consisted of POS tag unigrams, bigrams, and trigrams.

Parse trees

A much richer representation of syntax is available through the use of parse trees. In addition to being richer, however, this representation is more complex, and the challenge exists in properly exploiting the information within the parse trees for classification. Our approach consisted of using parse subtrees identified with a tree-based classifier, BACT [16], as features in our classifier. BACT (Boosting Algorithm for the Classification of Trees), identifies the subtrees of structured or semi-structured text and uses them in a boosting algorithm that employs subtree-based decision stumps as weak learners. At testing time, all subtrees triggered for a given example are output and these are the features used in our classifier. This approach facilitates the combination of the subtree information learned by the BACT classifier with other features of interest. Utterance parse trees were generated using the Berkeley Parser[1] trained on Wall Street Journal data. To better match the experimental data, case and punctuation information was removed.

B. Turn-related features

Utterance length

Utterance length has also been shown to help discriminate between various dialog acts, including questions and statements. Shriberg et al. in [5], for example, observed statements to be longer than questions. This holds true to an extent for the MRDA corpus, in which the average question is of length 6.8 while the average statement is of length 9.

Speaker change information

Beyond those related to an utterance in isolation, there exist additional cues for questions which derive from the turn-taking behavior of human conversation. For example, if the previous and following utterances of a given utterance are both produced by the speaker of the current utterance, it is unlikely that this utterance is a question. To encode this information we use two feature values, one which indicates speaker identity matching for the previous utterance and another for the following one.

[1] http://code.google.com/p/berkeleyparser

C. Pitch-related features

F0 statistics

Given that pitch information at the end of an utterance may serve as a question cue, the maximum, minimum, mean, range, and standard deviation F0 statistics in the last 200 ms of the utterance were used as a feature.

Final F0 slope

To explicitly capture the rising pitch trajectory believed to be present at the end of certain question types, the final F0 slope was also included as a feature. This slope was computed simply using the beginning and end points of the sequence of pitch values for the last 500 ms of the utterance, or the entire utterance, if shorter than 500 ms.

V. Experiments

Classification experiments involved identifying each pre-segmented utterance as either a question or statement. This was done both for ASR and reference word and word timing data. These experiments were carried out using icsiboost[2], an implementation of the well-known AdaBoost algorithm [17]. AdaBoost produces a classification rule by combining a set of weak classifiers (or "learners"). The algorithm finds a set of weak hypotheses by calling these weak learners in a series of iterations (or "rounds") and finally combining the weak hypotheses into a single rule. At each round, a distribution of weights over training examples is updated such that incorrectly classified examples receive more weight and the new classifier focuses more on these examples.

For this work, the weak learners used were one-level decision trees (stumps). One thousand rounds of training were performed, after which a decision threshold, optimized for maximum F-measure, was determined using development data. To evaluate the classification performance, test data was scored using precision, recall, and F-measure, with F-measure being the primary figure of merit.

TABLE III
MRDA individual and group performance results for ASR output.

Feature	Recall	Precision	F-measure
Words	47.11	57.13	51.64
POS Tags	35.96	34.58	35.25
Parse Trees	48.71	51.04	49.80
Combo	47.41	53.97	50.48
Utt. Length	60.16	14.62	23.52
Speaker change info	59.96	16.10	25.39
Combo	66.14	17.65	27.86
F0 Stats	42.73	18.56	25.87
Final F0 slope	48.11	21.56	29.78
Combo	37.95	19.71	25.94

Table III shows the results for the individual features as well as combinations of related feature subsets when using ASR output. With regard to the lexico-syntactic features, word n-grams give the best performance, followed by parse trees and POS tags. The parse trees far outperform the POS tags, differing in F-measure by 14.5% absolute, while the difference between parse trees and word n-grams is 1.84% absolute. The lexical and syntactic information contained in the parse trees appears to be an improvement over the POS tags, but falls short of the word n-grams. In addition, the combination of these features yields a slight reduction in F-measure over the single best feature. A possible reason for this is that the features possess redundant information. The turn-related features—utterance length and speaker change information—perform rather poorly, particularly in terms of precision. These two features, however, combine to produce gains over the individual features, suggesting they possess complementary information. In addition, speaker change information yields better results in isolation than utterance length. The pitch-related features, too, have low precision and, consequently, low F-measure. Furthermore, final F0 slope appears to be a better feature than general F0 statistics. Combination, here, too, fails to outperform this best individual feature.

The same experiments were carried out using reference transcripts and these results are presented in Table IV. First of note is the substantial increase in precision, recall, and F-measure for the lexico-syntactic features. The word n-gram F-measure, for example, differs by nearly 16% absolute from the automatic output case; the noise introduced by ASR is considerable. Again, we observe the trend that words and parse trees yield similar performance, whereas POS tags significantly underperform these other two features. In addition, as the POS tag relative improvement between ASR and reference is highest, these features appear to be least robust to errorful transcriptions. Unsurprisingly, the turn- and pitch-related features remain largely unchanged in terms of their performance. The small difference can be attributed to start and end time differences between ASR and reference words.

TABLE IV
MRDA individual and group performance results for reference transcripts.

Feature	Recall	Precision	F-measure
Words	66.32	68.68	67.48
POS Tags	40.46	65.78	50.11
Parse Trees	69.16	64.55	66.78
Combo	67.32	67.83	67.57
Utt. Length	75.69	13.45	22.84
Speaker change info	59.96	16.10	25.39
Combo	73.37	16.24	26.59
F0 Stats	34.88	20.02	25.44
Final F0 slope	45.62	22.11	29.78
Combo	40.38	18.95	25.80

Having observed the interaction of features within each subset, additional feature combination analysis was performed using features between subsets. As combinations within the lexico-syntactic and pitch-related feature subsets produced no gains, the best performing feature from each of these groups was selected as a candidate for combination, while the two turn-related features as a unit served as the third candidate. Results for ASR output are presented in Table V and those for reference transcripts appear in Table VI.

Similar trends can be observed for these two evaluation conditions. In all cases, combination yields improved F-measure

[2]http://code.google.com/p/icsiboost

TABLE V
MRDA feature combination results for ASR output.

Feature	Recall	Precision	F-measure
Lex + Turn	48.41	58.41	52.94
Lex + Pitch	45.22	62.79	52.58
Turn + Pitch	42.43	28.12	33.82
Lex + Turn + Pitch	52.89	55.25	54.05

results over the individual candidate features. Changes in recall and precision results vary, and this is due to optimizing the threshold for maximum F-measure, as previously mentioned. In addition, the single best combination involves all candidate features. For ASR output, this combination represents an improvement of 2.41% absolute in F-measure over the best single-feature performance while in the reference case the improvement is 2.23% absolute. With regard to two-feature combination, it appears that turn- and pitch-related features complement lexical features similarly but can produce still more gains when both are combined with the lexical features.

TABLE VI
MRDA feature combination results for reference transcripts.

Feature	Recall	Precision	F-measure
Lex + Turn	62.80	76.23	68.86
Lex + Pitch	63.83	75.74	69.28
Turn + Pitch	48.45	26.84	34.55
Lex + Turn + Pitch	66.84	72.85	69.71

VI. Discussion

The results presented in Section V clearly demonstrate the dominance of lexico-syntactic features in the detection of questions in English. But what, exactly, is the classifier learning? A simple analysis can be done by looking at the weights of the decision stumps contained in the classifier model. In our case, the more positive the weight, the more the decision is moved in the direction of classifying the test example as a question. Multiplying these weights by the training example frequencies can produce a rough ranking of the relative importance of each decision stump in identifying questions.

TABLE VII
Frequency-weighted ranking of top six question-discriminating decision stumps for the lexico-syntactic features. Results obtained using ASR output.

Words	POS Tags	Parse Trees
what	RB	you
how	WRB	NP
mean	RB_END	(S(NP)(VP))
BEGIN_what	BEGIN_WP	WP
do_you	BEGIN_RB_PRP	(ADVP(RB(right)))
BEGIN_you	VBZ_PRP	WRB

Table VII presents such a ranking for the top six decision stumps for each of the lexico-syntactic features. In the case of words, we observe *wh*-words, *wh*-fronting, the auxiliary verb "do" with subject-verb inversion, as well as the second-person pronoun "you". Indeed, most of the lexical and syntactic cues discussed in Section II appear to have been identified. The POS tag decision stumps show a less convincing learning of question cues; a *wh*-adverb (WRB) and a *wh*-pronoun (WP) are among the top six. The parse subtrees lie somewhere between these two, with the second-person pronoun "you", a *wh*-pronoun (WP) and a *wh*-adverb (WRB) in the top six. This interestingly parallels the performance of the three features on the evaluation data. Recall, though, that this ranking is approximate, and the presence of non-cue decision stumps may be in part due to disproportionately large frequencies, as is most certainly the case for RB with POS tags and (S(NP)(VP)) with parse trees.

The limitation of representing syntax using POS tags, in contrast with parse trees, seems to have been demonstrated by our results as well. That being said, word n-grams, which lack most of the structural information of parse trees, outperform them. Why is this the case? A likely reason is that the parses, being automatically generated, are errorful, much like the ASR output. And just as the ASR errors can degrade classification performance, so, too, can the parsing ones. Table VIII shows classification performance results using parse

TABLE VIII
OntoNotes word and parse tree performance results.

Feature	Recall	Precision	F-measure
Words (ref)	62.55	62.55	62.55
Parse Trees (auto)	63.35	76.44	69.28
Parse Trees (ref)	82.07	81.42	81.75

trees on the Broadcast Conversation subset of release 2.9 of the OntoNotes data set (LDC2009E05) [18]. In addition to reference transcriptions (derived from closed captions), the data contains manually annotated parse trees of some 13,000 utterances. Approximately 9% of the utterances examined were questions. Here we see that a significant performance gap (12.47% absolute) exists between automatic and reference parse trees. Furthermore, in this case, both features outperform word n-grams produced using *reference* data. Clearly, much stands to be gained from applying our efforts to improving parsing accuracy for the meetings genre.

On the other hand, the limited ability of end-of-utterance pitch information to identify questions appears to be confirmed as well. Pitch information, however, may be valuable in specific cases where lexico-syntactic features are insufficient. Single-word utterances—such as "right", "yeah", and "okay"—possess little lexical and syntactic information to distinguish their question and statement forms. Because of this ambiguous nature, too, these utterances may exhibit pronounced intonation that can be used to correctly classify them. Table IX presents classification results on single-word utterances using ASR words both with and without pitch. The roughly 3% absolute improvement in F-measure when using the F0 slope pitch-related feature demonstrates its usefulness in this particular context. Thus, one could envision an approach in which single-word utterances are classified with this pitch information as a feature and multi-word utterances are not.

Lastly, it should be noted that a comparison of the performance of the approaches examined in this work to others is

TABLE IX
MRDA performance results for single-word utterances using ASR output.

Feature	Recall	Precision	F-measure
Lex	71.03	61.29	65.80
Lex + Pitch	65.89	71.94	68.78

difficult. As previously mentioned, only a few related studies have utilized the same data set, and these have focused mainly on overall dialog act tagging. Consequently, performance specific to question detection was not presented. While not directly comparable, the performance in these studies is only slightly higher, which is expected as some DA units, such as backchannels, are less challenging to detect; Stolcke et al. in [6], for example, included a set of regular expressions for backchannel detection.

VII. Conclusions and Future Work

In this paper we presented an automatic approach to the detection of questions among English utterances from multiparty conversational speech. An analysis of features used for classification revealed that lexico-syntactic features are most useful for this task, with turn- and pitch-related features providing complementary information in combination. The lexico-syntactic features seem to enable the classifier to correctly identify the cues that signal a question. This is particularly the case for word n-grams, which slightly outperform parse trees and significantly outperform POS tags. Furthermore, the added syntactic information in parse trees as compared to POS tags was shown to greatly improve performance. Additional experiments using reference parse trees on an alternate data set demonstrated the potential of these parse trees to outperform word n-grams as well.

With this in mind, one of the likely next steps for this work is to improve the quality of the parser output. Though efforts were made to reduce the mismatch between the evaluation and parser training data, it is clear that there is still work to be done in this regard. Another possible extension is to combine this approach with sentence segmentation. In this work utterances were pre-segmented, but it is possible to independently segment and classify. It would be of interest to see how this two-stage approach compares to a joint segmentation/classification one. Lastly, the analysis performed here should be extended to other languages. The existence of lexico-syntactic cues to questions is not unique to English—other languages, for example, possess a set of *wh*-words—suggesting the approach should generalize. The extent to which it does, however, is an open area of investigation.

Acknowledgements

We would like to thank Liz Shriberg and James Fung for their helpful commments and suggestions. This work is supported by the Defense Advanced Research Projects Agency (DARPA) GALE project, under Contract No. HR0011-06-C-0023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] A. Kathol and G. Tur, "Extracting question/answer pairs in multi-party meetings," in *Proc. ICASSP 2008*, 2008, pp. 5053–5056, Las Vegas, NV.

[2] J. Liscombe, J. Venditti, and J. Hirschberg, "Detecting question-bearing turns in spoken tutorial dialogues," in *Proc. Interspeech 2006*, 2006, pp. 69–72, Pittsburgh, PA.

[3] J. Yuan and D. Jurafsky, "Detection of questions in Chinese conversational speech," in *Proc. ASRU 2005*, 2005, pp. 47–52, San Juan, Puerto Rico.

[4] V. M. Quang, L. Besacier, and E. Castelli, "Automatic question detection: prosodic-lexical features and cross-lingual experiments," in *Proc. Interspeech 2007*, 2007, pp. 2257–2260, Antwerp, Belgium.

[5] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, 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.

[6] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, pp. 339–371, 2000.

[7] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP 2005*, 2005, pp. 1061–1064, Philadelphia, PA.

[8] M. Zimmermann, D. Hakkani-Tür, E. Shriberg, and A. Stolcke, "Text based dialog act classification for multiparty meetings," in *MLMI 2006, Lecture Notes in Computer Science*, S. Renals, S. Bengio, and J. Fiscus, Eds. Springer Berlin/Heidelberg, 2006, vol. 4299, pp. 190–199.

[9] M. Šafářová and M. Swerts, "On recognition of declarative questions in English," in *Proc. of Speech and Prosody*, 2004, pp. 313–316, Nara, Japan.

[10] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SIGDIAL*, 2004.

[11] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "A* based joint segmentation and classification of dialog acts in multi-party meetings," in *Proc. ASRU 2005*, 2005, pp. 215–219, San Juan, Puerto Rico.

[12] Q. Zhu, A. Stolcke, B. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proc. Interspeech 2005*, 2005, pp. 2141–2144, Lisbon, Portugal.

[13] A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schulz, H. Soltau, H. Yu, and K. Zechner, "Advances in automatic meeting record creation and access," in *Proc. ICASSP 2001*, 2001, Seattle, WA.

[14] S. M. Thede and M. P. Harper, "A second-order hidden markov model for part-of-speech tagging," in *Proc. of the 28th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 175–182, Baltimore, MD.

[15] Z. Huang, M. Harper, and W. Wang, "Mandarin part-of-speech tagging and discriminative reranking," in *Proc. EMNLP CoNLL 2007*, 2007, pp. 1093–1102, Prague, Czech Republic.

[16] T. Kudo and Y. Matsumoto, "A boosting algorithm for the classification of semi-structured text," in *Proc. EMNLP 2004*, 2004, pp. 301–308, Barcelona, Spain.

[17] R. E. Schapire and Y. Singer, "Boostexter: a boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, May 2000.

[18] S. Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: A unified relational semantic representation," in *Proc. ICSC 2007*, 2007, pp. 517–526.