Question Detection in Spoken Conversations Using Textual Conversations

Anna Margolis and **Mari Ostendorf**
Department of Electrical Engineering
University of Washington
Seattle, WA, USA
{amargoli,mo}@ee.washington.edu

Abstract

We investigate the use of textual Internet conversations for detecting questions in spoken conversations. We compare the text-trained model with models trained on manually-labeled, domain-matched spoken utterances with and without prosodic features. Overall, the text-trained model achieves over 90% of the performance (measured in Area Under the Curve) of the domain-matched model including prosodic features, but does especially poorly on declarative questions. We describe efforts to utilize unlabeled spoken utterances and prosodic features via domain adaptation.

1 Introduction

Automatic speech recognition systems, which transcribe words, are often augmented by subsequent processing for inserting punctuation or labeling speech acts. Both prosodic features (extracted from the acoustic signal) and lexical features (extracted from the word sequence) have been shown to be useful for these tasks (Shriberg et al., 1998; Kim and Woodland, 2003; Ang et al., 2005). However, access to labeled speech training data is generally required in order to use prosodic features. On the other hand, the Internet contains large quantities of textual data that is already labeled with punctuation, and which can be used to train a system using lexical features. In this work, we focus on question detection in the Meeting Recorder Dialog Act corpus (MRDA) (Shriberg et al., 2004), using text sentences with question marks in Wikipedia "talk" pages. We compare the performance of a question detector trained on the text domain using lexical features with one trained on MRDA using lexical features and/or prosodic features. In addition, we experiment with two unsupervised domain adaptation methods to incorporate unlabeled MRDA utterances into the text-based question detector. The goal is to use the unlabeled domain-matched data to bridge stylistic differences as well as to incorporate the prosodic features, which are unavailable in the labeled text data.

2 Related Work

Question detection can be viewed as a subtask of speech act or dialogue act tagging, which aims to label functions of utterances in conversations, with categories as question/statement/backchannel, or more specific categories such as request or command (e.g., Core and Allen (1997)). Previous work has investigated the utility of various feature types; Boakye et al. (2009), Shriberg et al. (1998) and Stolcke et al. (2000) showed that prosodic features were useful for question detection in English conversational speech, but (at least in the absence of recognition errors) most of the performance was achieved with words alone. There has been some previous investigation of domain adaptation for dialogue act classification, including adaptation between: different speech corpora (MRDA and Switchboard) (Guz et al., 2010), speech corpora in different languages (Margolis et al., 2010), and from a speech domain (MRDA/Switchboard) to text domains (emails and forums) (Jeong et al., 2009). These works did not use prosodic features, although Venkataraman

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 118–124,
Portland, Oregon, June 19-24, 2011.

et al. (2003) included prosodic features in a semi-supervised learning approach for dialogue act labeling within a single spoken domain. Also relevant is the work of Moniz et al. (2011), who compared question types in different Portuguese corpora, including text and speech. For question detection on speech, they compared performance of a lexical model trained with newspaper text to models trained with speech including acoustic and prosodic features, where the speech-trained model also utilized the text-based model predictions as a feature. They reported that the lexical model mainly identified *wh* questions, while the speech data helped identify *yes-no* and *tag* questions, although results for specific categories were not included.

Question detection is related to the task of automatic punctuation annotation, for which the contributions of lexical and prosodic features have been explored in other works, e.g. Christensen et al. (2001) and Huang and Zweig (2002). Kim and Woodland (2003) and Liu et al. (2006) used auxiliary text corpora to train lexical models for punctuation annotation or sentence segmentation, which were used along with speech-trained prosodic models; the text corpora consisted of broadcast news or telephone conversation transcripts. More recently, Gravano et al. (2009) used lexical models built from web news articles on broadcast news speech, and compared their performance on written news; Shen et al. (2009) trained models on an online encyclopedia, for punctuation annotation of news podcasts. Web text was also used in a domain adaptation strategy for prosodic phrase prediction in news text (Chen et al., 2010).

In our work, we focus on spontaneous conversational speech, and utilize a web text source that is somewhat matched in style: both domains consist of goal-directed multi-party conversations. We focus specifically on question detection in pre-segmented utterances. This differs from punctuation annotation or segmentation, which is usually seen as a sequence tagging or classification task at word boundaries, and uses mostly local features. Our focus also allows us to clearly analyze the performance on different question types, in isolation from segmentation issues. We compare performance of textual- and speech-trained lexical models, and examine the detection accuracy of each question type. Finally, we compare two domain adaptation approaches to utilize unlabeled speech data: bootstrapping, and Blitzer et al.'s Structural Correspondence Learning (SCL) (Blitzer et al., 2006). SCL is a feature-learning method that uses unlabeled data from both domains. Although it has been applied to several NLP tasks, to our knowledge we are the first to apply SCL to both lexical and prosodic features in order to adapt from text to speech.

3 Experiments

3.1 Data

The Wiki talk pages consist of threaded posts by different authors about a particular Wikipedia entry. While these lack certain properties of spontaneous speech (such as backchannels, disfluencies, and interruptions), they are more conversational than news articles, containing utterances such as: "Are you serious?" or "Hey, that's a really good point." We first cleaned the posts (to remove URLs, images, signatures, Wiki markup, and duplicate posts) and then performed automatic segmentation of the posts into sentences using MXTERMINATOR (Reynar and Ratnaparkhi, 1997). We labeled each sentence ending in a question mark (followed optionally by other punctuation) as a question; we also included parentheticals ending in question marks. All other sentences were labeled as non-questions. We then removed all punctuation and capitalization from the resulting sentences and performed some additional text normalization to match the MRDA transcripts, such as number and date expansion.

For the MRDA corpus, we use the manually-transcribed sentences with utterance time alignments. The corpus has been hand-annotated with detailed dialogue act tags, using a hierarchical labeling scheme in which each utterance receives one "general" label plus a variable number of "specific" labels (Dhillon et al., 2004). In this work we are only looking at the problem of discriminating questions from non-questions; we consider as questions all complete utterances labeled with one of the general labels *wh*, *yes-no*, *open-ended*, *or*, *or-after-yes-no*, or *rhetorical question*. (To derive the question categories below, we also consider the specific labels *tag* and *declarative*, which are appended to one of the general labels.) All remaining utterances, in-

cluding backchannels and incomplete questions, are considered as non-questions, although we removed utterances that are very short (less than 200ms), have no transcribed words, or are missing segmentation times or dialogue act label. We performed minor text normalization on the transcriptions, such as mapping all word fragments to a single token.

The Wiki training set consists of close to 46k utterances, with 8.0% questions. We derived an MRDA training set of the same size from the training division of the original corpus; it consists of 6.6% questions. For the adaptation experiments, we used the full MRDA training set of 72k utterances as unlabeled adaptation data. We used two meetings (3k utterances) from the original MRDA development set for model selection and parameter tuning. The remaining meetings (in the original development and test divisions; 26k utterances) were used as our test set.

3.2 Features and Classifier

Lexical features consisted of unigrams through trigrams including start- and end-utterance tags, represented as binary features (presence/absence), plus a total-number-of-words feature. All ngram features were required to occur at least twice in the training set. The MRDA training set contained on the order of 65k ngram features while the Wiki training set contained over 205k. Although some previous work has used part-of-speech or parse features in related tasks, Boakye et al. (2009) showed no clear benefit of these features for question detection on MRDA beyond the ngram features.

We extracted 16 prosody features from the speech waveforms defined by the given utterance times, using stylized F0 contours computed based on Sönmez et al. (1998) and Lei (2006). The features are designed to be useful for detecting questions and are similar or identical to some of those in Boakye et al. (2009) or Shriberg et al. (1998). They include: F0 statistics (mean, stdev, max, min) computed over the whole utterance and over the last 200ms; slopes computed from a linear regression to the F0 contour (over the whole utterance and last 200ms); initial and final slope values output from the stylizer; initial intercept value from the whole utterance linear regression; ratio of mean F0 in the last 400-200ms to that in the last 200ms; number of voiced frames; and number of words per frame. All 16 features were z-normalized using speaker-level parameters, or gender-level parameters if the speaker had less than 10 utterances.

For all experiments we used logistic regression models trained with the LIBLINEAR package (Fan et al., 2008). Prosodic and lexical features were combined by concatenation into a single feature vector; prosodic features and the number-of-words were z-normalized to place them roughly on the same scale as the binary ngram features. (We substituted 0 for missing prosody features due to, e.g., no voiced frames detected, segmentation errors, utterance too short.) Our setup is similar to (Surendran and Levow, 2006), who combined ngram and prosodic features for dialogue act classification using a linear SVM. Since ours is a detection problem, with questions much less frequent than non-questions, we present results in terms of ROC curves, which were computed from the probability scores of the classifier. The cost parameter C was tuned to optimize Area Under the Curve (AUC) on the development set ($C = 0.01$ for prosodic features only and $C = 0.1$ in all other cases.)

3.3 Baseline Results

Figure 1 shows the ROC curves for the baseline Wiki-trained lexical system and the MRDA-trained systems with different feature sets. Table 2 compares performance across different question categories at a fixed false positive rate (16.7%) near the equal error rate of the MRDA (lex) case. For analysis purposes we defined the categories in Table 2 as follows: *tag* includes any yes-no question given the additional *tag* label; *declarative* includes any question category given the *declarative* label that is not a tag question; the remaining categories (*yes-no*, *or*, etc.) include utterances in those categories but not included in *declarative* or *tag*. Table 1 gives example sentences for each category.

As expected, the Wiki-trained system does worst on *declarative*, which have the syntactic form of statements. For the MRDA-trained system, prosody alone does best on *yes-no* and *declarative*. Along with lexical features, prosody is more useful for *declarative*, while it appears to be somewhat redundant with lexical features for *yes-no*. Ideally, such redundancy can be used together with unla-

yes-no	did did you do that?
declarative	you're not going to be around this afternoon?
wh	what do you mean um reference frames?
tag	you know?
rhetorical	why why don't we do that?
open-ended	do we have anything else to say about transcription?
or	and @frag@ did they use sigmoid or a softmax type thing?
or-after-YN	or should i collect it all?

Table 1: Examples for each MRDA question category as defined in this paper, based on Dhillon et al. (2004).

beled spoken utterances to incorporate prosodic features into the Wiki system, which may improve detection of some kinds of questions.

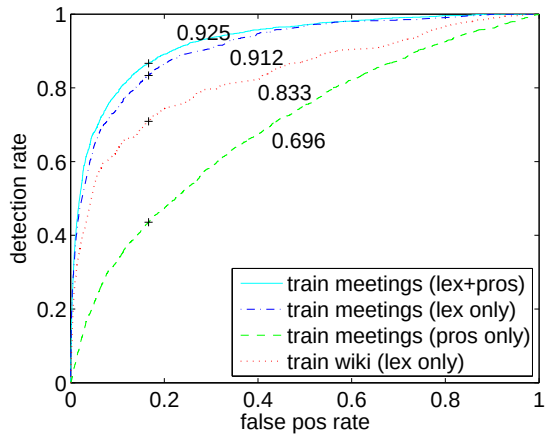

Figure 1: ROC curves with AUC values for question detection on MRDA; comparison between systems trained on MRDA using lexical and/or prosodic features, and Wiki talk pages using lexical features.

3.4 Adaptation Results

For bootstrapping, we first train an initial baseline classifier using the Wiki training data, then use it to label MRDA data from the unlabeled adaptation set. We select the k most confident examples for each of the two classes and add them to the training set using the guessed labels, then retrain the classifier using the new training set. This is repeated for r rounds. In order to use prosodic features, which are

type (count)	MRDA (L+P)	MRDA (L)	MRDA (P)	Wiki (L)
yes-no (526)	**89.4**	86.1	59.3	77.2
declar. (417)	**69.8**	59.2	49.4	25.9
wh (415)	**95.4**	93.0	42.2	92.8
tag (358)	89.7	**90.5**	26.0	79.1
rhetorical (75)	88.0	90.7	25.3	**93.3**
open-ended (50)	88.0	**92.0**	16.0	80.0
or (38)	97.4	**100**	29.0	89.5
or-after-YN (32)	**96.9**	**96.9**	25.0	90.6

Table 2: Question detection rates (%) by question type for each system (L=lexical features, P=prosodic features.) Detection rates are given at a false positive rate of 16.7% (starred points in Figure 1), which is the equal error rate point for the MRDA (L) system. Boldface gives best result for each type.

type (count)	baseline	bootstrap	SCL
yes-no (526)	77.2	**81.4**	**83.5**
declar. (417)	25.9	**30.5**	**32.1**
wh (415)	92.8	92.8	**93.5**
tag (358)	79.1	**79.3**	**80.7**
rhetorical (75)	93.3	*88.0*	*92.0*
open-ended (50)	80.0	*76.0*	80.0
or (38)	89.5	89.5	89.5
or-after-YN (32)	90.6	90.6	90.6

Table 3: Adaptation performance by question type, at false positive rate of 16.7% (starred points in Figure 2.) Boldface indicates adaptation results better than baseline; italics indicate worse than baseline.

available only in the bootstrapped MRDA data, we simply add 16 zeros onto the Wiki examples in place of the missing prosodic features. The values $k = 20$ and $r = 6$ were selected on the dev set.

In contrast with bootstrapping, SCL (Blitzer et al., 2006) uses the unlabeled target data to learn domain-independent features. SCL has generated much interest lately because of the ability to incorporate features not seen in the training data. The main idea is to use unlabeled data in both domains to learn linear predictors for many "auxiliary" tasks, which should be somewhat related to the task of interest. In particular, if $\mathbf{x}$ is a row vector representing the original feature vector and y_i represents the label for auxiliary task i, the linear predictor $\mathbf{w}_i$ is learned to predict $\hat{y}_i = \mathbf{w}_i \cdot \mathbf{x}'$ (where $\mathbf{x}'$ is a modified version of

$\mathbf{x}$ that excludes any features completely predictive of y_i.) The learned predictors for all tasks $\{\mathbf{w}_i\}$ are then collected into the columns of a matrix $\mathbf{W}$, on which singular value decomposition $\mathbf{USV}^T = \mathbf{W}$ is performed. Ideally, features that behave similarly across many y_i will be represented in the same singular vector; thus, the auxiliary tasks can tie together features which may never occur together in the same example. Projection of the original feature vector onto the top h left singular vectors gives an $h-$dimensional feature vector $\mathbf{z} \equiv \mathbf{U}_{1:h}^T \cdot \mathbf{x}'$. The model is then trained on the concatenated feature representation $[\mathbf{x}, \mathbf{z}]$ using the labeled source data.

As auxiliary tasks y_i, we identify all initial words that begin an utterance at least 5 times in each domain's training set, and predict the presence of each initial word ($y_i = 0$ or 1). The idea of using the initial words is that they may be related to the interrogative status of an utterance— utterances starting with "do" or "what" are more often questions, while those starting with "i" are usually not. There were about 250 auxiliary tasks. The prediction features $\mathbf{x}'$ used in SCL include all ngrams occuring at least 5 times in the unlabeled Wiki or MRDA data, except those over the first word, as well as prosody features (which are zero in the Wiki data.) We tuned $h = 100$ and the scale factor of $\mathbf{z}$ (to 1) on the dev set.

Figure 2 compares the results using the bootstrapping and SCL approaches, and the baseline unadapted Wiki system. Table 3 shows results by question type at the fixed false positive point chosen for analysis. At this point, both adaptation methods improved detection of *declarative* and *yes-no* questions, although they decreased detection of several other types. Note that we also experimented with other adaptation approaches on the dev set: bootstrapping without the prosodic features did not lead to an improvement, nor did training on Wiki using "fake" prosody features predicted based on MRDA examples. We also tried a co-training approach using separate prosodic and lexical classifiers, inspired by the work of Guz et al. (2007) on semi-supervised sentence segmentation; this led to a smaller improvement than bootstrapping. Since we tuned and selected adaptation methods on the MRDA dev set, we compare to training with the labeled MRDA dev (with prosodic features) and Wiki data together. This gives superior results compared to adaptation; but note that the adaptation process did not use labeled MRDA data to train, but merely for model selection. Analysis of the adapted systems suggests prosody features are being utilized to improve performance in both methods, but clearly the effect is small, and the need to tune parameters would present a challenge if no labeled speech data were available. Finally, while the benefit from 3k labeled MRDA utterances added to the Wiki utterances is encouraging, we found that most of the MRDA training utterances (with prosodic features) had to be added to match the MRDA-only result in Figure 1, although perhaps training separate lexical and prosodic models would be useful in this respect.

4 Conclusion

This work explored the use of conversational web text to detect questions in conversational speech. We found that the web text does especially poorly on *declarative* questions, which can potentially be improved using prosodic features. Unsupervised adaptation methods utilizing unlabeled speech and a small labeled development set are shown to improve performance slightly, although training with the small development set leads to bigger gains. Our work suggests approaches for combining large amounts of "naturally" annotated web text with unannotated speech data, which could be useful in other spoken language processing tasks, e.g. sentence segmentation or emphasis detection.

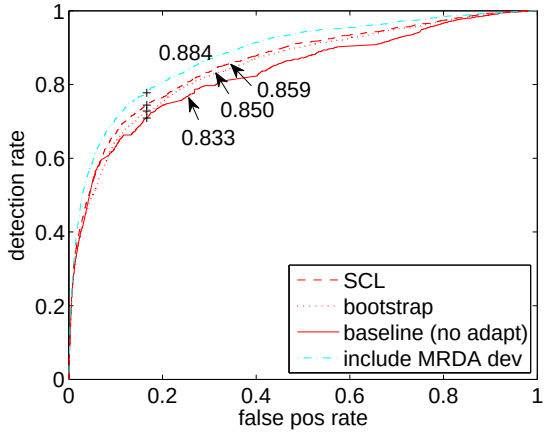

Figure 2: ROC curves and AUC values for adaptation, baseline Wiki, and Wiki + MRDA dev.

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. Int. Conference on Acoustics, Speech, and Signal Processing*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July. Association for Computational Linguistics.

Kofi Boakye, Benoit Favre, and Dilek Hakkini-tür. 2009. Any questions? Automatic question detection in meetings. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*.

Zhigang Chen, Guoping Hu, and Wei Jiang. 2010. Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction. In *Proc. Interspeech*.

Heidi Christensen, Yoshihiko Gotoh, and Steve Renals. 2001. Punctuation annotation using statistical prosody models. In *in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 35–40.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proc. of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, November.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, ICSI Tech. Report.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, August.

Agustin Gravano, Martin Jansche, and Michiel Bacchiani. 2009. Restoring punctuation and capitalization in transcribed speech. In *Proc. Int. Conference on Acoustics, Speech, and Signal Processing*.

Umit Guz, Sébastien Cuendet, Dilek Hakkani-Tür, and Gokhan Tur. 2007. Co-training using prosodic and lexical information for sentence segmentation. In *Proc. Interspeech*.

Umit Guz, Gokhan Tur, Dilek Hakkani-Tür, and Sébastien Cuendet. 2010. Cascaded model adaptation for dialog act segmentation and tagging. *Computer Speech & Language*, 24(2):289–306, April.

Jing Huang and Geoffrey Zweig. 2002. Maximum entropy model for punctuation annotation from speech. In *Proc. Int. Conference on Spoken Language Processing*, pages 917–920.

Minwoo Jeong, Chin-Yew Lin, and Gary G. Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259, Singapore, August. Association for Computational Linguistics.

Ji-Hwan Kim and Philip C. Woodland. 2003. A combined punctuation generation and speech recognition system and its performance enhancement using prosody. *Speech Communication*, 41(4):563–577, November.

Xin Lei. 2006. *Modeling lexical tones for Mandarin large vocabulary continuous speech recognition*. Ph.D. thesis, Department of Electrical Engineering, University of Washington.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech, and Language Processing*, 14(5):1526–1540, September.

Anna Margolis, Karen Livescu, and Mari Ostendorf. 2010. Domain adaptation with unlabeled data for dialog act tagging. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 45–52, Uppsala, Sweden, July. Association for Computational Linguistics.

Helena Moniz, Fernando Batista, Isabel Trancoso, and Ana Mata. 2011. Analysis of interrogatives in different domains. In *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, volume 6456 of *Lecture Notes in Computer Science*, chapter 12, pages 134–146. Springer Berlin / Heidelberg.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proc. 5th Conf. on Applied Natural Language Processing*, April.

Wenzhu Shen, Roger P. Yu, Frank Seide, and Ji Wu. 2009. Automatic punctuation generation for speech. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 586–589, December.

Elizabeth Shriberg, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech (Special Double Issue on Prosody and Conversation)*, 41(3-4):439–487.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100.

Kemal Sönmez, Elizabeth Shriberg, Larry Heck, and Mitchel Weintraub. 1998. Modeling dynamic prosodic variation for speaker verification. In *Proc. Int. Conference on Spoken Language Processing*, pages 3189–3192.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden Markov models. In *Proc. Interspeech*, pages 1950–1953.

Anand Venkataraman, Luciana Ferrer, Andreas Stolcke, and Elizabeth Shriberg. 2003. Training a prosody-based dialog act tagger from unlabeled data. In *Proc. Int. Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 272–275, April.