Detection of Agreement and Disagreement in Broadcast Conversations

Wen Wang[1] **Sibel Yaman**[2†*] **Kristin Precoda**[1] **Colleen Richey**[1] **Geoffrey Raymond**[3]
[1]SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA
[2]IBM T. J. Watson Research Center P.O.Box 218, Yorktown Heights, NY 10598, USA
[3]University of California, Santa Barbara, CA, USA
{wwang,precoda,colleen}@speech.sri.com, syaman@us.ibm.com, graymond@soc.ucsb.edu

Abstract

We present Conditional Random Fields based approaches for detecting agreement/disagreement between speakers in English broadcast conversation shows. We develop annotation approaches for a variety of linguistic phenomena. Various lexical, structural, durational, and prosodic features are explored. We compare the performance when using features extracted from automatically generated annotations against that when using human annotations. We investigate the efficacy of adding prosodic features on top of lexical, structural, and durational features. Since the training data is highly imbalanced, we explore two sampling approaches, random downsampling and ensemble downsampling. Overall, our approach achieves 79.2% (precision), 50.5% (recall), 61.7% (F1) for agreement detection and 69.2% (precision), 46.9% (recall), and 55.9% (F1) for disagreement detection, on the English broadcast conversation data.

1 Introduction

In this work, we present models for detecting agreement/disagreement (denoted (dis)agreement) between speakers in English broadcast conversation shows. The Broadcast Conversation (BC) genre differs from the Broadcast News (BN) genre in that it is more interactive and spontaneous, referring to free speech in news-style TV and radio programs and consisting of talk shows, interviews, call-in programs, live reports, and round-tables. Previous work on detecting (dis)agreements has been focused on meeting data. (Hillard et al., 2003), (Galley et al., 2004), (Hahn et al., 2006) used spurt-level agreement annotations from the ICSI meeting corpus (Janin et al., 2003). (Hillard et al., 2003) explored unsupervised machine learning approaches and on manual transcripts, they achieved an overall 3-way agreement/disagreement classification accuracy as 82% with keyword features. (Galley et al., 2004) explored Bayesian Networks for the detection of (dis)agreements. They used adjacency pair information to determine the structure of their conditional Markov model and outperformed the results of (Hillard et al., 2003) by improving the 3-way classification accuracy into 86.9%. (Hahn et al., 2006) explored semi-supervised learning algorithms and reached a competitive performance of 86.7% 3-way classification accuracy on manual transcriptions with only lexical features. (Germesin and Wilson, 2009) investigated supervised machine learning techniques and yields competitive results on the annotated data from the AMI meeting corpus (McCowan et al., 2005).

Our work differs from these previous studies in two major categories. One is that a different definition of (dis)agreement was used. In the current work, a (dis)agreement occurs when a responding speaker agrees with, accepts, or disagrees with or rejects, a statement or proposition by a first speaker. Second, we explored (dis)agreement detection in broadcast conversation. Due to the difference in publicity and intimacy/collegiality between speakers in broadcast conversations vs. meetings, (dis)agreement may have different character-

[†]This work was performed while the author was at ICSI.

istics. Different from the unsupervised approaches in (Hillard et al., 2003) and semi-supervised approaches in (Hahn et al., 2006), we conducted supervised training. Also, different from (Hillard et al., 2003) and (Galley et al., 2004), our classification was carried out on the utterance level, instead of on the spurt-level. Galley et al. extended Hillard et al.'s work by adding features from previous spurts and features from the general dialog context to infer the class of the current spurt, on top of features from the current spurt (*local* features) used by Hillard et al. Galley et al. used *adjacency pairs* to describe the interaction between speakers and the relations between consecutive spurts. In this preliminary study on broadcast conversation, we directly modeled (dis)agreement detection without using adjacency pairs. Still, within the conditional random fields (CRF) framework, we explored features from preceding and following utterances to consider context in the discourse structure. We explored a wide variety of features, including lexical, structural, durational, and prosodic features. To our knowledge, this is the first work to systematically investigate detection of agreement/disagreement for broadcast conversation data. The remainder of the paper is organized as follows. Section 2 presents our data and automatic annotation modules. Section 3 describes various features and the CRF model we explored. Experimental results and discussion appear in Section 4, as well as conclusions and future directions.

2 Data and Automatic Annotation

In this work, we selected English broadcast conversation data from the DARPA GALE program collected data (GALE Phase 1 Release 4, LDC2006E91; GALE Phase 4 Release 2, LDC2009E15). Human transcriptions and manual speaker turn labels are used in this study. Also, since the (dis)agreement detection output will be used to analyze social roles and relations of an *interacting* group, we first manually marked soundbites and then excluded soundbites during annotation and modeling. We recruited annotators to provide manual annotations of speaker roles and (dis)agreement to use for the supervised training of models. We defined a set of speaker roles as follows. *Host/chair* is a person associated with running the discussions or calling the meeting. *Reporting participant* is a person reporting from the field, from a subcommittee, etc. *Commentator participant/Topic participant* is a person providing commentary on some subject, or person who is the subject of the conversation and plays a role, e.g., as a newsmaker. *Audience participant* is an ordinary person who may call in, ask questions at a microphone at e.g. a large presentation, or be interviewed because of their presence at a news event. *Other* is any speaker who does not fit in one of the above categories, such as a voice talent, an announcer doing show openings or commercial breaks, or a translator.

Agreements and disagreements are composed of different combinations of initiating utterances and responses. We reformulated the (dis)agreement detection task as the sequence tagging of 11 (dis)agreement-related labels for identifying whether a given utterance is initiating a (dis)agreement opportunity, is a (dis)agreement response to such an opportunity, or is neither of these, in the show. For example, a *Negative tag question* followed by a negation response forms an agreement, that is, *A: [Negative tag] This is not black and white, is it? B: [Agreeing Response] No, it isn't.* The data sparsity problem is serious. Among all 27,071 utterances, only 2,589 utterances are involved in (dis)agreement as initiating or response utterances, about 10% only among all data, while 24,482 utterances are not involved.

These annotators also labeled shows with a variety of linguistic phenomena (denoted *language use constituents*, *LUC*), including discourse markers, disfluencies, person addresses and person mentions, prefaces, extreme case formulations, and dialog act tags (DAT). We categorized dialog acts into statement, question, backchannel, and incomplete. We classified disfluencies (DF) into filled pauses (e.g., *uh*, *um*), repetitions, corrections, and false starts. Person address (PA) terms are terms that a speaker uses to address another person. Person mentions (PM) are references to non-participants in the conversation. Discourse markers (DM) are words or phrases that are related to the structure of the discourse and express a relation between two utterances, for example, *I mean*, *you know*. Prefaces (PR) are sentence-initial lexical tokens serving functions close to discourse markers (e.g., *Well, I think* *that...*). Extreme case formulations (ECF) are lexical patterns emphasizing extremeness (e.g., *This is the best book I have ever read*). In the end, we manually annotated 49 English shows. We preprocessed English manual transcripts by removing transcriber annotation markers and noise, removing punctuation and case information, and conducting text normalization. We also built automatic rule-based and statistical annotation tools for these LUCs.

3 Features and Model

We explored lexical, structural, durational, and prosodic features for (dis)agreement detection. We included a set of "lexical" features, including n-grams extracted from all of that speaker's utterances, denoted *ngram* features. Other lexical features include the presence of negation and acquiescence, yes/no equivalents, positive and negative tag questions, and other features distinguishing different types of initiating utterances and responses. We also included various lexical features extracted from LUC annotations, denoted *LUC* features. These additional features include features related to the presence of prefaces, the counts of types and tokens of discourse markers, extreme case formulations, disfluencies, person addressing events, and person mentions, and the normalized values of these counts by sentence length. We also include a set of features related to the DAT of the current utterance and preceding and following utterances.

We developed a set of "structural" and "durational" features, inspired by conversation analysis, to quantitatively represent the different participation and interaction patterns of speakers in a show. We extracted features related to pausing and overlaps between consecutive turns, the absolute and relative duration of consecutive turns, and so on.

We used a set of prosodic features including pause, duration, and the speech rate of a speaker. We also used pitch and energy of the voice. Prosodic features were computed on words and phonetic alignment of manual transcripts. Features are computed for the beginning and ending words of an utterance. For the duration features, we used the average and maximum vowel duration from forced alignment, both unnormalized and normalized for vowel identity and phone context. For pitch and energy, we calculated the minimum, maximum, range, mean, standard deviation, skewness and kurtosis values. A decision tree model was used to compute posteriors from prosodic features and we used cumulative binning of posteriors as final features , similar to (Liu et al., 2006).

As illustrated in Section 2, we reformulated the (dis)agreement detection task as a sequence tagging problem. We used the Mallet package (McCallum, 2002) to implement the linear chain CRF model for sequence tagging. A CRF is an undirected graphical model that defines a global log-linear distribution of the state (or label) sequence E conditioned on an observation sequence, in our case including the sequence of sentences S and the corresponding sequence of features for this sequence of sentences F. The model is optimized globally over the entire sequence. The CRF model is trained to maximize the conditional log-likelihood of a given training set $P(E|S, F)$. During testing, the most likely sequence E is found using the Viterbi algorithm. One of the motivations of choosing conditional random fields was to avoid the label-bias problem found in hidden Markov models. Compared to Maximum Entropy modeling, the CRF model is optimized globally over the entire sequence, whereas the ME model makes a decision at each point individually without considering the context event information.

4 Experiments

All (dis)agreement detection results are based on n-fold cross-validation. In this procedure, we held out one show as the test set, randomly held out another show as the dev set, trained models on the rest of the data, and tested the model on the held-out show. We iterated through all shows and computed the overall accuracy. Table 1 shows the results of (dis)agreement detection using all features except prosodic features. We compared two conditions: (1) features extracted completely from the automatic LUC annotations and automatically detected speaker roles, and (2) features from manual speaker role labels and manual LUC annotations when manual annotations are available. Table 1 showed that running a fully automatic system to generate automatic annotations and automatic speaker roles produced comparable performance to the system using features from manual annotations whenever available.

Table 1: Precision (%), recall (%), and F1 (%) of (dis)agreement detection using features extracted from manual speaker role labels and manual LUC annotations when available, denoted *Manual Annotation*, and automatic LUC annotations and automatically detected speaker roles, denoted *Automatic Annotation*.

	Agreement		
	P	**R**	**F1**
Manual Annotation	81.5	43.2	56.5
Automatic Annotation	79.5	44.6	57.1

	Disagreement		
	P	**R**	**F1**
Manual Annotation	70.1	38.5	49.7
Automatic Annotation	64.3	36.6	46.6

We then focused on the condition of using features from manual annotations when available and added prosodic features as described in Section 3. The results are shown in Table 2. Adding prosodic features produced a 0.7% absolute gain on F1 on agreement detection, and 1.5% absolute gain on F1 on disagreement detection.

Table 2: Precision (%), recall (%), and F1 (%) of (dis)agreement detection using manual annotations without and with prosodic features.

	Agreement		
	P	**R**	**F1**
w/o prosodic	81.5	43.2	56.5
with prosodic	81.8	44.0	57.2

	Disagreement		
	P	**R**	**F1**
w/o prosodic	70.1	38.5	49.7
with prosodic	70.8	40.1	51.2

Note that only about 10% utterances among all data are involved in (dis)agreement. This indicates a highly *imbalanced* data set as one class is more heavily represented than the other/others. We suspected that this high imbalance has played a major role in the high precision and low recall results we obtained so far. Various approaches have been studied to handle imbalanced data for classifications, trying to balance the class distribution in the training set by either oversampling the minority class or downsampling the majority class. In this preliminary study of sampling approaches for handling imbalanced data for CRF training, we investigated two approaches, *random downsampling* and *ensemble downsampling*. *Random downsampling* randomly downsamples the majority class to equate the number of minority and majority class samples. *Ensemble downsampling* is a refinement of *random downsampling* which doesn't discard any majority class samples. Instead, we partitioned the majority class samples into N subspaces with each subspace containing the same number of samples as the minority class. Then we train N CRF models, each based on the minority class samples and one disjoint partition from the N subspaces. During testing, the posterior probability for one utterance is averaged over the N CRF models. The results from these two sampling approaches as well as the baseline are shown in Table 3. Both sampling approaches achieved significant improvement over the baseline, i.e., training on the original data set, and ensemble downsampling produced better performance than downsampling. We noticed that both sampling approaches degraded slightly in precision but improved significantly in recall, resulting in 4.5% absolute gain on F1 for agreement detection and 4.7% absolute gain on F1 for disagreement detection.

Table 3: Precision (%), recall (%), and F1 (%) of (dis)agreement detection without sampling, with random downsampling and ensemble downsampling. Manual annotations and prosodic features are used.

	Agreement		
	P	**R**	**F1**
Baseline	81.8	44.0	57.2
Random downsampling	78.5	48.7	60.1
Ensemble downsampling	79.2	50.5	61.7

	Disagreement		
	P	**R**	**F1**
Baseline	70.8	40.1	51.2
Random downsampling	67.3	44.8	53.8
Ensemble downsampling	69.2	46.9	55.9

In conclusion, this paper presents our work on detection of agreements and disagreements in English broadcast conversation data. We explored a variety of features, including lexical, structural, durational, and prosodic features. We experimented these features using a linear-chain conditional random fields model and conducted supervised training. We observed significant improvement from adding prosodic features and employing two sampling approaches, random downsampling and ensemble downsampling. Overall, we achieved 79.2% (precision), 50.5% (recall), 61.7% (F1) for agreement detection and 69.2% (precision), 46.9% (recall), and 55.9% (F1) for disagreement detection, on English broadcast conversation data. In future work, we plan to continue adding and refining features, explore dependencies between features and contextual cues with respect to agreements and disagreements, and investigate the efficacy of other machine learning approaches such as Bayesian networks and Support Vector Machines.

Acknowledgments

The authors thank Gokhan Tur and Dilek Hakkani-Tür for valuable insights and suggestions. This work has been supported by the Intelligence Advanced Research Projects Activity (IARPA) via Army Research Laboratory (ARL) contract number W911NF-09-C-0089. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, ARL, or the U.S. Government.

References

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of ACL*.

S. Germesin and T. Wilson. 2009. Agreement detection in multiparty conversation. In *Proceedings of International Conference on Multimodal Interfaces*.

S. Hahn, R. Ladner, and M. Ostendorf. 2006. Agreement/disagreement classification: Exploiting unlabeled data using constraint classifiers. In *Proceedings of HLT/NAACL*.

D. Hillard, M. Ostendorf, and E. Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of HLT/NAACL*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *Proc. ICASSP*, Hong Kong, April.

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 Transactions on Audio, Speech, and Language Processing*, 14(5):1526–1540, September. Special Issue on Progress in Rich Transcription.

Andrew McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The AMI meeting corpus. In *Proceedings of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*.