Identifying Agreement/Disagreement in Conversational Speech: A Cross-lingual Study

Wen Wang[1], *Kristin Precoda*[1], *Colleen Richey*[1], *Geoffrey Raymond*[2]

[1]Speech Technology and Research Lab
SRI International, Menlo Park, CA, USA
[2]University of California, Santa Barbara, CA, USA
{wwang,precoda,colleen}@.speech.sri.com, graymond@soc.ucsb.edu

Abstract

This paper presents models for detecting agreement/disagreement between speakers in English and Arabic broadcast conversation shows. We explore a variety of features, including lexical, structural, durational, and prosodic features. We experiment with these features using Conditional Random Fields models and conduct systematic investigations on efficacy of various feature groups across languages. Sampling approaches are examined for handling highly imbalanced data. 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; and 89.2% (precision), 30.1% (recall), 45.1% (F1) for agreement detection and 75.9% (precision), 28.4% (recall), and 41.3% (F1) for disagreement detection, on Arabic broadcast conversation data.

Index Terms: agreement/disagreement detection, sampling approaches, Conditional Random Fields, feature analysis

1. Introduction

In this work, we present models for detecting agreement/disagreement (denoted (dis)agreement) between speakers in English and Arabic 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. [1], [2], [3] used spurt-level agreement annotations from the ICSI meeting corpus [4]. [1] explored unsupervised machine learning approaches. [2] explored Bayesian Networks for the detection of (dis)agreements. They used adjacency pair information to determine the structure of conditional Markov models and outperformed the results of [1]. [3] explored semi-supervised learning algorithms and reached a competitive performance on manual transcriptions with only lexical features. [5] investigated supervised machine learning techniques and yielded competitive results on the annotated data from the AMI meeting corpus [6].

Our work differs from these previous studies in three major categories. First, 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, the genre of conversations we worked on is broadcast conversation in two languages, instead of English meeting data. Due to the difference in publicity and intimacy/collegiality between speakers in broadcast conversations vs. meetings, (dis)agreement may have different characteristics. Finally, different from the unsupervised approaches in [1] and semi-supervised approaches in [3], we conducted supervised training and systematic cross-lingual feature analysis. Also, different from [1] and [2], 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 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.

2. Data and Automatic Annotation

Human transcriptions and manual speaker turn labels are used in this study. Also, since the final goal of this study is to analyze social roles and relations of an *interacting* group, we first manually marked soundbites and then excluded soundbites during annotation and modeling. Soundbites are clip of speech played in broadcast shows that contain speech or interview quotations from speakers other than reporters and anchors in the show. The speakers in soundbites do not have interactions with speakers in the show. 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 tag-

ging 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. Table 1 shows the distribution of these labels in our manual annotations. Among all 27,071 English BC 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, revisions, and restarts. 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 and 74 Arabic shows. We preprocessed English manual transcripts by removing transcriber annotation markers and noise, removing punctuation and case information, and conducting text normalization. Additional preprocessing for Arabic includes Buckwalter romanization. We developed automatic annotation tools for discourse markers, extreme case formulations, and prefaces using rule-based systems integrating an HMM-based Part-of-Speech (POS) tagger, predefined tables, and heuristic rules. The automatic disfluency detection model was a hybrid system combining hidden-event language models, CRF based models, and rule-based models, for predicting fillers, repetitions, revisions, and restarts, following the approaches described in [7].

Table 1: Number of utterances with (dis)agreement-related labels in manual annotations for English and Arabic.

Label	English	Arabic
Positive declarative	331	452
Negative declarative	30	71
Positive interrogative	791	510
Interrogative which is negative	45	16
Negative interrogative	28	44
Positive tag question	55	16
Negative tag question	22	8
Agreeing response	673	524
Disagreeing response	278	178
Other response	336	275
Not involved in (dis)agreement	24,482	35,910

3. Features and Model

We explored lexical, structural, durational, and prosodic features for (dis)agreement detection. We included a set of "lexical" features composed of three categories, denoted *Ngram*, *LUC*, and *Disagr*. *Ngram* features include n-grams extracted from all of that speaker's utterances. *LUC* features were extracted from various LUC annotations, for example, the occurrence locations and counts of discourse markers, filled pauses, and disfluencies in the response, the dialog acts of the initiating utterance and the response, and occurrences and counts of ECFs in the initiating utterance and response. *Disagr* features include words presenting and indicating negative and acquiescence, yes/no equivalents, positive and negative tag questions, and a large variety of features distinguishing different types of initiating utterances and responses.

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. Examples of structural features include values corresponding to the following questions: whether there is overlap in time between initiating utterance and response, and speaker roles for preceding, current, and following utterance. Examples of duration features include values corresponding to the following questions: how long is the time between the end of the initiating utterance and the response; how long is the initiating utterance in seconds; how long is the response in seconds; and what is the ratio of the length of the initiating utterance to the length of the response.

We used a set of prosodic features including pause, fundamental frequency (F0), duration, and energy. Prosodic features were computed on words and phonetic forced alignment of manual transcripts. Features are computed for the beginning and ending words of an utterance. Similar to [1], we computed average, maximum, and initial pause duration features. For the duration features, we used the average and maximum vowel duration from the forced alignment, both unnormalized and normalized for vowel identity and phone context. F0 features include the distance from the average pitch in the word to the speaker's pitch floor and change in the averaged stylized pitch across a word boundary etc [8]. Energy features were computed similarly. A decision tree model was used to compute posteriors from prosodic features and we used cumulative binning of posteriors as final features, similar to the approach in [8]. In this work, we used the Mallet package [9] to implement the linear chain CRF model for sequence tagging.

4. Experiments

All (dis)agreement detection results are based on n-fold leave-one-out cross-validation. In this procedure, we held out one show as the test set, trained models on the rest of the data, and tested the model on the test set show. We iterated through all shows and computed the overall accuracy. First, we explored sampling approaches to handle the highly imbalanced data. As can be seen from Table 1, only about 10% utterances in the English data are involved in (dis)agreement as initiating or response utterances. The situations are similar in Arabic data. This indicates a highly *imbalanced* data set as one class is more heavily represented than the other/others. 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. We investigated two approaches for CRF training, *random downsampling* and *ensemble downsampling*. *Ran-*

dom downsampling randomly downsamples the majority class to equate the number of minority and majority class samples. Poorer performance on the majority class might be obtained as the sample space for the majority class gets reduced. *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 2, using lexical, structural, durational, and prosodic features. Both sampling approaches achieved significant improvement over the baseline, i.e., training on the original data set, and ensemble downsampling produced better performance than random downsampling. We noticed that both sampling approaches degraded slightly in precision but improved significantly in recall, resulting in 4.4% absolute gain on F1 for agreement detection and 4.7% absolute gain on F1 for disagreement detection. All of the results reported in the rest of this section are based on model training with ensemble downsampling. Note that we also compared the performance between CRF and Support Vector Machines for this task after applying downsamplings for both models and observed better performance from CRF. Since the CRF model is optimized globally over the entire sequence and better models dependencies for sequence tagging, the observation is reasonable.

Next, we systematically investigated the contributions from various feature groups on English. Results are shown in Table 3. Note that for *Ngram* features, we used unigrams and bigrams for English since we observed adding trigrams and higher order ngrams degraded the performance on both agreement and disagreement detection. We used unigram, bigram, and trigram for the Arabic system. As can be observed from the table, on English data, for agreement detection, Ngram features provided more gain when added to the Disagr baseline, than LUC or structural plus durational features; whereas, for disagreement detection, LUC features provided more complementarity. Adding structural and durational features on top of lexical features produced significant improvement on both agreement and disagreement detection. Adding prosodic features produced only minor improvement in agreement detection but improved disagreement F1 by 5% absolute.

A similar study of feature efficacy on the Arabic data is shown in Table 4. Different from their behaviors on the English data, LUC features provided more complementarity to Disagr than other features. And prosodic features improved both agreement and disagreement detection, 0.7% absolute for the former and 1% absolute for the latter. On both languages, lexical features perform much better than structural and durational features, consistent with the findings on the ICSI meeting data from [1, 2]. On both languages, structural, durational, and prosodic features provided complementarity to lexical features as adding them sequentially produced improvement in recall and F1 for both agreement and disagreement detection.

Table 5 compares (dis)agreement detection performance on English and Arabic between 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. It can be seen from the table that on English, running a fully automatic system to generate automatic annotations and automatic speaker roles yielded 1% and 2% absolute degradation on (dis)agreement F1 compared to the system using features from manual annotations whenever available. On Arabic, this degradation is increased to about 10% absolute on F1 for both agreement and disagreement detection. We will investigate approaches to improve performance of automatic annotation tools as well as exploring more robust features for Arabic (dis)agreement detection, for example, employing morphological analysis in annotations and feature extractions.

In conclusion, this paper presents our work on detection of agreements and disagreements in English and Arabic 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. We conducted systematic investigations on efficacy of various feature groups and examined two sampling approaches for handling imbalanced data. 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; and 89.2% (precision), 30.1% (recall), 45.1% (F1) for agreement detection and 75.9% (precision), 28.4% (recall), and 41.3% (F1) for disagreement detection, on Arabic 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, investigate the efficacy of other machine learning approaches such as Bayesian networks, and study the effectiveness of our modeling approaches on other genres such as meeting data and other languages.

Acknowledgments 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.

5. References

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

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

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

[4] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI Meeting Corpus", *in Proc. ICASSP*, vol. 1, pp. 364–367, Hong Kong, Apr. 2003.

[5] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation", *in Proceedings of International Conference on Multimodal Interfaces*, 2009.

[6] 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, "The AMI meeting corpus", *in Proceedings of Measuring Behavior 2005, the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[7] W. Wang, G. Tur, J. Zheng, and N. F. Ayan, "Automatic disfluency removal for improving spoken language translation", *in Proc. ICASSP*, 2010.

[8] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1526–1540, Sep. 2006, Special Issue on Progress in Rich Transcription.

[9] A. McCallum, "Mallet: A machine learning for language toolkit", 2002, http://mallet.cs.umass.edu.

Table 2: Precision, recall, and F1 of (dis)agreement detection on **English** data without sampling, with random downsampling and ensemble downsampling, using lexical, structural, durational, and prosodic features. Manual annotations are used for computing LUC and structural features.

	Agreement			**Disagreement**		
	Precision (%)	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
Baseline	81.81	44.02	57.24	70.78	40.12	51.21
Random Downsampling	78.46	48.67	60.07	67.31	44.77	53.77
Ensemble Downsampling	79.18	50.52	61.68	69.18	46.94	55.93

Table 3: Precision (%), recall (%), and F1 (%) of **English** (dis)agreement detection using ensemble downsampling and features extracted from manual speaker role labels and manual LUC annotations when available.

	English					
	Agreement			**Disagreement**		
Feature Sets	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
a. Structural + Durational	87.62	17.56	29.25	69.39	19.10	29.96
b. Disagr	72.54	45.91	56.23	62.35	36.81	46.29
c. Disagr + Ngram	80.94	46.06	58.71	66.24	36.11	46.74
d. Disagr + LUC	74.87	43.83	55.30	63.95	38.19	47.83
e. Disagr + Structural + Durational	72.60	44.87	55.46	55.43	35.42	43.22
f. Lexical (= c. + LUC)	79.84	45.32	57.82	66.47	38.54	48.79
g. Lexical + Structural + Durational	83.29	48.89	61.61	67.82	40.97	51.08
h. All (g. + Prosodic)	79.18	50.52	61.68	69.18	46.94	55.93

Table 4: Precision (%), recall (%), and F1 (%) of **Arabic** (dis)agreement detection using ensemble downsampling and features extracted from manual speaker role labels and manual LUC annotations when available.

	Arabic					
	Agreement			**Disagreement**		
Feature Sets	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
a. Structural + Durational	80.23	13.17	22.62	79.52	12.60	21.75
b. Disagr	81.63	22.47	35.24	68.23	19.42	30.23
c. Disagr + Ngram	82.33	23.29	36.31	69.73	20.00	31.08
d. Disagr + LUC	82.35	24.31	37.54	69.92	22.66	34.23
e. Disagr + Structural + Durational	80.01	24.16	37.22	62.33	23.64	34.28
f. Lexical (= c. + LUC)	81.56	27.25	40.85	70.62	25.38	37.34
g. Lexical + Structural + Durational	91.35	29.28	44.35	75.43	27.51	40.32
h. All (g. + Prosodic)	89.22	30.13	45.05	75.89	28.40	41.33

Table 5: Precision, recall, and F1 of (dis)agreement detection on **English** and **Arabic** data using all features and ensemble sampling on two conditions of using manual and completely automatic annotations for computing LUC and structural features, respectively.

	English					
	Agreement			**Disagreement**		
	Precision (%)	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
Manual annotation	79.18	50.52	61.68	69.18	46.94	55.93
Automatic annotation	80.94	48.59	60.72	66.58	45.02	53.72
	Arabic					
	Agreement			**Disagreement**		
	Precision (%)	**Recall (%)**	**F1 (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
Manual annotation	89.22	30.13	45.05	75.89	28.40	41.33
Automatic annotation	81.63	22.47	35.24	86.61	18.51	30.50