DETECTING LEADERSHIP AND COHESION IN SPOKEN INTERACTIONS

Wen Wang, Kristin Precoda, Raia Hadsell, Zsolt Kira, Colleen Richey, Gabriel Jiva

SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA
{wwang,precoda,colleen}@speech.sri.com, {raia.hadsell,zsolt.kira,gabriel.jiva}@sri.com

ABSTRACT

We present a system for detecting leadership and group cohesion in multiparty dialogs and broadcast conversations in English and Mandarin. We systematically investigate the impact of features and designs of the prediction systems, the relationships between features and their individual significance in logistic regressions, and the contributions of feature groupings as predictors for leader and group cohesion, across genres and languages. We achieve 73.0% to 94.7% F1 accuracy for leader detection and around 80% F1 accuracy for group cohesion detection, on all data sets.

***Index Terms*—** multiparty dialog, broadcast conversation, leader detection, group cohesion, feature analysis

1. INTRODUCTION

There has been much effort recently to develop methods to correlate the social goals of the members of a group with the language that they use, including the work on modelling and automatically detecting speaker roles, social relations, and speaker characteristics from spoken interactions [1, 2, 3, 4, 5, 6, 7]. The goal of this work is to develop systems for automatically analyzing group and social relations of an interacting group, that is, who is the leader of a group, whether a group exhibits cohesive, divisive, or mixed relations. In a multiparty conversation, the leader is the speaker who guides the group toward an outcome, controls the group discussion, manages interactions of the group, and is recognized as such by other members of the group. Generally, most speakers in a multiparty conversation are not the leader. Leaders set the agenda, make decisions, run the discussion, choose who will speak next, and call and end to the interaction. Other participants may occasionally perform one or another of these functions but do not consistently do so throughout the interaction. We categorized cohesiveness of a group into cohesive, divisive, or mixed interactions. Cohesive interactions consist mostly of agreement and alignment with very few and relatively minor disagreements and other forms of rejection. Divisive interactions are characterized by strong disagreement in places or by more disagreements than agreements. Other interactions may have a mixture of agreement and disagreement, with no particularly charged or heated disagreements, or offer insufficient information to make a determination. Our work is informed by the insights and theories of conversation analysis and built upon machine learning approaches.

2. SYSTEM

We explored three data resources, namely, the ICSI meeting corpus [8], the U.S. Nuclear Waste Technical Review Board discussion transcriptions (denoted NWTRB)[1], and the Mandarin broadcast conversation transcriptions released under the DARPA GALE program. Note that the NWTRB transcripts do not have any timing information. Also, the transcripts include regions when a single speaker made reports and hence do not present spoken interactions. Therefore, we developed a tool to extract NWTRB excerpts based on speaker interactivities. We manually annotated leaders and group cohesion relations for conducting supervised training. Since conversation analysis studies suggest that agreement/disagreement (denoted (dis)agreement) detection is essential for analysis of group cohesion, we also manually annotated (dis)agreement and built an automatic (dis)agreement detection system using linear chain Conditional Random Fields (CRF). 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. 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*. In the end, for ICSI meeting data, we manually annotated 75 meetings for leader, 37 meetings for (dis)agrement, and 46 meetings for group cohesion. For NWTRB, we manually annotated 186 excerpts for leader, 35 excerpts for (dis)agreement, and 35 excerpts for group cohesion. For Mandarin BC data, we manually annotated 36 shows for leader, 52 shows for (dis)agremeent, and 36 shows for group cohesion.

We explored features based on a variety of linguistic phenomena (denoted *language use constituents*, or *LUC*), including discourse markers, disfluencies, 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. 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*. Extreme case formulations (ECF) are lexical patterns emphasizing extremeness (e.g., *This is the best book I have ever read*). We developed automatic annotation tools for discourse markers and extreme case formulations using rule-based systems integrating an HMM-based part-of-speech (POS) tagger, predefined tables, and heuristic rules. We developed an automatic dialog act tagger using AdaBoost and word n-gram features. 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 [9]. We also designed features inspired by various conversation analysis hypotheses and a close review of existing features and their interrelationships across multiple data sets. For example, for leader detection, we

[1] http://www.nwtrb.gov/meetings/meetings.html

explored features related to pausing and interruptions, and interaction management lexical cues. For cohesion detection, besides features related to interactivity and presence of (dis)agreements, we also explored features related to usage of time and extreme case formulations. The full feature set for leader detection is 80 dimension besides word n-gram features, and 40 dimension for group cohesion. Studies on feature analyses are presented in Section 3.

We used the AdaBoost algorithm for classifying leader and group cohesion both for interpretability and for its high discriminative classification performance. Boosting aims at combining weak base classifiers to come up with a strong classifier. The learning algorithm is iterative. During each iteration, a different distribution or weighting over the training samples is used to give more emphasis on samples that were often misclassified by the preceding weak classifiers. We used the BoosTexter tool [10], which handles both discrete and continuous features and enables a convenient incorporation of various features without the need for binning. We used CRF for detecting (dis)agreements, as described in detail in [11].

3. FEATURE ANALYSIS

In our leader and group cohesion features, there are multiple variants of a feature that are expected to be highly correlated and we needed to determine the most appropriate operationalization of the feature. In other cases, there are features that turn out to be highly correlated because of characteristics of the data. In both of these situations, the machine learning algorithms may not have any theoretical basis for choosing among highly correlated features and their choice may not be as sensible as possible. Therefore, we aimed at discarding features and manually selecting a smaller number of features to use before employing machine learning algorithms.

Some of the most significant predictors for leaders in the ICSI meetings are the number of { statements, questions, backchannels, incomplete utterances} per unit time, the percentage of { statements, questions, backchannels, incomplete utterances} by this speaker out of all such utterances in the interaction; the percentage of sentences and of turns by this speaker in the interaction; the number of wh-questions and non-wh- questions; the average number of discourse markers per sentence; the average number of words per sentence; the average number of times per turn a speaker was interrupted, percent of interruptions produced by this speaker, percent of interruptions experienced by this speaker; percent produced by this speaker out of all (dis)agreement initiating utterances, percent produced by this speaker out of {all, positive, negative} responses to (dis)agreement opportunities, and percent of positive, other responses received by this speaker in response to an initiating utterance; and the ratio of the degree and weighted degree of this speaker to the total number of speakers. Some of the most significant predictors for leaders in the NWTRB excerpts are the average number of discourse markers per sentence; the average number of words per sentence and per turn; the ratio of the degree and out-degree of this speaker to the total number of speakers. The modest overlap between these two lists of features could be caused by several factors. One factor is that several of the significant predictors for ICSI meetings cannot be calculated for NWTRB excerpts because of the loss of data in the NWTRB transcripts (e.g., disfluencies, time stamps). Predictors which cannot be calculated include the features relating to interruptions and accurate counts of backchannels and incomplete utterances. Another factor is the different nature of the interactions in the ICSI meetings and the NWTRB excerpts.

Features that were found to be of greatest significance for both the ICSI and NWTRB corpora for (individually) predicting cohesion include ratio of the number of agreements to the sum of the number of agreements and disagreements, total number of agreements per hour in the interaction, difference between number of agreements and number of disagreements per hour, total number of disagreements in the interaction divided by total number of speakers, total number of disagreements in the interaction divided by approximate entropy of speaker distribution, total number of unique extreme case formulations used per hour in an interaction, and total number of speakers in the interaction.

There were also several features that were significant for the ICSI meetings but were undefined for the NWTRB excerpts, because they rested on information that is not available in the NWTRB transcripts. For example, features relating to the amount of silence in an interaction were by far the most significant predictors of cohesion for the ICSI meetings. It is understandable that silence could be a predictor of cohesiveness, in that participants do not have to choose their words carefully and can contribute quickly and without hesitation. Unfortunately these features could not be used for NWTRB excerpts, which have no timing information.

We grouped features into several language uses based on the design of the features, as shown in Table 1. The contributions of language uses are a combination of the contributions from each of potentially many features, normalized for presentation as a percentage. In some cases a feature may in fact contribute negatively to a prediction decision, which is therefore made in spite of that feature. These are shown as a zero contribution, as a negative percentage is not interpretable. The language use contribution scores for leader and group cohesion are shown in Table 3. These contribution scores should be helpful to analysts in understanding not just what aspects of language behavior led to a decision, but also some sense of the relative importance of the different aspects.

4. EXPERIMENTS

(Dis)agreement-related features are used in both the speaker role and group cohesion detection models, and the speaker role detection outputs are also used in the group cohesion models. In single-pass system I, we built a leader detection system in the first step when the automatic (dis)agreement labels are not available. Hence, we didn't use (dis)agreement-related features for speaker role labeling. After leader labeling, we built a (dis)agreement detection system and then used the automatic leader and (dis)agremeent labels for building the cohesion classification system with AdaBoost. However, in the feature analysis with manual (dis)agreement labels described in Section 3, we found (dis)agreement-related features significantly correlating to speaker role modelling. Hence, we designed the alternative single-pass system II, where first (dis)agreement models are trained, and then those automatic labels are used in speaker role model training. Next, the automatic (dis)agreement and leader labels are used for building the cohesion detection system. The results are shown in Table 2. We observed significant improvement on leader detection in the single-pass system II over single-pass system I. We then extended the single-pass system II to a multi-pass cohesion detection system, by conducting another pass of (dis)agreement model training after leader detection and another pass of cohesion detection model training at the end. Note that the second round of (dis)agreement model training can use leader-related features and hence might improve its own performance and benefit second-pass cohesion detection. We observed significant improvement in second-stage cohesion detection on ICSI meeting data, while there was no gain on NWTRB excerpts and Mandarin BC data. We found that this is due to the sparseness of (dis)agreement in NWTRB and Mandarin BC compared to ICSI meeting data.

We computed the average contribution score for each LU for

Table 1. Language uses as clusters of features for leadership and cohesiveness detection.

Leader detection: Language Use Name	**Examples of features**
1 Quantity of verbal contributions	Total number of questions by this speaker per unit time with wh-question words
2 Interactivity with other speakers	Total number of unique speakers that this speaker talked to
3 Pausing	Average length of a pause within a turn by this speaker
4 Interruptions	Percent of this speaker's utterances that interrupt an utterance by someone else
5 Participation in disagreements	Percent of this speaker's utterances that contain an utterance initiating a (dis)agreement opportunity
6 Use of extreme case formulations	Average number of unique extreme case formulations per sentence used by this speaker
7 Interaction management phrases	Total number of times a speaker used words or phrases associated with interaction management per unit time
8 Sentence types used	Average number of disfluencies per sentence produced by this speaker
9 Discourse markers used	Average number of discourse markers per sentence by this speaker
10 Lexical cues	N-grams by this speaker
Cohesiveness detection: Language Use Name	**Examples of features**
1 Interactivity and presence of (dis)agreements	Total number of agreements in the interaction, per hour
2 Types of utterances initiating (dis)agreements	Number of declarative statements disagreed with or rejected by a response, per hour
3 Usage of time	Total silence time during the interaction (no one talking), per hour
4 Use of extreme case formulations	Total number of unique extreme case formulations used in an interaction, per hour

Table 2. Comparison of single-pass and multi-pass Precision (%), recall (%), and F1 (%) of Leader and Cohesion detection on ICSI meetings (with a subset of manual annotations for supervised training).

	ICSI meetings					
	Leader			**Cohesion**		
System	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Single-pass system I	86.0	56.9	68.5	77.8	87.5	82.4
Single-pass system II	84.0	64.6	73.0	63.3	79.2	70.4
Multi-pass cohesion detection	84.0	64.6	73.0	77.8	87.5	82.4

leader and cohesion prediction, on ICSI meetings, NWTRB, and Mandarin BC data, as shown in Table 3. For leader detection, *lexical cues* contribute significantly to leader detection across the three corpora. *Quantity of verbal contributions* is another significant contributor across the three corpora. *Participation in (dis)agreements* contributes significantly to ICSI meeting data leader detection, but its contribution was reduced on the NWTRB and Mandarin BC data. *Interactivity with other speakers* plays an important role for NWTRB and Mandarin BC data, but quite minor for the ICSI meeting data. For cohesion detection, it is interesting to observe that *usage of time* contributes the most on the ICSI and Mandarin BC data, which have valid timing information. Nevertheless, *Interactivity and presence of (dis)agreements* and *Types of utterances initiating (dis)agreements* contribute significantly across the three corpora, except Mandarin BC, on which we obtained the lowest automatic (dis)agreement detection performance due to data sparsity and possibly insufficient feature design for (dis)agreement detection. Hence, the automatic (dis)agreement labels on Mandarin BC data are not quite reliable.

Table 4 shows the comparison between using the full feature set and the manually selected subset of features for leader and cohesion detection, in ICSI meetings, NWTRB excerpts, and Mandarin BC data sets. Since both NWTRB and Mandarin BC models are relatively undertrained due to sparsity of training samples, we noticed comparable performance from the subset features (an improvement on cohesion detection on NWTRB). On the other hand, for the better-trained ICSI meeting models, the full feature sets outperformed the subset features.

The manual cohesiveness labeling of the Mandarin BC shows is highly imbalanced. Among 36 shows, only two shows were labeld as "mixed" and the rest were labled as "cohesive". Building on the observation that certain feature subsets generalize better than others even within the same genre, we decided to explore feature subsets that achieve high recall for the cohesive class (hence not missing the dominant class of this genre) while at the same time classifying obvious cases of less-than-cohesiveness with high precision. After exploration of several subsets, we decided upon 18 features such as (dis)agreement and silences that are well-separated between cohesive and less-than-cohesive instances. In order to validate the subset, we used two tests. First, we verified high recall for cohesive instances for both ICSI and NWTRB, using the same feature subset and same classifier learned with ICSI-only data. This verifies that the subset is general enough that it transfers across genres. Since most Mandarin BC shows are labeled as cohesive, we also verified that the classifier labeled as many of these data as cohesive as pos-

Table 3. Average LU contributions to leader and cohesion detection, represented as percentages.

Leader Detection			
LU	ICSI meetings (%)	NWTRB (%)	Mandarin BC (%)
1 Quantity of verbal contributions	11.3	12.3	19.4
2 Interactivity with other speakers	1.5	27.4	8.1
3 Pausing	4.0	N/A	1.1
4 Interruptions	2.4	N/A	1.7
5 Participation in (dis)agreements	15.7	5.2	4.4
6 Use of extreme case formulations	3.5	1.1	0.3
7 Interaction management phrases	6.6	10.5	7.2
8 Sentence types used	2.9	5.4	2.2
9 Discourse markers used	0.1	1.4	4.4
10 Lexical cues	52.0	36.7	51.2
Cohesion Detection			
LU	ICSI meetings (%)	NWTRB (%)	Mandarin BC (%)
1 Interactivity and presence of (dis)agreements	30.3	40.9	30.7
2 Types of utterances initiating (dis)agreements	18.4	36.0	0.6
3 Usage of time	34.5	N/A	35.8
4 Use of extreme case formulations	16.8	19.5	1.7

Table 4. Impact of subsetting features on SC detection in ICSI meetings, NWTRB excerpts, and Mandarin BC data. Results are presented in Precision (P) (%), Recall (R) (%), and F1 (%).

	ICSI meetings					
	Leader			**Cohesion**		
	P	**R**	**F1**	**P**	**R**	**F1**
Full	84.0	64.6	73.0	79.2	79.2	79.2
Subset	88.6	60.0	71.6	61.5	66.7	64.0
	NWTRB excerpts					
	Leader			**Cohesion**		
	P	**R**	**F1**	**P**	**R**	**F1**
Full	96.9	78.8	86.9	77.8	73.7	75.7
Subset	95.4	79.5	86.7	83.3	78.9	81.1
	Mandarin BC					
	Leader			**Cohesion**		
	P	**R**	**F1**	**P**	**R**	**F1**
Full	95.2	94.1	94.7	94.1	94.1	94.1
Subset	96.3	91.8	94.0	94.1	94.1	94.1

sible. While we have obtained a small amount of less-than-cohesive Mandarin data to train on subsequent to these experiments, the analysis here represents an interesting study of making up for sparsity of data in one language by utilizing a classifier learned in another language.

In conclusion, we systematically investigated the contributions of language uses to social construct prediction, and the impact of single-pass or multi-pass social construct predictions, for ICSI, NWTRB, and Mandarin genres. We examined the leader and cohesion features for their relationships to each other and for their individual significance as predictors in logistic regressions, in both the ICSI meeting data and a set of NWTRB excerpts. Using a manually selected smaller set of features, we obtained comparable or better precision, recall, and F1-score on both the NWTRB and Mandarin BC data than with a larger set of features. We believe that using a smaller set of features based on feature correlation study may increase the generality of the learned models. We are currently exploring semi-supervised training approaches and exploring new features for leader and cohesion detection, for example, features related to topic shift for leader detection. We will also explore the effectiveness of the systems on other genres and languages.

Acknowledgments The authors thank the effort of our annotators. This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through Army Research Laboratory (ARL) contract number W911NF-09-C-0089. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

5. REFERENCES

[1] Regina Barzilay, Michael Collins, Julia Hirschberg, and Steve Whittaker, "The rules behind roles: Identifying speaker role in radio broadcasts," in *Proceedings of AAAI*, 2000, pp. 679–684.

[2] Yang Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proceedings of HLT/NAACL*, 2006, pp. 81–84.

[3] Brian Hutchinson, Bin Zhang, and Mari Ostendorf, "Unsupervised broadcast conversation speaker role labeling," in *Proceedings of ICASSP*, 2010, pp. 5322–5325.

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

[5] 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.

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

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

[8] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters, "The ICSI Meeting Corpus," in *Proc. ICASSP*, Hong Kong, Apr. 2003, vol. 1, pp. 364–367.

[9] Wen Wang, Gokhan Tur, Jing Zheng, and Necip Fazil Ayan, "Automatic disfluency removal for improving spoken language translation," in *Proc. ICASSP*, 2010.

[10] Robert E. Schapire and Yoram Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.

[11] Wen Wang, Kristin Precoda, Colleen Richey, and Geoffrey Raymond, "Identifying agreement/disagreement in conversational speech: A cross-lingual study," in *Proceedings of Interspeech*, 2011.