DETECTING TARGETS OF ALIGNMENT MOVES IN MULTIPARTY DISCUSSIONS

Alex Marin, Wei Wu, Bin Zhang, Mari Ostendorf

Department of Electrical Engineering, University of Washington, Seattle, WA 98195

ABSTRACT

In analyzing goal-oriented multiparty discussions, one challenge is to determine who is responding to whom when they are supporting or opposing a remark put forward by another participant. This paper looks at algorithms for detecting the target discussant, comparing findings of important features for three genres of text and spoken discussions. Comparing to the common baseline of "previous speaker," we find gains from considering content and semantic similarity in target detection, but with substantial differences in accuracy and important features across genres.

***Index Terms*—** Target detection, alignment moves, agreement/disagreement, semantic similarity

1. INTRODUCTION

There is increasing availability of multiparty discussions, both from online interactions and teleconference recordings, and therefore interest in automatic analysis of group interactions. In goal-oriented discussions, participants frequently support or oppose ideas or claims put forward by another participant to improve their group standing or express solidarity with a subgroup of discussants; we refer to such interaction behavior as positive or negative alignment moves [1]. The alignment moves represent a subset of adjacency pairs, which have been studied as fundamental units of conversational organization [2]. (Adjacency pairs would also include question-answer pairs, offer-accept, etc., which we do not consider alignment moves.) Alignment moves are related to agreement and disagreement, but differ in that alignment is motivated by social interactions, whereas agreement can appear as part of information exchange, e.g. as in question answering. Alignment moves are also related to, but not the same as, entrainment (e.g. as discussed by Nenkova et al. [3]), in that entrainment may occur as part of an alignment move, but is not required.

As an example of alignment moves, consider the following example with a position statement followed by a negative alignment move from a Wikipedia discussion thread. [Atfyfe: *"Even if you do not qualify Iraq a war, the arguement that it is not because Congress has not formally declared war is bizzare."* Nescio: *"Bizarre or not, in the legal sense the assertion "there can be no war without a declaration of war" is technically correct."*]

In multi-party conversations, the strict requirement of elements of an adjacency pair to be adjacent often must be relaxed, to allow for interruptions from other speakers, backchannels, disfluencies, answers to a question from multiple speakers, and so on; some of these phenomena are also noted in conversational analysis literature [4]. Such phenomena are even more important in asynchronous conversations, such as online discussion forums; there, participants may not only interrupt each other, but also go back to earlier parts of a conversation to pick up an abandoned or postponed thread, the information about the past discussion points being readily available, unlike in live discussions. Since discussants often do not refer to the target of their alignment by name, detecting the target is an added challenge to detecting the alignment move.

The problem of detecting alignment moves is similar to work on agreement/disagreement detection in meeting data [5, 6, 7, 8] and in broadcast conversations [9, 10]. However, in these studies there was no target detection. Galley et al. [6] do explore target detection for the full set of adjacency pairs in the ICSI meeting corpus using lexical, structural, and temporal features in a maximum entropy model. Starting with a baseline accuracy of 80% using the simple previous speaker rule, they obtain 87% with backward-looking features as used here and 90% using information from future utterances, in all cases assuming that the responding part of the pair is known.

In this paper, we focus on identifying the target of alignment moves in three different goal-oriented, multi-party conversation genres: live in-person conversations (ICSI meetings), live online conversations (IRC chat), and asynchronous forum conversations (Wikipedia discussions). Starting with labels of only the utterances of the responding speaker (the "source" of the alignment move), we attempt to select the original speaker (the "target" of the alignment move) without trying to identify which specific utterance made by that speaker had been involved in the alignment.

2. DATA

We use three data sets in our experiments, chosen to give us different characteristics of language. While all three genres are conversational and goal-oriented in nature, we selected genres that allow us to compare synchronous and asynchronous discussions, as well as written and spoken genres. All data is annotated for alignment moves with targets at the utterance or sentence level, using the annotation scheme presented in [1].

The ICSI meetings corpus [11] is a collection of multi-party discussions held by ICSI research groups on a variety of technical topics. We select sections of each meeting with the highest amount of interactivity, based on the meeting acts annotations described by Bates et al. [12]. Specifically, meeting fragments labeled as containing negotiation, reporting, brainstorming, or other discussion were selected as candidates for annotation. Twenty hand-transcribed meeting segments were selected for annotation. One annotator added alignment move labels to each meeting fragment.

The chat corpus comprises four 45-minute unscripted, goal-oriented IRC conversations. All conversations have the same goal (planning a student party). There are four participants in each conversation. Each participant is assigned a role, which can be one of project manager, secretary, accountant, and publicity coordinator.

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). 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.

Exactly one participant is assigned to each role. Again, only one annotator labeled this corpus for alignment moves.

The Wikipedia discussions data was obtained from the AAWD corpus [1]. The data in the AAWD corpus consists of threads from discussion pages associated with various English Wikipedia articles. Each thread consists of all the posts related to a particular topic, as they appeared in a Wikipedia database dump from 2008. The discussion participants (Wikipedia editors) may or may not be familiar with each other online, but almost never know each other in person. The discussions are asynchronous. Three annotators labeled alignment moves in a subset of the corpus, resulting in 190 annotated files. The inter-annotator agreement was $\kappa = 0.86$.

Corpus statistics are presented in table 1. There are significant differences in the distribution of alignment moves as well as the average utterance length within a corpus. While there are more positive alignment moves than negative ones in the two synchronous discussion genres, there are more negative claims in the asynchronous discussion genre. The neutral class always dominates, with even the most dense alignment moves being outnumbered 4-1 by the utterances with no moves. The average utterance length is much longer in Wikipedia discussions than in both chat and meetings data.

Statistic	Meetings	Chat	Wiki
# Files	20	4	190
# Utterances	2639	2229	15799
# Positive Alignments	490	238	589
# Negative Alignments	174	78	1898
Ave Utterance Length	9	5	15

Table 1. Corpus Statistics

3. CONTENT-BASED DETECTION

Our first method attempts to select the target of an alignment move based on the amount of topical content in the potential target utterance, motivated by anecdotal evidence that most interruptions, backchannels, and fragments tend to be short and contain little topical content. A related feature in [6] is the number of content words in the target sentence. We implement a detection system which selects the closest utterance to the source of the alignment move made by a different speaker, which contains "sufficient" topical content. Thus, we score each utterance of all previous speakers, starting with the closest, until we find one whose score is above a given threshold. The threshold is a tunable parameter of the system.

3.1. Score Computation

Inspired by work in information retrieval, we use term frequency–inverse document frequency (TF-IDF) statistics to quantify the topical content of an utterance. We treat each utterance as a "document" and compute the TF-IDF score of word type w_i in document d_j as

$$\text{TF-IDF}(w_i, d_j) = c(w_i, d_j) \times \log \frac{|D|}{|\{d_k \in D : w_i \in d_k\}|}, \quad (1)$$

where D is the entire corpus and $c(w_i, d_j)$ denotes the number of times that the word w_i appears in the document d_j. The score of the utterance is the average TF-IDF score of all the words in the utterance, excluding some stop words.

Most content words are nouns, adjectives, or verbs, so we also explored whether better performance is obtained by computing the content score using only the subset of words in the utterance most likely to contain content information. Three filtering alternatives considered selects only: a) nouns; b) nouns and adjectives; or c) nouns, adjectives, and verbs.

As an alternative to, or to be used together with the part-of-speech filtering, we use information related to the structure of the parse tree of an utterance to determine whether the utterance should be considered as a suitable candidate or not. Again, driven by the intuition that most of the topical content is captured by nouns, adjectives, and verbs, we look at the number of noun phrases and verb phrases in the utterance, and discard utterances with too few noun and/or verb phrases.

We used the Berkeley parser [13] to obtain both part-of-speech and higher-level syntactic information. Since no training data was available in any of the genres we were interested in, we used parser models trained on the Wall Street Portion of the Penn Treebank [14]. The performance of this parser was found to be adequate for our needs, but some performance improvements could be obtained using a parser adapted to each domain, in particular the chat data, which is least matched in terms of punctuation, capitalization, and number of incomplete sentences to the Wall Street Journal data.

3.2. Classification Experiments

We performed a number of experiments on each genre to assess the efficiency of our different methods. In all experiments performed on the Wikipedia discussions, we split the data approximately 50-25-25 into training, development, and test subsets respectively. We used the development set to perform parameter tuning and reported results on the evaluation set. For the chat and ICSI meetings datasets, due to data sparsity, we performed parameter tuning and evaluation using cross-validation. We split the ICSI meetings into 5 partitions (with 4 meeting fragments each) and performed 5-fold cross-validation. On the chat data, we performed 4-fold cross-validation, with each chat discussion as a separate partition. All experimental results are reported using accuracy. Similar to [6], we select as the target of each alignment move the speaker whose utterance immediately preceded the turn containing the alignment move in question. The results for all three genres (ICSI meetings, IRC chat, and Wikipedia discussions) are summarized in table 2, where the accuracies of the target of positive and negative alignment moves are reported separately. The best results that outperform the baseline are boldfaced. The results on the meeting data are similar to those reported in [6], though they cannot be directly compared because the test set and tasks are slightly different.

Filter Type	Positive			Negative		
	Mtgs	Chat	Wiki	Mtgs	Chat	Wiki
Baseline	79.7	68.6	71.8	73.4	72.7	68.4
None	88.4	66.9	71.3	72.7	62.3	68.1
POS	89.9	67.8	71.3	80.5	**75.3**	68.1
Parse	**90.5**	**73.3**	71.3	**81.8**	71.4	67.9
Both	85.3	69.5	71.3	80.5	**75.3**	67.9

Table 2. Target detection accuracy, different content selection filters.

We detect significant improvement for both positive and negative alignment moves in the ICSI meetings corpus, and moderate improvement for both types of moves in the chat data. However, we find that there is no improvement over the baseline in Wikipedia discussions. This is not altogether surprising; due to the asynchronous nature of the Wikipedia discussions, sentences tend to be much more complex than in live discussion genres like face-to-face meetings or IRC discussions. Thus, the type of information captured by the TF-IDF content score is less likely to be correlated with the target of an

alignment move. Furthermore, since most sentences are reasonably complex, filtering by part-of-speech or higher-level parse structure is no longer as useful.

In all synchronous discussion cases, we find that some filtering (either POS or parse structure-driven) yields the best result. In the case of positive alignment moves, we find that filtering out sentences with insufficient higher-level parse structure yields the best result. In both ICSI meetings and the chat data, removing sentences whose parse tree does not contain at least one noun phrase (NP) yields the best results. This essentially removes sentence fragments or sentences containing disfluencies from being considered as potential targets. However, sentences in Wikipedia discussions are unlikely to consist of only sentence fragments or disfluencies, since the author of a post has no reason to finish posting until the post is complete.

We find that filtering using parse structure also helps with detection of targets of negative alignment moves in meetings. However, with both chat and Wikipedia discussions, the best results for negative alignment moves is obtained when using POS-based filtering, with the score computed over only nouns, adjectives, and verbs. This leads to the smallest degradation over the baseline in Wikipedia discussions, and a moderate improvement in performance in chat. The fact that combined POS and parse filtering does not help is probably due to too aggressive filtering when both methods are used.

4. LEVERAGING SEMANTIC SIMILARITY

Our second method is motivated by anecdotal observations on the errors made by the content-based classifier, which suggest that a number of errors made by the content-based classifier are made when the source utterance is long and contains a lot of content itself. This suggests that some of the content in the source sentence may also be required to identify the correct target. Rather than selecting the temporally closest target utterance with sufficient topical content, we will select among multiple potential target utterances with sufficient topical content, based on the similarity between the source utterance and the target utterance. This idea is captured in the features of [6] that look at the ratio of overlap of words (or content words) in the source and target utterances, but we take it further, allowing for alternative wording by using a semantic similarity score. Anecdotally, we also see alignment moves in which part of the source statement consists of or contains a paraphrase of a point made by the target speaker, further motivating the use of semantic similarity as a simple method of capturing some paraphrase-related information.

4.1. WordNet-based Semantic Similarity

We employ lexical semantic similarity derived from WordNet [15] to assess the similarity between utterance pairs as proposed in [16]. After stopword removal, for each word in the source utterance, we find the most similar word in the target utterance as its best match. The similarity score of the utterance pair is set to the sum of the semantic similarities of the matched word pairs. Since words may have different contributions on determining the semantic content of an utterance, the semantic similarity of each matched word pair is weighted with the IDF of the word. Hence, the directional similarity from the source utterance S_i to target utterance T_j is defined as

$$\text{Sim}(S_i, T_j)_{S_i} = \frac{\sum_{w_k \in S_i} \text{maxSim}(w_k, T_j) \times \text{idf}_{w_k}}{\sum_{w_k \in S_i} \text{idf}_{w_k}}, \quad (2)$$

where

$$\text{maxSim}(w_k, T_j) = \max_{w_l \in T_j} \text{Sim}(w_k, w_l), \quad (3)$$

is the similarity between w_k and its best match in utterance T_j.

Consequently, the bidirectional utterance similarity is defined as the combination of the two directional utterance similarity,

$$\text{Sim}(S_i, T_j) = \frac{\text{Sim}(S_i, T_j)_{S_i} + \text{Sim}(T_j, S_i)_{T_j}}{2}. \quad (4)$$

Here, we use the algorithm proposed by Lin [17] to extract lexical semantic similarity from the WordNet.

4.2. Classification Experiments

We implemented a separate target classifier using the WordNet similarity method. In this classifier, the WordNet similarity between the source utterance and each potential target utterance is computed, and the discussion participant corresponding to the target utterance with the highest similarity score is selected as the target of the source utterance. The target utterances are chosen from utterances preceding the source made by speakers other than the source speaker. Candidate utterances which differ by more than a factor of two in length (in either direction) from the source are discarded. The number of preceding turns considered is genre-dependent, selected based on corpus statistics. We discovered that more than 95% of the targets of an alignment move appear within 10 turns of the source utterance in the meetings data; within 5 turns in the chat data; and within 25 turns in the development portion of the Wikipedia discussions data.

Applying the WordNet-based similarity classifier (without content filtering) on data from each genre results in much lower performance than the baseline system or the TF-IDF classifiers. There are a couple reasons for this. First, in the synchronous conversation genres, in particular, many of the source utterances of an alignment move contain little to no topical content, in which case the similarity score is less meaningful. In addition, in the online discussions, alignment moves may be made in the context of a poll, with the participants simply expressing their vote in a succinct fashion.

As an oracle experiment, we applied the WordNet classifier to only those alignment moves whose target had been misclassified by the content-based classifier. We select in each case the classifier setup with the best performance using just TF-IDF content. We notice that, in all cases, the WordNet-based classifier improves the accuracy of the system over the content-based classifier alone. Therefore, we explored an automatic combination of the WordNet classifier and the content-based classifier implemented in section 3.

To combine the two classifiers, we propose the following scheme. We evaluate the content of the source utterance containing the alignment move to determine whether the utterance contains enough topical content to be used in the semantic similarity classifier. If it is, then we compare it with each non-empty utterance from the preceding turns within the range appropriate for that genre. We optionally also discard any candidates whose similarity score is below a defined threshold. The threshold is a tunable parameter, defaulting to 0. If the source utterance does not contain sufficient content, or if the highest score does not pass the threshold, we instead use the content-based classifier.

The results for the oracle experiment together with the classifier combination experiment are presented in table 3. As before, the best non-oracle results that outperform the baseline are boldfaced. We observe that the classifier combination results in a slight improvement for the detection of targets of both positive and negative alignment moves in Wikipedia discussions, with the best result for negative alignment moves slightly outperforming the baseline. The performance for positive alignment moves is on par with the baseline. We also detect a slight improvement in the detection of targets of

positive alignment moves in meetings. The performance of the combination in negative alignment moves in meetings and both types in the chat data did not improve compared to the best content-based classifier. In all cases, we find that the best performance is obtained using a fairly high threshold used to discard candidate utterances. For Wikipedia, the optimal threshold is 0.7. For meetings and chat, the optimal threshold is 0.9.

Filter Type	Positive			Negative		
	Mtgs	Chat	Wiki	Mtgs	Chat	Wiki
Baseline	79.7	68.6	71.8	73.4	72.7	68.4
Best Content	90.5	**73.3**	71.3	**81.8**	**75.3**	68.1
+ Combined Sem	**90.9**	72.9	71.8	**81.8**	74.0	**68.7**
+ Oracle Sem	94.8	78.0	73.6	88.9	77.9	75.7

Table 3. Target detection accuracy with content vs. similarity filters.

5. DISCUSSION

We find that target detection performance using the content-based classifier varies significantly with the genre of the data. In particular, the performance on the asynchronous conversations is significantly worse (and worse than the baseline), whereas we obtain a small to significant improvement in both the face-to-face and the online synchronous conversation genres. Looking at the statistics of each corpus may help explain this discrepancy. We find that sentences in Wikipedia are much longer (and thus correspondingly more likely to contain content) than sentences in either synchronous discussion genre. Whereas in meetings and chat data the topic-content classifier acts as a somewhat effective filter of contentless sentences, in Wikipedia discussions the content information does not provide sufficient discriminative information to select the correct target, even when paired with our various syntactic filters.

Despite the positive results when combining the content-based classifier with the semantic similarity classifier in the oracle case, the actual system combination yields mixed results. Error analysis of the results suggests that many errors are caused by using the output of the semantic similarity classifier (which was incorrect) instead of the output of the content-based classifier (which was correct). Thus, the challenge remains to find a better method for deciding when to use the semantic similarity classifier. Setting a high threshold on the minimum score required to use the result helped (significantly, in the case of chat data) but not enough to offset the errors made by the classifier.

Manual error analysis of the classification results shows other distinctions between the different genres in the use of names. The use of names as part of addressing someone is common during multiparty conversations, both online and in-person. However, names are used differently in online and face-to-face discussions. One difference is an artifact of data collection: since the ICSI meetings corpus was anonymized in terms of speaker labeling but not in the transcripts, neither human nor machine labelers could use names in determining targets. For the two online discussion genres, the names are used in determining the target, but the discussants often use abbreviations since the names are often quite long. For example, in our chat corpus, the secretary's name, "secretary2ne" would be shortened to "secretary" by another participant. In Wikipedia discussions, "Dubc0724" could be shortened to "Dubc", and "Commodore Sloat" to "CSloat" or "CS". Such abbreviations make using name information in automatic target detection more difficult, while not presenting much of a problem for people. The use of such information during annotation may explain why our systems perform better on the ICSI meetings corpus than on the chat and Wikipedia discussions corpora.

A major limitation of the work described here is a lack of hand-annotated data, which restricted the study to a small number of features: essentially two scores. With additional training data, it would be interesting to consider more complex methods for combing these features and explore additional structural, paraphrasing, and forward-looking cues.

6. REFERENCES

[1] E. M. Bender et al., "Annotating social acts: Authority claims and alignment moves in wikipedia talk pages," in *Proc. Workshop on Language in Social Media*, 2011, pp. 48–57.

[2] E. A. Schegloff and H. Sacks, "Opening up closings," *Semiotica*, vol. 8, no. 4, pp. 289–327, 1973.

[3] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proc. ACL*, 2008.

[4] S. C. Levinson, *Pragmatics*, Cambridge University Press, Cambridge [Cambridgeshire] ; New York :, 1983.

[5] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: training with unlabeled data," in *Proc. NAACL-HLT*, 2003, pp. 34–36.

[6] M. Galley et al., "Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies," in *Proc. ACL*, 2004, pp. 669–675.

[7] S. Germesin and T. Wilson, "Agreement detection in multiparty conversation," in *ICMI-MLMI '09*, November 2009.

[8] I. McCowan et al., "The AMI meeting corpus," in *Proc. Measuring Behavior*, 2005, p. 4.

[9] W. Wang et al., "Automatic identification of speaker role and agreement/disagreement in broadcast conversation," in *Proc. ICASSP*, 2011, pp. 5556–5559.

[10] W. Wang et al., "Detection of agreement and disagreement in broadcast conversations," in *Proc. ACL*, 2011, pp. 374–378.

[11] A. Janin et al., "The ICSI Meeting Corpus," in *Proc. ICASSP*, 2003, vol. 1, pp. I–364–I–367 vol.1.

[12] R. A. Bates et al., "Meeting acts: a labeling system for group interaction in meetings," in *Proc. INTERSPEECH*, 2005, pp. 1589–1592.

[13] S. Petrov et al., "Learning accurate, compact, and interpretable tree annotation," in *Proc. COLING-ACL*, 2006, pp. 433–440.

[14] M. Marcus et al., "The penn treebank: Annotating predicate argument structure," in *In ARPA Human Language Technology Workshop*, 1994, pp. 114–119.

[15] G. A. Miller et al., "WordNet: An on-line lexical database," *Intl. Journal of Lexicography*, vol. 3, pp. 235–244, 1990.

[16] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proc. ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 2005, pp. 13–18.

[17] D. Lin, "An information-theoretic definition of similarity," in *Proc. International Conference on Machine Learning*, 1998.