Communicative Gestures in Coreference Identification in Multiparty Meetings

Tyler Baldwin
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
baldwi96@msu.edu

Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, USA
jchai@cse.msu.edu

Katrin Kirchhoff
Department of Electrical Engineering
University of Washington
Seattle, WA, USA
katrin@ee.washington.edu

ABSTRACT

During multiparty meetings, participants can use non-verbal modalities such as hand gestures to make reference to the shared environment. Therefore, one hypothesis is that incorporating hand gestures can improve coreference identification, a task that automatically identifies what participants refer to with their linguistic expressions. To evaluate this hypothesis, this paper examines the role of hand gestures in coreference identification, in particular, focusing on two questions: (1) what signals can distinguish communicative gestures that can potentially help coreference identification from non-communicative gestures; and (2) in what ways can communicative gestures help coreference identification. Based on the AMI data, our empirical results have shown that the length of gesture production is highly indicative of whether a gesture is communicative and potentially helpful in language understanding. Our experiments on the automated identification of coreferring expressions indicate that while the incorporation of simple gesture features does not improve overall performance, it does show potential on expressions referring to participants, an important and unique component of the meeting domain. A further analysis suggests that communicative gestures provide both redundant and complementary information, but further domain modeling and world knowledge incorporation is required to take full advantage of information that is complementary.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic Processing*

General Terms

Experimentation

Keywords

Hand gesture, reference resolution, multiparty meetings

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

1. INTRODUCTION

As more and more multiparty meeting data becomes available, techniques to automatically process meetings to identify topics, summarize key points, and discover participant opinions become increasingly important. To enable these techniques, one fundamental aspect in automated meeting processing is reference resolution, a process that identifies what participants refer to with their linguistic referring expressions.

Previous work has investigated aspects of reference resolution in multiparty meetings [15], but mainly based on linguistic features. Compared to written text, multiparty meetings have several unique characteristics which bring new implications to the reference resolution problem. In meetings, participants engage in a multiparty conversation in a shared physical space. This situatedness influences the form of referring expressions, for example, by allowing participants to use non-verbal modalities such as hand gesture to make reference to the shared environment [14], as shown in the following excerpt:

> **A:** "Why is that?"
> **B:** "Because, um, based on what you've go- everybody's saying, right, [*gestures at Speaker D*] you want something simple. You [*gestures at Speaker C*] want basic stuff and [*gestures at Speaker A*] you want something that is easy to use. Speech recognition might not be the simplest thing."

In this example, speaker B utters the pronoun *you* three times, each time referring to a different participant. In order to make their intention clear, the speaker gestures towards the intended participant concurrently with the utterance of *you*. The use of gesture allows the listeners to interpret the speaker's intent unambiguously.

In this work, we examine the prevalence and distribution of gesture in a multiparty meeting setting and its potential role in meeting understanding. In particular, we investigate the role of gesture in coreference identification, an important subtask of reference resolution. This work focuses on two questions: (1) what signals can distinguish communicative gestures that can potentially help coreference identification from non-communicative gestures; and (2) in what ways can communicative gestures help coreference identification.

Our empirical results show that the length of gesture production is highly indicative of whether a gesture is communicative and potentially helpful in language understanding. Our experiments on the automated identification of corefer-

ring expressions indicate that while the incorporation of simple gesture features does not improve overall performance, it does show potential on expressions referring to participants, an important and unique component of the meeting domain. A further analysis suggests that communicative gestures provide both redundant and complementary information, but further domain modeling and world knowledge incorporation is required to take full advantage of information that is complementary.

In the following sections, we present the dataset and give an analysis of the distribution of gestures and referring expressions. We then attempt to identify which gestures can potentially facilitate language understanding and are thus considered communicative. Once we have separated out those gestures that are communicative, we explore their role in improving the performance on the coreference identification task.

2. RELATED WORK

The availability of corpora such as the AMI Meeting Corpus [19], the ICSI meeting Corpus [11], and the VACE Multimodal Meeting Corpus [4] have increased research interest in the multiparty meeting domain. Several meeting projects have included the use and analysis of multimodal information, e.g. video or pen input. This has primarily been used for tracking focus of attention [21], multimodal speaker identification [22] and tracking [2].

Various studies have analyzed multi-party interactions from a discourse-based perspective. These have mostly concentrated on specific isolated tasks, e.g. the annotation and automatic detection of dialogue acts in meetings [6, 1], or topic segmentation [10]. More comprehensive ontologies for multiparty communication have also been proposed [17].

Despite the growing body of interest in multiparty data, the reference resolution problem has yet to be fully explored in multiparty settings. Recent work has examined aspects of the pronoun resolution task, such as the resolution of the pronoun *you* [9] or of particular demonstratives [15]. However, previous work did not incorporate gesture or information from modalities other than speech. The use of hand gesture features in the coreference task has been studied by Eisenstein and Davis [7, 8] using data from lecture-style monologues.

3. AMI DATA

The AMI Meeting Corpus is a large, publicly available corpus of multiparty design meetings. It provides speech transcriptions as well as data from several other sources, such as the projected slides, writings on the whiteboard, and notes taken by participants. Annotations are provided for several other modalities, such as focus of attention and hand and head gestures. Figure 1 shows a snapshot of the video of a meeting provided in the corpus.

In our current investigation, we selected 6 AMI meeting segments for use as our data set. Three of these segments (AMI meeting IDs IS1008a, IS1008b, and TS3005a) were analyzed in our initial analysis step to better understand the distribution of the data. These three segments were kept as training data for our coreference evaluation, while the three unexamined segments (AMI meeting IDs IS1008c, IS1008d, ES2008a) were used as testing. Each meeting segment was 20-40 minutes in length and contained one full meeting between four participants. The style of each meeting was similar; all meeting segments contained design meetings in which participants designed a theoretical television remote control. Each meeting participant had a predefined role as either a project manager, marketing expert, industrial designer, or user interface designer. All six segments included manually annotated gesture information, which we utilized in our investigation.

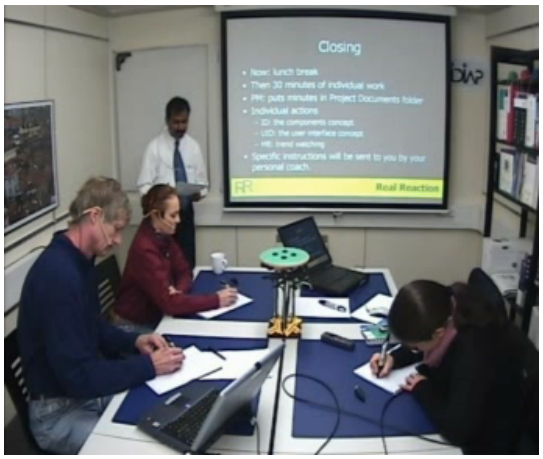

Figure 1: The AMI Meeting Corpus

The AMI corpus divides gestures broadly into 2 types: communicative and non-communicative gestures[1]. Communicative gestures are those that potentially carry some useful information in relation to the discourse, such as pointing gestures. Conversely, non-communicative gestures are those not thought to add relevant information to the discourse, such as beats or fiddling idly with objects.

Communicative gestures were further subdivided into three categories: pointing gestures, interacting gestures, and other gestures. Pointing gestures are a common communicative gesture in which participants highlight a certain object by gesturing towards it. Interaction gestures include such things as picking up or reaching for an object. All other gestures that are intended to communicate information are put in the other communicative gesture category. This includes gestures such as mimicking the shape of an object in the air, or emblematic gestures such as the "ok" sign.

We examined a small sample of 242 gestures and 1790 referring expressions from 3 AMI meetings segments for a preliminary analysis. The distribution of gestures is shown in Figure 2. Non-communicative gestures accounted for about half of all gestures made. Of those that were communicative, the majority were pointing gestures, which were split evenly between pointing at other participants and at inanimate objects. Interacting and other communicative gestures accounted for a relatively small portion of gestures produced.

This distribution gives us several insights relevant to understanding the role of gesture in multiparty meetings. First, the high number of non-communicative gestures suggests that some analysis to distinguish communicative gestures from non-communicative gestures may be necessary. This would include the analysis of both the form of the gesture

[1]The AMI annotations also contain a third gesture type, gestures that were off camera and thus have unknown intent. These were not considered in this study.

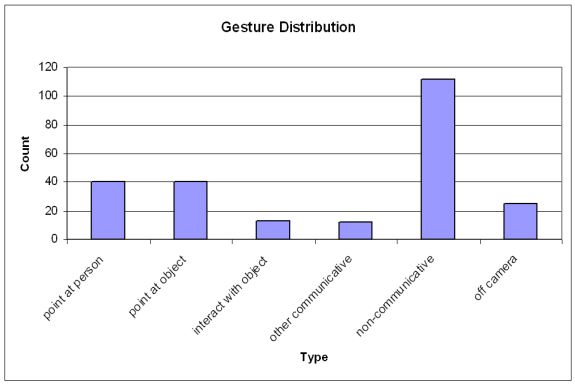

Figure 2: Distribution of Gestures

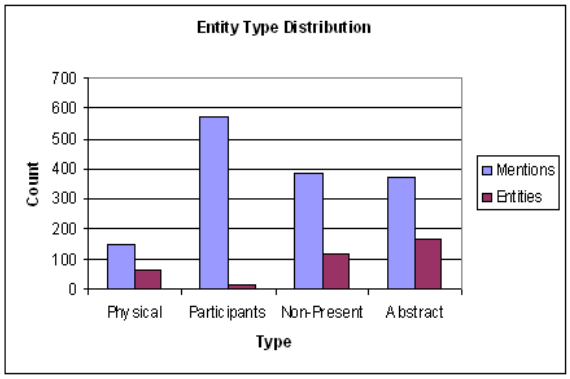

Figure 3: Distribution of Entity Types

produced and the surrounding linguistic context in order to determine the likelihood that the gesture carries information relevant to the discourse. We investigate this by looking at the classification problem of determining whether a gesture is communicative.

Secondly, Figure 2 suggests that the primary focus for communicative gesture analysis should be on pointing gestures, as they are more prevalent. We explore the role of pointing communicative gestures on coreference identification in Section 5.

In order to analyze the distribution of referring expressions, we examined how often real entities were mentioned in the text, as well as the distribution of entities themselves. Following the conventions set by the ACE information extraction evaluation[2], we call the real world objects that are referred to *entities* and the linguistic expressions from the text that refer to them *mentions*. We separated entities and their corresponding mentions into four distinct types: participants, physical, non-present, and abstract. The distribution of unique entities and mentions in our data set is shown in Figure 3.

Participants include not only each individual person present at the meeting, but also every combination of two or more individuals. The participant type is unique to dialogue, and has thus received less study than types that are present on written text. In multiparty meetings, understanding which participants are being referred to is a key aspect necessary to the understanding of the entire meeting. Because participants are present in the local environment and can be referred to extra-linguistically, gesture information has the potential to improve resolution of participant references.

Physical entities are those entities that have a physical presence in the local environment. This includes objects such as laptops and pens, but also less concrete physical objects such as pictures on projected slides or drawings on the whiteboard. Although participants are also physically present, they represent a special case of physical entity and were thus examined separately. These entities are important because they may reflect the topic of discussion. As with participants, physical entities can be referenced and kept salient by extra-linguistic cues, such as a pointing gesture.

Non-present entities are those entities that have or could have a physical presence, but do not appear in the local environment. This includes all entities that would be considered physical if they were present in the local environment, including non-present persons. Hypothetical entities that would have a physical presence are also considered to be non-present entities. For instance, in the design meeting setting of the AMI corpus, it is understood by the meeting participants that the object being designed has physical properties associated with it, even though it only exists hypothetically. Non-present entities differ from physical entities in that they cannot be referred to by most extra-linguistic means, such as pointing gestures. However, certain gestures, such as mimicking the shape of an object in the air, may be used to reference non-present entities.

Abstract entities are complex entities without a physical presence. Examples include "the cost to produce a product", "the goal of the meeting", or "the market in which a product is to be sold". Abstract entities are often hard for reference resolution systems to deal with and are often missed completely, in part because they are commonly expressed in forms other than noun phrases [3]. As with non-present entities, abstract entities have no physical presence in the local environment and thus could not be the target of pointing gestures.

Given the nature of pointing gestures, our hypothesis is that communicative gestures will be more likely to accompany mentions of the participant and physical types. Analysis of alignment between gestures and referring expressions seems to support this hypothesis. 38% of mentions to physical entities and 24% of mentions to participants occurred within 5 seconds of a communicative gesture. For non-present and abstract mentions, the percentages are only 9% and 17%, respectively. However, as illustrated in Figure 3, the physical and participant types make up less than half of all mentions and entities, limiting the effectiveness of gesture data. The relatively low number of mentions to physical entities may be due in part to the nature of the data. Because the data we examined was from design meetings, much of the focus of the discussion was on a hypothetical object, the object under design. Further study of other datasets is needed to understand if the distribution we observed is indicative of general multiparty meeting settings.

The distributions in Tables 2 and 3 suggest that the participant type requires particular attention. Not only is it more prevalent in our data than any other type, but half of the pointing gestures in our data are to people. Because of its central importance, a focused evaluation of the participant type is included in Section 5.

[2]http://www.itl.nist.gov/iad/mig/tests/ace/

4. COMMUNICATIVE GESTURES

Although many hand gestures may be produced by participants over the course of a meeting, not all are intended to aid in comprehension. Some, such as beats, are simply meant to add emphasis to the speaker's words. If we wish to utilize hand gestures to help coreference identification, the system must first be able to separate these instances from true communicative gestures. Therefore, our first task is to identify what features can be used to distinguish communicative gestures from non-communicative gestures.

4.1 Methodology

We formulate the problem as a binary classification problem. A given gesture is a member of the positive class if it is communicative, otherwise it is a member of the negative class. We drew features from two main sources: linguistic and gesture information.

Several linguistic features were considered. A few features were based on co-occurring referring expressions. Gestures were aligned with co-occurring expressions automatically, using insight from previous studies. Previous work [18] has suggested that the onset of gestures will come close to, but before, the onset of the expression. As such, we chose the closest noun phrase after the the onset of the gesture as our aligned expression. From this we extracted the type of the referring expression (either definite, indefinite, pronoun, or demonstrative; each stored as a binary feature) and the grammatical number (singular or plural).

One other linguistic feature unrelated to the aligned expression was used: the presence of disfluency in the speech. Disfluencies were identified by looking for filler phrases such as "um" and "er". A disfluency was said to be present if one of these phrases was uttered while the gesture was being produced.

Only one gesture-based feature was used: the amount of time spent producing the gesture. Gesture features were purposely kept shallow in order to ensure that the methods applied here would be applicable to automatically detected gestures. Certainly other visual features from gestures can be potentially helpful, but here we only focus on features that are readily available from the AMI corpus. Our assumption is that communicative gestures will be generally short so as to correspond most directly to co-occurring expressions, while non-communicative gestures will likely be longer, making gesture length an important feature.

4.2 Results

A total of six AMI meeting segments was used for evaluation. Three meeting segments (described in Section 3) were used for training and development, while the other three segments were used for testing. A logistic regression classifier provided by the Weka toolkit[3] was used for classification.

Configurations based on feature sets consisting of only linguistic, only gestural, and all features were examined. A majority class baseline was used for comparison. The results are shown in Table 1.

While the incorporation of linguistic features showed a large improvement over the baseline, the gesture length feature proved to be the most significant, showing even larger gains when it was the only feature considered. Combining the two feature sources did not result in a statistically significant increase in performance over the the gesture length feature alone. All other differences in Table 1 were significant ($p < 0.05$).

	Accuracy
Baseline	50.4
Linguistic Features Only	68.4
Gesture Length Only	86.4
All Features	87.6

Table 1: Results of Classification of Communicative and Non-communicative Gestures

These results show that the length of production is a primary difference between gestures that carry communicative intent and those that do not. Communicative gestures often correspond to a referring expression or keyword that is uttered concurrently, necessitating a concise gesture so as to aid listeners with the alignment of gesture and phrase. Furthermore, emblematic communicative gestures, such as the "ok" sign, convey meaning relevant to the discourse, and thus benefit from being timely and concise.

Conversely, the nature of non-communicative gestures makes them more likely to have longer productions. While some beats may be used to emphasize certain words or phrases, they often accompany whole utterances or turns taken by a speaker. Similarly, other non-communicative hand gestures such as fiddling with objects or writing notes are also likely to last longer than quick communicative gestures.

5. COREFERENCE IDENTIFICATION

Given a communicative gesture, our second task is to investigate whether the gesture is able to help interpret co-occurring linguistic referring expressions in the discourse. To facilitate this investigation, we manually annotated a small dataset of 3261 noun phrases from six AMI meeting segment transcripts[4]. Referring expressions were manually annotated with referents by two human annotators. Non-referential noun phrases were not considered. Inter-annotator agreement was acceptably high, $\kappa = 0.80$.

The coreference identification task is to identify whether two referring expressions corefer to each other. This is modeled as a pairwise classification task which makes a binary decision about whether each pair of NPs in the text are coreferential. This is a very important first step for coreference resolution, which identifies a chain of expressions that refer to the same entity [20].

In our experiments, we specifically examine pointing gestures to the exclusion of all other gestures. This choice stems largely from annotation limitations; we restrict ourselves to gestures that have an intended target, as this gives insight into its intent. Although interaction gestures may also have targets, AMI annotations did not specify what they were. Other communicative gestures, such as using the hands to mimic the shape of an object in the air, may not have a physical target at all.

5.1 Features

We applied a diverse set of features drawn from text, dialogue, and gesture information.

[3]http://www.cs.waikato.ac.nz/ml/weka/

[4]The annotated data is available at http://links.cse.msu.edu:8000/AMIrefdata

Text Features	
Gender	MATCH if the two mentions agree in gender, NOMATCH if they don't, UNKNOWN if not determinable
Number	TRUE if mentions agree in number, FALSE otherwise
Phrase Match	TRUE if mentions have the same surface text
Substring	TRUE if one mention's surface text is a substring of the other's
Apposition	TRUE if the two mentions exist in an apposition relationship
Mention Types	Features representing the type of each mention; either definite, indefinite, demonstrative, or pronoun
NP-Distance	The distance between the two mentions in NPs
Time-Distance	The distance between the two mentions in seconds
Dialogue Features	
Speaker	Features representing the speaker of each mention (1 per mention)
Is-*I*	TRUE if the mention is a form of the pronoun "I"
Is-*you*	TRUE if the mention is a form of the pronoun "you"
Gesture Features	
Gesture	TRUE if gestures co-occurring with the mentions point to same entity, FALSE if they don't, N/A if one or both mentions do not have accompanying gesture

Table 2: Coreference Feature Set.

Text Features. Since the coreference resolution task has been performed extensively on text, many different feature sets have been experimented with. In this work we choose six common textual features that are applicable to multiparty meetings, as listed in Table 2. Our text features are straightforward applications of the feature as they have been applied on text. One exception is the distance feature. While this is often measured as distance in sentences, the free-flowing nature of situated multiparty speech made it difficult to measure this accurately. Instead, we chose to measure relative distance in two ways: (1) we count the number of noun phrases between the two mentions and (2) the amount of time that elapses between the two mentions.

It is worth pointing out that several other features, such as alias features [20], have been explored in coreference resolution on text. We chose not to use these features because they are not applicable to spoken dialogue on multiparty meeting data.

Dialogue Features. Jing et. al. [12] note the need for speaker role identification when performing coreference in two-party dialogues. Since the pronoun *I* generally refers to the speaker, some knowledge of speaker identity is needed. Additional pronouns, such as *you*, also benefit from knowledge of the speaker. Although *you* can often be trivially resolved in two-party dialogue, it carries more ambiguity in multiparty contexts.

We incorporated several conversational speech features from Jing et. al., with slight modifications. Two *speaker* features were used, representing the speaker of each mention. Speakers were represented by their role in the meeting (i.e., "user interface designer"). Each speaker feature had four possible values, corresponding to the four participants of the meeting.

The other two features, the *is-I* and *is-You* features, indicate whether each of the mentions is in the form "I" or "you", respectively, since in dialogue the use of these expressions directly relate to speakers and listeners.

Gesture Features. We included a single feature in order to capture information from hand gesture. The gesture feature indicates whether the targets of the two gestures that accompany the pair of mentions match. Modeling gesture this way allows us to incorporate gesture without the additional domain knowledge needed to map individual mentions to physical entities (which are the targets of the gesture).

5.2 Empirical Results

Human annotators annotated all noun phrases from 6 AMI meeting segments with corresponding referents. Three meeting segments (see Section 3) were used for training and development, and the other three were used for testing. Those noun phrases that occurred within a small time window around the time of the gesture's production were considered co-occurring with that gesture. Previous psycholinguistic studies [13] have shown that the onset of the gesture generally precedes the onset of the expression. Analysis on our development data seems consistent with this finding; on average, the onset of communicative gestures occurred about one third of a second before the onset of the referring expression.

We applied a decision tree classifier for the binary classification task. For the classification task, every referring expression in the text that co-occurred with a gesture was compared to every other referring expression that co-occurred with a gesture and a binary decision was made as to whether the two referring expressions were coreferential. We restricted our data set to those instances that co-occur with gesture in order to focus on instances in which gestures have the potential to help. Although more sophisticated models have been shown to produce better results [5], we decided to use a simpler and more transparent model in this investigation, to serve our goal of understanding the role of gesture in coreference identification in multiparty meetings.

Since most referring expressions in the text are only coreferential with a few others, performing pairwise classification results in unbalanced data. Specifically, there are far more negative instances than positive. In our data, 95% of the pairs examined were not coreferential, which is consistent with previous work [16]. In order to not bias the classifier towards the negative class, only a small subset of negative examples were used in training to produce balanced data; this is a standard step taken in coreference identification [20, 16]. Because the number of pairs that do not corefer is so high, a trivial majority class classifier will produce high overall accuracy, while being completely useless for resolving coreference. As such, we use F-measure on the positive class (predicting a pair as coreferential) as our performance metric.

To examine the utility of gesture, two feature configurations were run: one that included the gesture feature and

	All Entity Types		
	Precision	Recall	F-measure
-Gesture	0.527	0.240	0.330
+Gesture	0.456	0.256	0.328
	Participants Only		
-Gesture	0.146	0.105	0.122
+Gesture	0.313	0.175	0.225

Table 3: Precision, Recall, and F-measure Results for the Positive Class for Coreference Identification

one that did not. The results are shown in Table 3. Across all entity types, the incorporation of the gesture feature did not provide improvement over the text and dialogue features alone. Inspection of the learned model shows that gesture was not a significant feature, appearing only in the deepest levels of the decision tree.

As we observed in Section 3, the participant type is the most prevalent entity type in our data, and gestures towards participants make up half of all communicative pointing gestures. To examine the role of gesture on these important instances, we further evaluate the subset of data that consisted solely of participant instances. Results on the participant type are shown in Table 3. Overall, the evaluation with our set of linguistic and dialogue features (and even the gesture feature) results in pretty poor performance. These results are based on a limited number of instances (57 positive instance) available in our data. Although more data is necessary to make it conclusive, the current results suggest that the incorporation of the simple gesture feature can potentially improve processing expressions referring to participants.

The results in Table 3 suggest the ability of gesture to aid in coreference identification is at least partially dependent on the type of referring expression that gesture co-occurs with. Because gestures target objects in the environment, they are most helpful in cases in which the referring expression also targets a physically present object, such as a meeting participant.

It is worth noting that, because the gesture data used was gold-standard manual annotations, these results represent an upper bound on results in which gestures were automatically determined. Our poor results when all entity types are considered seem to suggest that gesture information is largely redundant with other linguistic information. However, another explanation is possible: that the simple way in which gesture was incorporated was insufficient to extract any meaningful non-redundant information. This may be due to a lack of domain knowledge or how the gesture information was fused with dialogue information. To help further understand these issues, we conducted additional analysis.

5.3 Analysis

In order to better understand the potential role of gesture in coreference identification, we examined a subset of data in which a mention was accompanied by a communicative gesture. In the coreference identification task, our classifier provides us with confidence values for each coreferential pair. From these, we can rank the instances for any given mention to obtain an N-best list of coreferential mentions. Using the ground truth mapping from mentions to entities from our annotations, we are able to produce an N-best list of potential entities for each mention. We then examine whether gesture information can help identify the correct referent for the referring expression from the N-best list, via re-ranking. Instead of being represented as part of the largely text-based coreference task, gesture information is instead incorporated at a later step to attempt to correct classification mistakes.

We examined N-Best lists for 172 instances where gestures were aligned with referring expressions. Of these, 70 were given the correct referent by text and dialogue information alone, implying that the gesture information was redundant. Of the remaining 102 cases, 59 provided complementary information, while the other 43 were unhelpful in resolving the referent. However, only 21 of those with complementary information referred to the exact target of the gesture. While the remaining 38 instances held complementary data, they required additional domain knowledge to be useful.

The complementary information provided by gesture to the co-occurring referring expressions can be split into two broad categories: direct matches and partial matches. Direct match instances represent those cases in which the gestures are most clearly helpful; those in which the target of the gesture corresponds to the same physical object being referred to.

However, there are several instances where the gesture does not directly match the referent of the expression, but still provides information relevant to the expression. An analysis of these instances in our data found that they fall into 3 basic types: possessives, bridging, and target ambiguity.

Possessives are those instances in which a referring expression of the possessive form is used and the gesture refers to the possessor. An example of this would be the expression *your laptop* co-occurring with a gesture towards the laptop's owner. Although the gesture does not directly identify the object that is referred to, it gives some additional information (the object's owner) and may be able to help identify the object if domain knowledge is present. In our data, possessives are the least common type of partial match. This type of complementary information appeared to be quite rare; of the 38 referring expressions helped by partial match instances, only one was a possessive.

Bridging instances are those in which the gesture is made towards an object that is a representative of the referring expression. Often, the objects referred to by the gesture and expression are in a part-whole relationship. An example of this would be gesturing towards a member of the marketing team while referring to the marketing team as a whole. As with possessives, bridging gestures have the potential to provide useful information, but may require domain knowledge to do so. Fifteen examples of bridging were observed in our data.

The third type of potentially helpful gesture are those that suffer from *target ambiguity*. In these cases, the target of the gesture is too general to directly identify the object that is referred to. An example of this is reference to an object drawn on the whiteboard co-occurring with a gesture whose target is the whiteboard itself. In this case, the target of the gesture is not specific enough to identify the exact referent. Although these instances may be seen as a limitation to the annotation schema, they represent a real problem in gesture analysis; it may be implausible that the target of the gesture can be identified exactly due to limitations of gesture recognition and the dynamic nature of the meeting

setting (from changing of slides and writing and erasing of whiteboards). Target ambiguity was the most prevalent type of partial information encountered, with 22 instances in our data.

The relatively small number of instances in which gesture data is able to provide help beyond the discourse level suggests that a large portion of the information provided by gesture may be redundant. However, as greater levels of domain knowledge become available, the role of gesture information expands. Although further analysis is needed to draw a confident conclusion, our analysis suggests that incorporating gesture information into coreference identification tasks may require domain modeling and world knowledge. Additionally, because our data set was small, it is likely that the three types of partial matches identified here represent only a subset of ways in which gesture could provide partial information.

6. DISCUSSION

There are two issues related to our current investigation that could impact the findings from this work. The first issue is the amount of domain modeling performed. This had the most direct impact on our method of incorporating gesture data into our experiments.

We chose to incorporate gesture as a single feature, representing whether the gestures that accompanied a pair of referring expressions had the same target. This method of incorporation was very simple, and may not have accounted for all of the relevant information the gesture could offer. For instance, the gesture feature was only relevant if both referring expressions in the text had an accompanying gesture, significantly reducing the number of cases in which gesture could potentially help. Most significantly, the exact target of the gesture was not considered.

Our choice to use this type of incorporation stems from a conscious effort to examine the role of gesture without modeling the domain. In order to associate gesture targets with mentions from the text, we must be able to link these mentions to the physical objects that an entity represents. In order to do so, we must have some model of each physical object in the environment. This requires us to do extensive modeling of a dynamic domain. Since participants are continually changing projected slides and drawing and erasing whiteboard images, new objects that can be gestured towards are entering and leaving the domain regularly. Because this would require keeping a constantly updating computationally expensive model of the domain, we felt that such incorporation was impractical.

However, much of the analysis presented herein seems to suggest that extensive modeling may be necessary for gesture incorporation to be fruitful. Our analysis of partial match cases in Section 5.3 suggests that meeting participants use pointing gestures for more than just pointing to the exact referent of their expression, and they count on listeners to use their world knowledge to handle these more complex gestures. For instance, bridging references count on listeners' ability to resolve the part-whole relationship using general world and semantic knowledge, while possessive instances rely on the listeners' knowledge of the current domain and of the other participants. These instances present a challenge to the successful incorporation of gesture, but have the potential to provide performance gains if overcome.

The second issue to be addressed is the data. We observed two key factors of gesture usage in our data that limited its utility. One of these issues is that much of the gesture information is unnecessary, as the particular cases that gesture could potentially help can already be resolved by text and dialogue features alone. While this redundancy suggests that gesture information may not be *required* to correctly resolve these references, it still has the potential to aid in reinforcing the correct referent.

The other usage issue encountered in our corpus was poor coverage. In the data examined, the number of communicative pointing gestures produced is only around 6% of the number of referring expressions. Since each gesture is aligned with at most one expression, the low coverage of gesture equates to a limit on the performance gains possible from gesture incorporation.

It is not clear from our study if issues with redundancy and coverage are common factors of gesture usage in multiparty meetings as a whole, or are unique to the AMI corpus or this type of design meeting. It may be that the design meeting setting of the AMI corpus is less friendly towards the usage of gesture than other meeting settings, and future work is needed to explore this issue. However, no other comparable dataset is currently available.

7. CONCLUSIONS

This work presents an empirical investigation of hand gesture behavior relative to referring expression usage in multiparty meeting data. It first explores the task of understanding which aspects of a gesture's production and the surrounding linguistic context suggest that the gesture is communicative and can be potentially useful for language processing. It was shown that while examining the linguistic context does lead to above baseline performance, the most indicative feature for understanding whether the gesture is meant to be communicative was the length of the gesture presentation. In particular, communicative gestures tend to have significantly shorter presentations than non-communicative gestures (as defined by the AMI corpus).

Understanding that a gesture is defined as communicative does not guarantee that this gesture will be helpful in coreference identification. We explored this fact by investigating the role of gesture in coreference identification. Our initial attempt at incorporating gesture as a simple feature showed no improvement over a feature set that did not include gesture information when all entity types were considered. However, gesture incorporation did appear to aid in identification of coreference on entities of the participant type, an entity type that our analysis showed was both prevalent in our data and often gestured towards. To understand whether the lack of overall improvement arose from poor incorporation or redundancy, we performed a further analysis of those mentions that had an accompanying gesture.

Our further analysis suggested that while much of the information provided by the gesture was redundant, in several instances gesture data was seen to provide correct answers where text and dialogue data could not. Many of these complementary instances involved gestures interacting with referring expressions in more complex ways than a simple direct match. We identified three common potential uses of hand gesture data: possessives, bridging, and target ambiguity. In order to utilize gesture in these instances,

more extensive domain modeling is necessary. Our future work will investigate how to incorporate domain knowledge within these three directions to help resolve linguistic referring expressions.

8. ACKNOWLEDGMENTS

This work was supported by IIS-0840538 from the National Science Foundation. The authors would like to thank anonymous reviewers for their valuable comments and suggestions.

9. REFERENCES

[1] J. Ang, Y. Liu, and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1061–1064, 2005.

[2] K. Bernardin, H. K. Ekenel, and R. Stiefelhagen. Multimodal identity tracking in a smart room. *Personal Ubiquitous Comput.*, 13(1):25–31, 2009.

[3] D. K. Byron. Resolving pronominal reference to abstract entities. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[4] L. Chen, R. Travis, F. Parrill, X. Han, J. Tu, Z. Huang, M. Harper, F. Quek, D. McNeill, R. Tuttle, and T. Huang. VACE multimodal meeting corpus. In *Proceedings of the second Joint Workshop on Multmodal Interaction and Related Machine Learning Algorithms*, July 2005.

[5] A. Culotta, M. Wick, and A. McCallum. First-order probabilistic models for coreference resolution. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 81–88, Rochester, New York, April 2007. Association for Computational Linguistics.

[6] A. Dielmann and S. Renals. Multistream recognition of dialogue acts in meetings. In *Machine Learning for Multimodal Interaction*, pages 178–189. 2006.

[7] J. Eisenstein and R. Davis. Gesture improves coreference resolution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 37–40, New York City, USA, June 2006. Association for Computational Linguistics.

[8] J. Eisenstein and R. Davis. Conditional modality fusion for coreference resolution. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 352–359, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[9] S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky. Resolving you in multi-party dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[10] P. Y. Hsueh, J. Moore, and S. Renals. Automatic segmentation of multiparty dialogue. In *Proc. EACL*, 2006.

[11] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede. The ICSI meeting project: Resources and research. In *ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada*, 2004.

[12] H. Jing, N. Kambhatla, and S. Roukos. Extracting social networks and biographical facts from conversational speech transcripts. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 1040–1047, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[13] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. 1980.

[14] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[15] C. Müller. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 816–823, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[16] V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[17] J. Niekrasz and M. Purver. A multimodal discourse ontology for meeting understanding. In *Proceedings of MLMIŠ05. LNCS*, pages 162–173. Springer Verlag, 2005.

[18] S. Oviatt, A. DeAngeli, and K. Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422, New York, NY, USA, 1997. ACM.

[19] A. Popescu-Belis and P. Estrella. Generating usable formats for metadata and annotations in a large meeting corpus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 93–96, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[20] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Comput. Linguist.*, 27(4):521–544, 2001.

[21] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13:928–938, 2002.

[22] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel. Multimodal people ID for a multimedia meeting browser. In *in ACM Multimedia*, pages 159–168, 1999.