Journal of Artificial Intelligence Research 31 (2008) 353-398 Submitted 09/07; published 02/08

Gesture Salience as a Hidden Variable for Coreference Resolution and Keyframe Extraction

Jacob Eisenstein JACOBE@CSAIL.MIT.EDU
Regina Barzilay REGINA@CSAIL.MIT.EDU
Randall Davis DAVIS@CSAIL.MIT.EDU
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139 USA

Abstract

Gesture is a non-verbal modality that can contribute crucial information to the understanding of natural language. But not all gestures are informative, and non-communicative hand motions may confuse natural language processing (NLP) and impede learning. People have little difficulty ignoring irrelevant hand movements and focusing on meaningful gestures, suggesting that an automatic system could also be trained to perform this task. However, the informativeness of a gesture is context-dependent and labeling enough data to cover all cases would be expensive. We present *conditional modality fusion*, a conditional hidden-variable model that learns to predict which gestures are salient for coreference resolution, the task of determining whether two noun phrases refer to the same semantic entity. Moreover, our approach uses only coreference annotations, and not annotations of gesture salience itself. We show that gesture features improve performance on coreference resolution, and that by attending only to gestures that are salient, our method achieves further significant gains. In addition, we show that the model of gesture salience learned in the context of coreference accords with human intuition, by demonstrating that gestures judged to be salient by our model can be used successfully to create multimedia keyframe summaries of video. These summaries are similar to those created by human raters, and significantly outperform summaries produced by baselines from the literature.[1]

1. Introduction

Gesture is a nearly ubiquitous feature of face-to-face natural language communication and may be used to supplement speech with additional information or to reinforce the meaning already conveyed (McNeill, 1992). In either case, gesture can increase the robustness of natural language processing (NLP) systems to the inevitable disfluency of spontaneous

1. This article is an extension and unification of two conference publications (Eisenstein, Barzilay, & Davis, 2007; Eisenstein & Davis, 2007). It extends prior published work with several new, unpublished results: stability analysis with respect to initialization weights (Section 4.3); analysis of verbal features in terms of centering theory (Section 5.1); interrater agreement analysis on coreference annotations (Section 6); evaluation of coreference using a global metric (Section 6.2); expanded empirical evaluation on the coreference task with additional fusion models (Section 6.2); analysis of different types of gesture features for multimodal coreference resolution (Section 6.4.1); a study of the interaction between gestural and verbal features (Section 6.4.2); and interrater agreement on keyframe extraction (Section 7.3). Source code and data are available at `http://mug.csail.mit.edu/publications/2008/Eisenstein_JAIR/`

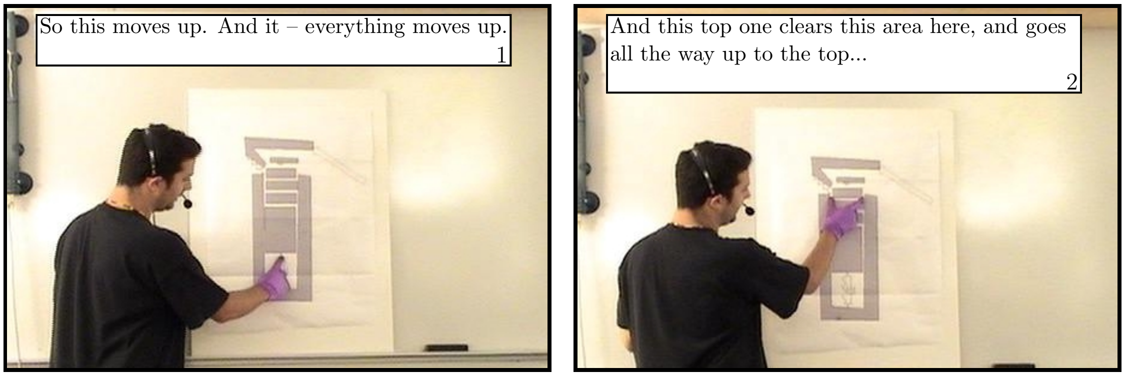


Figure 1: An excerpt of an explanatory narrative in which gesture helps to disambiguate meaning.

speech. For example, consider the following excerpt from a presentation in which the speaker describes a mechanical device:

> "So this moves up, and it – everything moves up. And this top one clears this area here, and goes all the way up to the top."

The references in this passage are difficult to disambiguate, but meaning becomes clearer when set in the context of the accompanying hand gestures (Figure 1).

Despite the apparent benefits offered by gestural cues, obtaining concrete gains for natural language understanding is difficult. A key problem is how to combine gesture and linguistic features. Existing systems typically address this issue by directly concatenating low-level visual information (e.g., hand position and speed) with traditional textual features (Eisenstein & Davis, 2006), or by combining the posteriors from separately-trained models (Chen, Harper, & Huang, 2006). An appealing alternative is to consider the inherent linguistic quality of gesture, as distinguished from other hand movements that may be meaningful for the desired language understanding task (Goodwin & Goodwin, 1986).[2] We show that better results can be obtained by focusing on the hand movements that are likely to correspond to relevant gestures.

To move beyond a low-level representation of gesture, one could attempt to develop a general-purpose taxonomy of gestures based on their relation to language. Such taxonomies have proved useful for psychological and linguistic research on gesture, but their application to corpus-based statistical language processing is not immediately practical. Gesture is a multifaceted phenomenon, and the key features for understanding a gesture's meaning may be highly context-dependent (Lascarides & Stone, 2006). For example, the flexion of a

2. All hand motions – or even the absence of hand motion – may be "meaningful" in some sense. However, for a specific language processing problem, only some gestures will be directly relevant. In the remainder of this article, the terms "meaningful" and "meaningless" should always be assumed to be framed within the context of a specific language processing task. Hand motions that are meaningless for coreference resolution may indeed be quite useful for another problem, such as sentiment classification.

single finger might be a crucial component of one gesture and an irrelevant detail in another context. Is it possible to create a formal annotation scheme expressive enough to capture all such details, yet compact enough to be tractable? This is a topic of ongoing research. But even if possible, such annotation would be very time-consuming, particularly on the scale necessary for corpus-based NLP.

In this paper we propose a middle path: a model that learns to attend to salient gestures without explicit gesture annotation. Instead of a top-down approach that attempts to analyze gestures according to a universal taxonomy, we work bottom-up from a specific language understanding problem: coreference resolution. When a speaker produces similar, meaningful deictic[3] gestures during two noun phrases, it is a good indication that the noun phrases are coreferent (Eisenstein & Davis, 2006). We automatically identify the gestures that are relevant for coreference resolution, from among all hand motions that co-occur with noun phrases. This approach is shown to enhance the contribution of low-level gesture features towards coreference.

More concretely, we employ a conditional model with a hidden variable that governs whether gesture features are included in the determination of coreference for each pair of noun phrases. With this model, it is possible to learn gesture salience jointly with coreference. As a baseline, we demonstrate that even a low-level concatenative approach to gesture-speech fusion[4] yields a small but statistically significant improvement for coreference resolution, compared to textual features alone. More importantly, we show that the contribution of the gesture features increases substantially when gesture and speech are combined using our structured model.

If the model of gesture salience that we learn were relevant only for coreference resolution, it would be useful only from an engineering perspective. An interesting question is whether these estimates of gesture salience are related to how humans perceive multimodal communication. To answer this, we examine whether our model of gesture salience is relevant to other language processing tasks. We demonstrate that the model learned for coreference resolution can be applied to the selection of keyframes for generating visual summaries of instructional presentations. Without any explicit training on the keyframe extraction task, our approach selects keyframes that cohere meaningfully with those chosen by human annotators.

The main contributions of this paper can be summarized as follows.

New applications of gesture: We demonstrate the benefits of incorporating gesture in two tasks: coreference resolution and video keyframe extraction. On the coreference task, we substantially improve on our previous work that showed that gesture similarity can help predict coreference resolution (Eisenstein & Davis, 2006); the application of linguistic analysis of gesture to video keyframe extraction is novel. In previous research, gesture information has been shown to boost performance of sentence segmentation, a local syntactic phenomenon. Our work demonstrates gesture's usefulness to non-local, discourse-level tasks. To this end, we introduce a novel set of features that tightly combine linguistic and gestural information.

3. Deictic gestures are those that communicate meaning through spatial location (McNeill, 1992).
4. We use the term "speech" to indicate that we are dealing with spoken language, but note that hand transcriptions rather than automatic speech recognition (ASR) are used throughout our experiments. The applicability of our techniques in the context of noisy ASR transcripts is a topic for future work.

Gesture salience in language: We develop the idea that gesture information should be considered for language processing only when the gesture is salient. As in prior research (e.g., Chen, Liu, Harper, & Shriberg, 2004), our model uses low-level features extracted directly from a vision-based hand tracker, avoiding the need for manual annotation of gesture features. However, the relevance of these low-level features depends on the linguistic context. By modeling this relationship through gesture salience, we obtain significant performance gains. In addition, we present a set of features designed to capture the salience of gesture to the associated speech.

Hidden-variable modeling of gesture salience: We develop a framework in which gesture salience can be modeled jointly with coreference resolution. We show that gesture salience can be expressed as a hidden variable and learned without explicit labels, leveraging only coreference annotations. This novel framework is realized within a conditional model, enabling the use of arbitrary and possibly non-independent features. Further experiments demonstrate that the estimates of gesture salience obtained by this model can be applied to extract keyframes containing salient deictic gestures.

In Section 2, we consider prior work relating to existing models and taxonomies of gesture from the psychology literature, as well as previous efforts to incorporate gesture into natural language understanding. In Section 3, we describe the dataset on which we conduct our experiments. In Section 4, we present our model, *conditional modality fusion*. We show how gesture salience can be treated as a hidden variable and learned without explicit annotations. Section 5 includes a description of the textual and gestural features that we use for our experiments. In Section 6 we present experimental results showing that our model improves performance on coreference resolution. In Section 7, we show that our estimates of gesture salience are more general, and can be applied to select useful keyframes from video. Finally, in Section 8 we discuss implications of this research, and conclude.

2. Related Work

In this section, we describe four general areas of related work that provide background and context for our efforts. Section 2.1 discusses models of gesture and language from the psychology and linguistics communities. Section 2.2 describes projects that have employed gesture in natural language processing. Section 2.3 describes more general modality fusion techniques, particularly ones that have been used to incorporate prosodic features into NLP. Finally, Section 2.4 considers models from the machine learning literature that are related to conditional modality fusion.

2.1 Models of Gesture and Language

Psychology research has explored the problem of modeling gesture and its relation to language. We discuss two frequently-cited models: Kendon's taxonomy (1980), which focuses on the kinematic structure of individual gestures; and McNeill's (1992), which identifies the ways in which gesture communicates meaning within discourse.

According to Kendon, gestures are constructed from a set of movement phases: prepare, stroke, hold, and retract. The prepare and retract phases initiate and terminate the gesture, respectively. The stroke is the content-carrying part of the gesture, and a hold is a pause that may occur immediately before or after the stroke. These phases provide an essentially kinematic description of gesture. McNeill focuses on the way gestures communicate meaning, identifying four major types of conversational gestures: deictic, iconic, metaphoric, and beat. Deictics communicate by reference to spatial locations, iconics and metaphorics create imagery using the form of the gesture, and beats communicate using timing and emphasis.[5]

While these taxonomies have proved useful for psychological and linguistic research, substantial effort would be required to create a corpus of such annotations for statistical natural language processing. We circumvent this problem by learning a model of gesture directly from automatically-extracted visual features. The relationship between gesture and language semantics is learned in the context of a specific language phenomenon, using annotations only of verbal language semantics. This approach may not capture all meaningful hand gestures, but it will capture those that a) are relevant to coreference resolution and b) can be identified using our feature set.

2.2 Multimodal Natural Language Processing

Early computational work on the relationship between language and gesture focused on identifying examples of connections between discourse elements and automatically-recognized properties of hand gesture (Quek et al., 2000, 2002a). Quek et al. (2002a) show examples in which similar gestures are used in connection with repetitions of the associated discourse elements. We exploit this idea by using features that quantify gesture similarity to predict noun phrase coreference. Follow-up papers attempt to capture the contribution of individual gesture features, such as spatial location (Quek, McNeill, Bryll, & Harper, 2002b) and symmetric motion (Xiong & Quek, 2006), using a similar methodology. This line of work provides a helpful framework for understanding the relationship between gesture and natural language.

On the engineering side, several papers report work on exploiting the relationship between gesture and language. In one line of research, linguistic features are used to improve gesture processing (Poddar, Sethi, Ozyildiz, & Sharma, 1998; Kettebekov, Yeasin, & Sharma, 2005). These papers evaluate performance on human-human language in the domain of weather broadcasts, but with the stated goal of developing techniques for gesture-based human-computer interaction. The authors note that in the domain of weather broadcasts, many hand motions are well-described by a relatively small taxonomy of gestures, which identify points, contours, and regions. Lexical features from ASR transcripts are shown to improve gesture recognition (Poddar et al., 1998), and prosodic features are used to identify key gestural segments (Kettebekov et al., 2005).

Similarly, linguistic analysis has been shown to have important consequences for gesture generation in animated agents. Nakano, Reinstein, Stocky, and Cassell (2003) present an empirical study of human-human interaction, showing a statistical relationship between hand-coded descriptions of head gestures and the discourse labels for the associated utterances (e.g., “acknowledgment,” “answer,” and “assertion”). It is then demonstrated that these findings can be encoded in a model to generate realistic conversational “grounding”

5. McNeill notes that these types should not be thought of a mutually exclusive bins, but as features that may be present in varying degrees.

behavior in an animated agent. In addition to this discourse-moderating function, gestures are also shown to be useful for supplementing the semantic content of verbal explanations. Kopp, Tepper, Ferriman, and Cassell (2007) describe a system in which animated agents give navigation directions, using hand gestures to describe the physical properties of landmarks along the route. While this research describes interesting relationships between gesture and language that can be exploited for generation, we focus on recognition of multimodal communication.

Where the research cited above uses linguistic context to supplement gesture generation and recognition, other work has used gesture features to supplement natural language processing. Much of this research has taken place in the context of spoken language dialogue systems incorporating pen gestures and automatically recognized speech. An early example of such a system is Quickset (Cohen et al., 1997), in which pen gestures and speech input are used to plan military missions. Working in this domain, Johnston and Bangalore (2000) describe a multimodal integration algorithm that parses entire utterances and resolves ambiguity in both the speech and gesture modalities. Chai and Qu (2005) present an alternative take on a similar problem, showing that speech recognition can be improved by increasing the salience of entities targeted by gestures. Both research projects differ from our own in that they assume the ontology of possible referents is known in advance. In addition, because gestures are performed with a pen rather than the free hand, gesture segmentation can be inferred from the contact of the pen with the sensing surface. Finally, the dialogue in both cases is human-computer, rather than human-human, so the language usage probably differs.

The research most similar to our own involves using gesture features to improve language processing in spontaneous human-to-human discourse. Gesture is shown to improve the task of sentence segmentation using automatically recognized features (Chen et al., 2004), and more successfully with manual gesture annotations (Chen et al., 2006). A hidden Markov model (HMM) is used to capture the relation between lexical tokens and sentence boundaries. They then train a maximum entropy model using a feature vector of the posterior probability estimates of the HMM and a set of gesture features based on the Kendon-McNeill taxonomies described above. In earlier work, we show that gesture features can also improve coreference resolution; we describe a system in which a classifier is trained for coreference, using a joint feature vector of gesture and textual features (Eisenstein & Davis, 2006). In both approaches, gesture and speech are combined in an unstructured way, such that even irrelevant hand movements may influence classification decisions. The approach that we present in this paper includes gesture features only when they are likely to be relevant, substantially improving performance above our previous reported results (Eisenstein & Davis, 2006).

2.3 Model Combination Techniques for NLP

There is a large literature on integrating non-verbal features into NLP, much of it relating to prosody. For example, Shriberg, Stolcke, Hakkani-Tur, and Tur (2000) explore the use of prosodic features for sentence and topic segmentation. The first modality combination technique that they consider trains a single classifier with all modalities combined into a single feature vector; this is sometimes called *early fusion.* They also consider training separate classifiers and combining their posteriors, either through weighted addition or multiplication; this is sometimes called *late fusion* (see also Liu, 2004). Experiments on multimodal fusion with prosodic features find no conclusive winner among early fusion, additive late fusion, and multiplicative late fusion (Shriberg et al., 2000; Kim, Schwarm, & Osterdorf, 2004). These techniques have also been employed for gesture-speech fusion. In prior work, we employed early fusion for gesture-speech combination (Eisenstein & Davis, 2006); late fusion has also been applied to gesture-speech combination (Chen et al., 2004, 2006).

Toyama and Horvitz (2000) introduce a Bayesian network approach to modality combination for speaker identification. As in late fusion, modality-specific classifiers are trained independently. However, the Bayesian approach also learns to predict the reliability of each modality on a given instance, and incorporates this information into the Bayes net. While more flexible than early or late fusion, training modality-specific classifiers separately is still suboptimal compared to training them jointly, because independent training of the modality-specific classifiers forces them to account for data that they cannot possibly explain. For example, if the speaker's gestures are not relevant to the language processing task, it is counterproductive to train a gesture-modality classifier on the features at this instant; doing so can lead to overfitting and poor generalization.

Our approach combines aspects of both early and late fusion. As in early fusion, classifiers for all modalities are trained jointly. But as in Toyama and Horvitz's Bayesian late fusion model, modalities can be weighted based on their predictive power for specific instances. In addition, our model is trained to maximize conditional likelihood, rather than joint likelihood.

2.4 Related Machine Learning Approaches

From a machine learning perspective, our research relates to three general areas: domain adaptation, co-training, and hidden-variable conditional models.

In *domain adaptation*, one has a small amount of "in-domain" data that is relevant to the target classification task, and a large amount of "out-of-domain" data from a related, but different task (Blitzer, McDonald, & Pereira, 2006; Chelba & Acero, 2006). The goal is to use the out-of-domain data to improve performance in the target domain. In one recent approach, each feature is replicated and separate weights are learned for the general and domain-specific applications of the feature (Daumé III, 2007). In a sense, the model learns which features are relevant generally, and which are relevant only in specific domains. Our task is somewhat similar, in that we are interested in learning when to apply the gesture features, while simultaneously learning how they predict coreference. However, one key difference is that in domain adaptation, the data is partitioned into separate domains in advance, while our model must learn to identify cases in which the gesture is salient.

Co-training is another technique for combining multiple datasets (Blum & Mitchell, 1998). In co-training, a small amount of labeled data is supplemented by a large amount of unlabeled data. Given sets of features that are each sufficient to predict the desired label – called "views" – separate classifiers can be trained such that the predictions of one classifier provide the labeled data for the other classifier. Such an approach is shown to yield better performance than using only the labeled data in some applications, such as parsing (Sarkar, 2001). If large amounts of unlabeled data were available, co-training could be applied here, using the gesture and verbal features for the independent views. In our research, acquiring data is a greater bottleneck than creating the coreference annotations. In addition, previous attempts to apply co-training to textual coreference resolution proved largely unsuccessful (Müller, Rapp, & Strube, 2002), possibly because the views were not independently sufficient to predict the label. While further investigation on this topic is merited, our approach does not make use of any unlabeled data; instead, we treat gestures salience as a hidden variable within our existing dataset.

From a methodological standpoint, our work is most closely related to the literature on hidden variables in conditionally trained models. Quattoni, Collins, and Darrell (2004) improve object recognition through the use of a hidden variable indicating which "part" of the object contains each localized visual feature. Part-based object recognition had previously been performed in a generative framework, but the conditional approach permits the use of a broader feature set, without concern for whether the features are mutually independent. Subsequent work has shown how conditional hidden-variable models can be used in gesture recognition (Wang, Quattoni, Morency, Demirdjian, & Darrell, 2006) and language processing (Koo & Collins, 2005; Sutton, McCallum, & Rohanimanesh, 2007). Wang et al. (2006) employ a model that is similar to HMM-based gesture recognition, with the hidden variable encoding different phases of the gesture to be recognized; again, the conditional approach is shown to improve performance. Hidden variables are applied to statistical parsing by Koo and Collins (2005), assigning lexical items to word clusters or word senses. Finally, Sutton et al. (2007) use hidden variables to encode intermediate levels of linguistic structure that are relevant to the overall language-processing task. For example, in one application, hidden variables encode part-of-speech tags, which are then used for noun phrase chunking. We continue this line of work, extending hidden-variable conditional models with a novel, linguistically-motivated hidden-variable architecture for gesture-speech combination.

3. Dataset

The research described in this paper is based on a corpus of multimodal presentations. There are a few existing corpora that include visual data, but none are appropriate for our research. The Ami corpus (Carletta et al., 2005) includes video and audio from meetings, but participants are usually seated and their hands are often not visible in the video. The Vace corpus (Chen et al., 2005) also contains recordings of meetings, with tracking beacons attached to the speakers providing very accurate tracking. This corpus has not been publicly released at the time of this writing.

Both corpora address seated meeting scenarios; we have observed that gesture is more frequent when speakers give standing presentations, as in classroom lectures or business presentations. There are many such video recordings available, but they have typically been filmed under circumstances that frustrate current techniques for automatic extraction of visual features, including camera movement, non-static background, poor lighting, and occlusion of the speaker. Rather than focusing on these substantial challenges for computer vision, we chose to gather a new multimodal corpus.

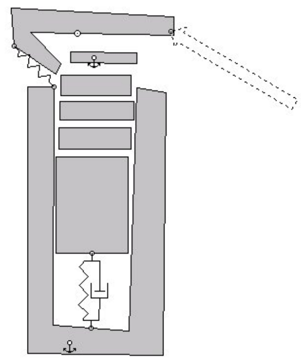


Figure 2: An example pre-printed diagram used in gathering the corpus. The diagram is a schematic depiction of a candy dispenser.

In gathering our corpus, we aimed to capture conversations in which gesture was frequent and direct, but also natural and unsolicited. We sought a middle ground between task-oriented dialogues such as Trains (Allen et al., 1995) and completely open-ended discussions such as Switchboard (Godfrey, Holliman, & McDaniel, 1992). In our work, participants were given specific topics for discussion (usually the function of mechanical devices), but were then permitted to converse without outside interference. The speakers were given pre-printed diagrams to aid their explanations. The interpretation of gestures in this condition is usually relatively straightforward; many, if not most of the gestures involve pointing at locations on the diagram. Visual aids such as printed or projected diagrams are common to important application areas, including business presentations, classroom lectures, and weather reports. Thus, this restriction does not seem overly limiting to the applicability of our work. We leave the presumably more challenging problem of understanding the gestures that are produced without visual aids to future work.

Figure 1 shows two still frames from our corpus, with the accompanying text. The visual aid is shown in more detail in Figure 2. Our corpus includes sixteen short videos from nine different speakers. A total of 1137 noun phrases were transcribed; this is roughly half the number found in the MUC6 training set, a text-only dataset that is also used for coreference resolution (Hirschman & Chinchor, 1998). Building a multimodal corpus is a time-consuming task requiring substantial manpower, but we hope that this initial work will lead to larger future corpora that are well-suited for the study of gesture in natural language processing. Corpus statistics can be found in Appendix C, and the data is available on-line at: `http://mug.csail.mit.edu/publications/2008/Eisenstein_JAIR/`

Finally, we draw the reader's attention to the differences between this corpus and commonly-used textual corpora in coreference resolution, such as MUC (Hirschman & Chinchor, 1998). Topically, this corpus focuses on description of mechanical devices, rather than news articles. Consequently, the emphasis is less on disambiguating entities such as people and organizations, and more on resolving references to physical objects. The corpora also differ in genre, with our corpus comprised of spontaneous speech, while the MUC corpus includes edited text. Such genre distinctions are known to play an important role in patterns of reference (Strube & Müller, 2003) and language use generally (Biber, 1988). Four different mechanical devices were used as topics of discussion: a piston, candy dispenser (Figure 2), latch box (shown in Appendix B), and pinball machine.

3.1 Data Gathering Protocol

Fifteen pairs of participants joined the study by responding to posters on our university campus; their ages ranged from 18-32, and all participants were university students or staff. A subset of nine pairs of participants was selected on the basis of recording quality,[6] and their speech was transcribed and annotated. The corpus is composed of two videos from each of the nine pairs; audio recording problems forced us to exclude two videos, yielding 16 annotated documents, each between two and three minutes in duration.

One participant was randomly selected from each pair to be the "speaker," and the other to be the "listener." The speaker's role was to explain the behavior of a mechanical device to the listener. The listener's role was to understand the speaker's explanations well enough to take a quiz later. Prior to each discussion, the speaker privately viewed a simulation of the operation of the relevant device.

The speaker was limited to two minutes to view the video or object and three minutes to explain it; the majority of speakers used all of the time allotted. This suggests that we could have obtained more natural data by not limiting the explanation time. However, we found in pilot studies that this led to problematic ordering effects, where participants devoted a long time to the early conditions, and then rushed through later conditions. With these time constraints, the total running time of the experiment was usually around 45 minutes. The data used in this study is part of a larger dataset initially described by Adler, Eisenstein, Oltmans, Guttentag, and Davis (2004).

3.2 Speech Processing

Speech was recorded using headset microphones. An integrated system controlled the synchronization of the microphones and video cameras. Speech was transcribed manually, and audio was hand-segmented into well-separated chunks with duration not longer than twenty seconds. The chunks were then force-aligned by the Sphinx-II speech recognition system (Huang, Alleva, Hwang, & Rosenfeld, 1993).

A wide range of possibilities exist regarding the fidelity and richness of transcribed speech. Choices include transcription quality, existence of punctuation and capitalization, the presence of sentence boundaries and syntactic annotations. We assume a perfect transcription of words and sentence boundaries,[7] but no additional punctuation. This is similar to much of the NLP research on the Switchboard corpus, (e.g., Kahn, Lease, Charniak, Johnson, & Ostendorf, 2005; Li & Roth, 2001), although automatic speech recognition (ASR) transcripts are also used (e.g., Shriberg et al., 2000). Using ASR may more accurately replicate the situation for an application developer. However, such an approach would also introduce a certain arbitrariness, as results would depend heavily on the amount

6. Difficulties with the microphones prevented us from getting suitable audio recordings in several cases; in other cases there were difficulties in synchronizing the two microphones and two video cameras.
7. Sentence boundaries were annotated according to the NIST Rich Transcription Evaluation (NIST, 2003).

of effort spent tuning the recognizer. In particular, if the recognizer is not well-tuned, this approach risks overstating the relative contribution of gesture features, because the verbal features would then be of little value.

Our natural language task of coreference resolution requires noun phrase boundaries as a preprocessing step, and we provide gold-standard noun phrase annotation. Our goal is to isolate the contribution of our model for gesture-speech combination on the coreference task, and thus we did not wish to deliberately introduce noise in the noun phrase boundaries. Gold standard noun phrase annotations have been used in previous research on coreference resolution, (e.g., McCallum & Wellner, 2004; Haghighi & Klein, 2007).[8] In addition, automatic noun phrase chunking is now possible with high accuracy. F-measures exceeding .94 have been reported on textual corpora (Kudo & Matsumoto, 2001; Sha & Pereira, 2003); on transcripts of the Switchboard corpus, state-of-the-art performance exceeds .91 (Li & Roth, 2001).[9]

The annotation of noun phrases followed the MUC task definition for "markable" NPs (Hirschman & Chinchor, 1998). Personal pronouns were not annotated, as the discourse focused on descriptions of mechanical devices. Such pronouns could easily be filtered out automatically. Annotation attempted to transcribe all other noun phrases. A total of 1137 markable NPs were transcribed. This is roughly half the size of the MUC6 training set, which includes 2072 markable NPs over 30 documents. The gold standard coreference and markable annotation was performed by the first author, using both the audio and video information.

An additional rater performed coreference annotations to help assess validity. The rater is a native speaker of English and is not an author on this paper. She annotated two documents, comprising a total of 270 noun phrases. Using the interrater agreement methodology described by Passonneau (1997), a score of .65 is obtained on Krippendorf's alpha. This is comparable to some of the results from the MUC textual corpus (Passonneau, 1997), but higher than the agreement reported on a corpus of multi-party spoken dialogues (Müller, 2007).

Finally, we assume gold standard sentence boundaries, but no additional punctuation.

3.3 Vision Processing

Video recording was performed using standard digital camcorders. Participants were given two different colored gloves to facilitate hand tracking. Despite the use of colored gloves, a post-study questionnaire indicated that only one of the thirty participants guessed that the study was related to gesture. The study was deliberately designed so that participants had very little free time to think; when not actually conducting the dialogue, the speaker was busy viewing the next mechanical system, and the participant was busy being tested on the previous conversation. We also presented consent forms immediately after the gloves, which may have diverted attention from the gloves' purpose.

8. The cited references do not include noun phrase that unless they participate in coreference relations; we include all noun phrases regardless.
9. This high accuracy on switchboard does not imply good performance on our data, since we do not have annotated data for noun phrase boundaries. Thus the overall impact of noisy preprocessing on coreference performance is unknown. In addition, it is possible that noisy noun phrase boundaries may pose particular problems for our approach, which assesses gesture features over the duration of the NP.

An articulated upper-body tracker was used to model the position of the speaker's torso, arms, and hands. By building a complete upper-body tracker, rather than simply tracking the individual hands, we were able to directly model occlusion of the hands and arms. At each frame, an annealed particle filter is used to search the space of body configurations. Essentially, the system performs a randomized beam search to simultaneously achieve three objectives: a) maximize the overlap between the model and pixels judged to be in the foreground, b) match the known glove color to the color observed at the hypothesized hand positions, c) respect physiological constraints and temporal continuity. The system was implemented using the OpenCV library.[10]

The tracker was inspired largely by the annealed particle filter of Deutscher, Blake, and Reid (2000); the main differences were that Deutscher et al. did not use color cues such as gloves, but did use multiple cameras to facilitate 3D tracking. We used only a single monocular camera and a 2.5D model (with just one degree of freedom in the depth plane, permitting body rotation). Parameters of the model, such as the body dimensions, are customized for each speaker. Each speaker provided two different explanations, and the segmentation of these videos was performed manually. No additional post-processing, calibration, or "cleaning" of the tracker output is performed.

From inspection, the lack of depth information appears the cause of many of our system's errors; bending of the arm joints in the depth dimension caused the arm length to appear to change in ways that were confusing to our model. Nonetheless, we estimate by manual examination of the tracking output that both hands were tracked accurately and smoothly over 90% of the time when not occluded. It is difficult to assess the tracker performance more precisely, as that would require ground truth data in which the actual hand positions were annotated manually at each time step.

4. Conditional Modality Fusion for Coreference Resolution

In this section we describe conditional modality fusion. In Section 4.1 we describe how hidden variables are incorporated in conditional models. Then in Section 4.2, we describe how various theories of model combination are expressed in this framework. In Section 4.3, we give details of our implementation.

4.1 Hidden Variables in Conditional Models

Our goal is to learn to use non-verbal features to make predictions when they are helpful, and ignore them when they are not. We call this approach *conditional modality fusion*. More formally, we are trying to predict a label $y \in \{-1, 1\}$, representing a single binary coreference decision of whether two noun phrases refer to the same entity.

The hidden variable $\mathbf{h}$ describes the salience of the gesture features. The observable features are written as $\mathbf{x}$, and our model is to learn a set of weights $\mathbf{w}$. Our hidden variable approach learns to predict y and $\mathbf{h}$ jointly, given labeled training data only for y. We use a conditional model, writing:

10. http://www.intel.com/technology/computing/opencv/

$$
\begin{aligned}
p(y|\mathbf{x};\mathbf{w}) &= \sum_{\mathbf{h}} p(y,\mathbf{h}|\mathbf{x};\mathbf{w}) \\
&= \frac{\sum_{\mathbf{h}} exp(\psi(y,\mathbf{h},\mathbf{x};\mathbf{w}))}{\sum_{y',\mathbf{h}} exp(\psi(y',\mathbf{h},\mathbf{x};\mathbf{w}))}.
\end{aligned}
$$

Here, ψ is a potential function representing the compatibility between the label y, the hidden variable $\mathbf{h}$, and the observations $\mathbf{x}$; this potential is parameterized by a vector of weights, $\mathbf{w}$. The numerator expresses the compatibility of the label y and observations $\mathbf{x}$, summed over all possible values of the hidden variable $\mathbf{h}$. The denominator sums over both $\mathbf{h}$ and all possible labels y', yielding the conditional probability $p(y|\mathbf{x};\mathbf{w})$.

This model can be trained by a gradient-based optimization to maximize the conditional log-likelihood of the observations. The unregularized log-likelihood and gradient are given by:

$$
\begin{aligned}
l(\mathbf{w}) &= \sum_i \log(p(y_i|\mathbf{x}_i;w)) && (1) \\
&= \sum_i \log \frac{\sum_{\mathbf{h}} \exp(\psi(y_i,\mathbf{h},\mathbf{x}_i;\mathbf{w}))}{\sum_{y',\mathbf{h}} \exp(\psi(y',\mathbf{h},\mathbf{x}_i;\mathbf{w}))} && (2)
\end{aligned}
$$

$$
\frac{\partial l_i}{\partial w_j} = \sum_{\mathbf{h}} p(\mathbf{h}|y_i,\mathbf{x}_i;\mathbf{w}) \frac{\partial}{\partial w_j}\psi(y_i,\mathbf{h},\mathbf{x}_i;\mathbf{w}) - \sum_{y',\mathbf{h}} p(\mathbf{h},y'|\mathbf{x}_i;\mathbf{w}) \frac{\partial}{\partial w_j}\psi(y',\mathbf{h},\mathbf{x}_i;\mathbf{w})
$$

The use of hidden variables in a conditionally-trained model follows Quattoni et al. (2004). However, while this reference gives the general outline for hidden-variable conditional models, the form of the potential function depends on the role of the hidden variable. This is problem-specific, and a novel contribution of our research is the exploration of several different potential functions, permitting different forms of modality fusion.

4.2 Models of Modality Fusion

The form of the potential function ψ is where our intuitions about the role of the hidden variable are formalized. We consider three alternative forms for ψ, capturing different theories of gesture-speech integration. The models range from a simple concatenation of gesture-speech features to a structured fusion model that dynamically assesses the relevance of gesture features for every noun phrase.

The models we consider are influenced by our goal, which is to determine whether two noun phrases (NPs) are coreferent. Gesture salience is assessed at each NP, to determine whether the gestural features should influence our decision about whether the noun phrases corefer. We set $\mathbf{h} = \langle h_1, h_2 \rangle$, with $h_1 \in \{1, -1\}$ representing gesture salience during the first noun phrase (antecedent), and $h_2 \in \{1, -1\}$ representing gesture salience during the second noun phrase (anaphor).

4.2.1 Same-Same model

In the trivial case, we ignore the hidden variable and always include the features from both gesture and speech. Since the weight vectors for both modalities are unaffected by the hidden variable, this model is referred to as the "same-same" model. Note that this is identical to a standard log-linear conditional model, concatenating all features into a single vector. This model is thus a type of "early fusion," meaning that the verbal and non-verbal features are combined prior to training.

$$\psi_{ss}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv y(\mathbf{w}_v^T \mathbf{x}_v + \mathbf{w}_{nv}^T \mathbf{x}_{nv}) \tag{3}$$

$\mathbf{x}_v$ and $\mathbf{w}_v$ refer to the features and weights for the verbal modality; $\mathbf{x}_{nv}$ and $\mathbf{w}_{nv}$ refer to the non-verbal modality.

4.2.2 Same-Zero Model

Next, we consider a model that treats the hidden variable as a gate governing whether the gesture features are included. This model is called the "same-zero" model, since the verbal features are weighted identically regardless of the hidden variable, and the gesture feature weights go to zero unless $h_1 = h_2 = 1$.

$$\psi_{sz}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_v^T \mathbf{x}_v + \mathbf{w}_{nv}^T \mathbf{x}_{nv}) + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}, & h_1 = h_2 = 1 \\ y \mathbf{w}_v^T \mathbf{x}_v + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}, & \text{otherwise.} \end{cases} \tag{4}$$

The features $\mathbf{x}_h$ and weights $\mathbf{w}_h$ contribute to the estimation of the hidden variable $\mathbf{h}$. They may include some or all of the features from $\mathbf{x}_v$ and $\mathbf{x}_{nv}$, or different features. These features are assessed independently at each noun phrase, yielding $\mathbf{x}_{h_1}$ for the antecedent and $\mathbf{x}_{h_2}$ for the anaphor.

This model reflects the intuition that gesture features (measured by $\mathbf{x}_{nv}$) are relevant only when the gestures during *both* noun phrases are salient. Thus, these features contribute towards the overall potential only when $h_1 = h_2 = 1$.

4.2.3 Different-Zero Model

We may add flexibility to our model by permitting the weights on the verbal features to change with the hidden variable. This model is called the "different-zero" model, since a different set of verbal weights ($\mathbf{w}_{v,1}$ or $\mathbf{w}_{v,2}$) is used depending on the value of the hidden variable. Such a model is motivated by empirical research showing language usage is different when used in combination with meaningful non-verbal communication than when it is used unimodally (Kehler, 2000; Melinger & Levelt, 2004).

The formal definition of the potential function is:

$$\psi_{dz}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_{v,1}^T \mathbf{x}_v + \mathbf{w}_{nv}^T \mathbf{x}_{nv}) + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}, & h_1 = h_2 = 1 \\ y \mathbf{w}_{v,2}^T \mathbf{x}_v + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}, & \text{otherwise.} \end{cases} \tag{5}$$

4.2.4 Other Models

We have presented three models of increasing complexity; the "different-different" model is one step more complex, including two pairs of weight vectors for both verbal and gestural features (see Equation 6). In this model, the distinction between verbal and non-verbal features ($\mathbf{x}_v$ and $\mathbf{x}_{nv}$) evaporates, and there is no reason that the hidden variable $\mathbf{h}$ should actually indicate the relevance of the non-verbal features. In addition, the high degree of freedom of this model may lead to overfitting.

$$\psi_{dd}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_{v,1}^T \mathbf{x}_v + \mathbf{w}_{nv,1}^T \mathbf{x}_{nv}) + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}, & h_1 = h_2 = 1 \\ y(\mathbf{w}_{v,2}^T \mathbf{x}_v + \mathbf{w}_{nv,2}^T \mathbf{x}_{nv}) + h_1 \mathbf{w}_h^T \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^T \mathbf{x}_{h_2}, & \text{otherwise.} \end{cases} \tag{6}$$

The models that we have considered assume that the verbal features are always relevant, while the gesture features may sometimes be ignored. In other words, we have not considered whether it might be necessary to assess the salience of the verbal features. One might consider alternative potential functions such as a "zero-same" model, in which the verbal features were sometimes ignored. We did not consider such models, as gesture unaccompanied by speech is extremely rare in our dataset.

4.3 Implementation

The objective function (Equation 1) is optimized using a Java implementation of L-BFGS, a quasi-Newton numerical optimization technique (Liu & Nocedal, 1989). Standard L2-norm regularization is employed to prevent overfitting, with cross-validation to select the regularization constant. Java source code is available online:
`http://rationale.csail.mit.edu/gesture`

Although standard logistic regression optimizes a convex objective, the inclusion of the hidden variable renders our objective non-convex. Thus, convergence to a global optimum is not guaranteed, and results may differ depending on the initialization. Nonetheless, non-convexity is encountered with many models in natural language processing and machine learning generally, such as Baum-Welch training of hidden Markov models (HMMs) (Rabiner, 1989) or hidden-state conditional random fields (Quattoni et al., 2004; Sutton & McCallum, 2006). Often, results can be shown to be reasonably robust to initialization; otherwise, multiple restarts can be used to obtain greater stability. We present an empirical evaluation in Section 6.2 showing that our results are not overly sensitive to initialization. In all other experiments, weights are initialized to zero, enabling the results to be reproduced deterministically.

5. Features

Coreference resolution has been studied for over thirty years in the AI community (Sidner, 1979; Kameyama, 1986; Brennan, Friedman, & Pollard, 1987; Lappin & Leass, 1994; Walker, 1998; Strube & Hahn, 1999; Soon, Ng, & Lim, 2001; Ng & Cardie, 2002). Based on this large body of work, there is a broad consensus on a core set of useful verbal features. This paper contributes to the literature with the study of gesture features, both for multimodal coreference resolution and for identifying salient gestures. We describe these

Pairwise verbal features		
feature	**type**	**description**
edit-distance	similarity	a numerical measure of the string similarity between the two NPs
exact-match	similarity	true if the two NPs are identical
str-match	similarity	true if the NPs are identical after removing articles
nonpro-str	similarity	true if the antecedent i and the anaphor j are not pronouns, and str-match is true
pro-str	similarity	true if i and j are pronouns, and str-match is true
j-substring-i	similarity	true if the j is a substring of the i
i-substring-j	similarity	true if i is a substring of j
overlap	similarity	true if there are any shared words between i and j
np-dist	centering-based	the number of noun phrases between i and j in the document
sent-dist	centering-based	the number of sentences between i and j in the document
both-subj	centering-based	true if both i and j precede the first verb of their sentences
same-verb	centering-based	true if the first verb in the sentences for i and j is identical
number-match	compatibility	true if i and j have the same number
Single-phrase verbal features		
pronoun	centering-based	true if the NP is a pronoun
count	centering-based	number of times the NP appears in the document
has-modifiers	centering-based	true if the NP has adjective modifiers
indef-np	centering-based	true if the NP is an indefinite NP (e.g., *a fish*)
def-np	centering-based	true if the NP is a definite NP (e.g., *the scooter*)
dem-np	centering-based	true if the NP begins with *this*, *that*, *these*, or *those*
lexical features	centering-based	lexical features are defined for the most common pronouns: *it*, *that*, *this*, and *they*

Table 1: The set of verbal features for multimodal coreference resolution. In this table, i refers to the antecedent noun phrase and j refers to the anaphor.

features in Sections 5.2 and 5.3, but we begin with a review of the verbal features which we have selected from the literature.

5.1 Verbal Features

Our selection of verbal features is motivated by the extensive empirical literature on text-based coreference resolution (Soon et al., 2001; Ng & Cardie, 2002; Strube & Müller, 2003; Daumé III & Marcu, 2005). The proliferation and variety of features that have been explored is a consequence of the fact that coreference is a complex discourse phenomenon. Moreover, realization of coreference is highly dependent on the type of discourse in which it appears; relevant factors include the modality (e.g., speech vs. language), genre (e.g., meeting vs. lecture) and topic (e.g., politics vs. scientific subject). Although certain feature types are application-specific, three classes of features – centering-based, similarity, and compatibility features – are useful across most coreference applications. These classes form a basis of the verbal features used by our model. Table 1 provides a brief description of our verbal feature set. We draw examples from the transcript in Appendix A to provide a more detailed explanation of these features and motivate their use in our application.

Pairwise gesture features	
focus-distance	the Euclidean distance in pixels between the average hand position during the two NPs
DTW-agreement	a measure of the agreement of the hand-trajectories during the two NPs, computed using dynamic time warping
same-cluster*	true if the hand positions during the two NPs fall in the same cluster
JS-div*	the Jensen-Shannon divergence between the cluster assignment likelihoods
Single-phrase gesture features	
dist-to-rest	distance of the hand from rest position
jitter	sum of instantaneous motion across NP
speed	total displacement over NP, divided by duration
rest-cluster*	true if the hand is usually in the cluster associated with rest position
movement-cluster*	true if the hand is usually in the cluster associated with movement

Table 2: The set of gesture features for multimodal coreference resolution. Features not used in prior work on gesture analysis are annotated with an asterisk (*).

- **Centering-related features:** This set of features captures the relative prominence of a discourse entity in a local discourse, and its likelihood to act as a coreferent for a given phrase. These features are inspired by linguistic analysis formalized in Centering Theory, which links coreferential status of an entity with its discourse prominence (Grosz, Joshi, & Weinstein, 1995; Walker, Joshi, & Prince, 1998; Strube & Hahn, 1999; Poesio, Stevenson, Eugenio, & Hitzeman, 2004; Kibble & Power, 2004). This theory hypothesizes that at any point of a coherent discourse, only one entity is in focus and it characterizes local discourse in terms of focus transitions between adjacent sentences.

  Most of the existing machine-learning based coreference systems do not attempt to fully implement Centering-style analysis.[11] Instead, a number of centering-related features are included. For instance, to identify focus-preserving transitions (i.e., CONTINUE transitions) a feature BOTH-SUBJ is introduced. According to the theory, such transitions are common in locally-coherent discourse, and therefore coreference assignments consistent with this principle may be preferable. We also characterize transitions in terms of their span (NP-DIST AND SENT-DIST). Transitions that involve short gaps are preferred over transitions with long gaps.

  Another important set of Centering-related features is defined at the level of a single phrase. The syntactic role of a phrase in a sentence – captured in features such as PRONOUN, HAS-MODIFIERS, INDEF-NP – indicates its discourse prominence and therefore its likelihood to be a coreference antecedent. For example, consider an utterance from lines 12 and 13: "and this spring is active meaning that its going up and down." Here, the anaphor "it" clearly refers to the antecedent "this spring." The

11. Such an implementation is challenging in several respects: one has to specify the "free parameters" of the system (Poesio et al., 2004) and to determine ways of combining the effects of various constraints. Additionally, an implementation of centering depends on obtaining detailed syntactic information, which is not available in our case.

fact that the antecedent is a demonstrative noun phrase (beginning with "this")[12] and that the anaphor is a pronoun are also centering-related features that suggest coreference is likely. In addition to the syntactic status, we also take into account the frequency of a noun phrase in a monologue (see count). Frequency information is commonly used to approximate topical salience of an entity in a text (Barzilay & Lapata, 2005).

- **Similarity features:** A simple yet informative set of coreference cues are based on string-level similarity between noun phrases. For instance, the reference between "this spring" in line 12 and the identical noun phrase in line 5 can be resolved by the exact match of the surface forms. In general, researchers in text-based coreference resolution have found that the string match feature is the single most predictive feature because a discourse entity is commonly described using identical or similar noun phrases (Soon et al., 2001).

  In our system, the similarity information is captured in seven features that quantify the degree of string overlap. For instance, the feature (exact-match) indicates full overlap between noun phrases, while the feature (overlap) captures whether two phrases share any common words. In the context of coreference resolution, noun phrase match is more informative than pronoun match, so we use distinct features for matching strings in these syntactic categories (e.g., nonpro-str vs. pro-str), following (Ng & Cardie, 2002). Surface similarity may also be quantified in terms of edit-distance (Strube, Rapp, & Müller, 2002).

- **Compatibility features:** An important source of coreference information is compatibility between two noun phrases. For instance, the utterance "the ball" in line 11 can *not* refer to the preceding noun phrase "these things," since they are incompatible in number. Feature number-match captures this information. Since the topic of discourse in our corpus relates to mechanical devices, almost all noun phrases are neuter-gendered. This eliminates the utility of features that measure gender compatibility. Finally, we note that more complex semantic compatibility features have previously been explored (Harabagiu, Bunescu, & Maiorano, 2001; Strube et al., 2002; Yang, Zhou, Su, & Tan, 2003; Ji, Westbrook, & Grishman, 2005; Daumé III & Marcu, 2005; Yang, Su, & Tan, 2005).

Some features that are traditionally used in coreference were avoided here. Features that depend on punctuation seem unlikely to be applicable in an automatic recognition setting, at least in the near future. In addition, while many systems in the MUC and ACE coreference corpora use "gazetteers" that list the names of nations and business entities, such features are not relevant to our corpus. Another possibility is to use features identifying the speaker, to capture individual variation in patterns of reference (Chai, Hong, Zhou, & Prasov, 2004; Jordan & Walker, 2005). However, we wished to develop an approach that was speaker-independent.

12. Simple string matching techniques are used to assess phrase types: definite noun phrases are those beginning with the article "the"; indefinite noun phrases begin with "a" or "an"; demonstrative noun phrases begin with "this." Bare plurals are not marked as indefinites, and proper names do not appear in the dataset.

5.2 Non-Verbal Features

Our non-verbal features attempt to capture similarity between the speaker's hand gestures, as similar gestures can suggest semantic similarity (McNeill, 1992; Quek et al., 2002a). For example, two noun phrases may be more likely to corefer if they are accompanied by identical pointing gestures. In this section, we describe features that quantify various aspects of gestural similarity.

In general, these features are computed over the duration of each noun phrase, yielding a single feature vector per NP. While it is not universally true that the beginning and end points of relevant gestures line up exactly with the beginning and end of the associated words, several experiments have demonstrated the close synchrony of gesture and speech (McNeill, 1992). In future work, we hope to explore whether more sophisticated gesture segmentation techniques can improve performance.

The most straightforward measure of similarity is the Euclidean distance between the average hand position during each noun phrase – we call this FOCUS-DISTANCE.[13] Euclidean distance captures cases in which the speaker is performing a gestural "hold" in roughly the same location (McNeill, 1992). However, Euclidean distance may not correlate directly with semantic similarity. For example, when gesturing at a detailed part of a diagram, very small changes in hand position may be semantically meaningful, while in other regions positional similarity may be defined more loosely. Ideally, we would compute a semantic feature capturing the *object* of the speaker's reference (e.g., "the red block"), but this is not possible in general, since a complete taxonomy of all possible objects of reference is usually unknown.

Instead, we perform a spatio-temporal clustering on hand position and velocity, using a hidden Markov model (HMM). Hand position and speed are used as observations, and are assumed to be generated by Gaussians, indexed by the model states. The states themselves correspond to clusters, and cluster membership can be used as a discretized representation of positional similarity. Inference of state membership and learning of model parameters are performed using the traditional forward-backward and Baum-Welch algorithms (Rabiner, 1989).

While a standard hidden Markov model may be suitable, we can increase robustness and make better use of available training data by reducing the model's degrees-of-freedom. Reducing the number of degrees-of-freedom means that we are learning simpler models, which are often more general. This is done through *parameter tying*: requiring some subsets of model parameters to take the same values (Bishop, 2006). We employ three forms of parameter tying:

1. Only one state is permitted to have an expected speed greater than zero. This state is called the "move" state; all other states are "hold" states, and their speed observations are assumed to be generated by zero-mean Gaussians. Only a single "move" state is used because we are most concerned about the location of hold gestures.

2. Transitions between distinct hold states are not permitted. This reflects the commonsense idea that it is not possible to transition between two distinct positions without moving.

13. In general, features are computed over the duration of individual noun phrases.

3. The outgoing transition probabilities from all hold states are assumed to be identical. Intuitively, this means that the likelihood of remaining within a hold state does not depend on where that hold is located. While it is possible to imagine scenarios in which this does not hold, it is a reasonable simplification that dramatically reduces the number of parameters to be estimated.

Two similarity features are derived from the spatio-temporal clustering. The SAME-CLUSTER feature reports whether the two gestures occupy the same state for the majority of the durations of the two noun phrases. This is a boolean feature that indicates whether two gestures are in roughly the same area, without need for an explicit discretization step. However, two nearby gestures may not be classified as similar by this method, if they are near the boundary between two states, or if both gestures move between multiple states. For this reason, we quantify the similarity of the state assignment probabilities using the Jensen-Shannon divergence, a metric on probability distributions (Lin, 1991). JS-DIV is a real-valued feature that provides a more nuanced view of the gesture similarity based on the HMM clustering. Both SAME-CLUSTER and JS-DIV are computed independently for models comprising five, ten, and fifteen states.

The gesture features described thus far largely capture the similarity between static gestures; that is, gestures in which the hand position is nearly constant. However, these features do not capture the similarity between gesture trajectories, which may also be used to communicate meaning. For example, a description of two identical motions might be expressed by very similar gesture trajectories. To measure the similarity between such dynamic gestures, we use dynamic time warping (Huang, Acero, & Hon, 2001); this is reported in the DTW-DISTANCE feature. Dynamic time warping has been used frequently in recognition of predefined gestures (Darrell & Pentland, 1993).

All features are computed from hand and body pixel coordinates, which are obtained automatically via computer vision, without manual post-processing of any kind (see Section 3.3). Our feature set currently supports only single-hand gestures, using the hand that is farthest from the body center. As with the verbal feature set, WEKA's default supervised binning class was applied to the continuous-valued features (Fayyad & Irani, 1993).[14] This method identifies "cut points" that minimize the class-wise impurity of each side of the cut, measured using average class entropy. A greedy top-down approach is used to recursively partition the domain of an attribute value. Partitioning is terminated by a criterion based on minimum description length.

5.3 Meta Features

Meta features are observable properties of the speech and gesture that give clues about whether the speaker is gesturing in a way that is meaningful for the language language processing task at hand. We hypothesize that the difference between relevant and irrelevant hand motions is apparent in a range of verbal and visual features. In equations 4-6, these features are represented by $\mathbf{x}_{h_1}$ and $\mathbf{x}_{h_2}$. Unlike the similarity-based features described above, meta features must be computable at a single instant in time, as they encode properties of individual gestures and their cotemporal NPs.

14. The specific class is `weka.filters.supervised.attribute.Discretize`.

Previous research has investigated which types of verbal utterances are likely to be accompanied by gestural communication (Melinger & Levelt, 2004; Kehler, 2000). However, this is the first attempt to formalize this relationship in the context of a machine-learning approach that predicts gesture salience.

5.3.1 Verbal Meta Features

Meaningful gesture has been shown to be more frequent when the associated speech is ambiguous (Melinger & Levelt, 2004). Kehler (2000) finds that fully-specified noun phrases are less likely to receive multimodal support. These findings lead us to expect that gestures should be likely to co-occur with pronouns, and unlikely to co-occur with noun phrases that begin with the determiner "the," particularly if they include adjectival modifiers. To capture these intuitions, all single-phrase verbal features (Table 1) are included as meta-features.

5.3.2 Non-verbal Meta Features

Research on gesture has shown that semantically meaningful hand motions usually take place away from "rest position," which is located at the speaker's lap or sides (McNeill, 1992). Effortful movements away from these default positions can thus be expected to predict that gesture is being used to communicate. We identify rest position as the center of the body on the x-axis, and at a fixed, predefined location on the y-axis. The DIST-TO-REST feature computes the average Euclidean distance of the hand from the rest position, over the duration of the NP.

Hand speed may also be related to gesture salience. The SPEED feature captures the overall displacement (in pixels) divided by the length of the noun phrase. Writing $\mathbf{x}$ for the hand position and $t \in \{1, 2, \ldots, T\}$ for the time index, we have $\text{SPEED} = \frac{1}{T}||\mathbf{x}_T - \mathbf{x}_1||_2$. The JITTER feature captures the average *instantaneous* speed: $\text{JITTER} = \frac{1}{T}\sum_{t=2}^{T}(\mathbf{x}_t - \mathbf{x}_{t-1})^T(\mathbf{x}_t - \mathbf{x}_{t-1})$. This feature captures periodic or jittery motion, which will not be quantified by the SPEED feature if the end position is not far from the original position. Also, high JITTER often indicates that the tracker has lost the hand position, which would be an excellent reason to ignore the gesture features.

As noted in the previous section, an HMM was used to perform a spatio-temporal clustering on the hand positions and velocities. The REST-CLUSTER feature takes the value "true" iff the most frequently occupied state during the NP is the closest to rest position. In addition, parameter tying is used in our HMM to ensure that all states but one are static holds, and this remaining state represents the transition movements between those holds. Only this last state is permitted to have an expected non-zero speed; if the hand is most frequently in this state during the NP, then the MOVEMENT-CLUSTER feature takes the value "true."

6. Evaluation on Coreference Resolution

In this evaluation, we assess whether gesture features improve coreference resolution, and we compare conditional modality fusion to other approaches for gesture-speech combination.

6.1 Evaluation Setup

We describe our procedure for evaluating the performance of our approach. This includes the evaluation metric (Section 6.1.1), baselines for comparison (Section 6.1.2), and parameter tuning (Section 6.1.3). The coreference annotations were described in Section 3.2.

6.1.1 Evaluation Metric

Coreference resolution is often performed in two phases: a binary classification phase, in which the likelihood of coreference for each pair of noun phrases is assessed; and a global partitioning phase, in which the clusters of mutually-coreferring NPs are formed. Our model does not address the global partitioning phase, only the question of whether each pair of noun phrases in the document corefer. Moving from pairwise noun phrase coreference to global partitioning requires a clustering step that may obscure performance differences on the level at which our model operates. Moreover, these results will depend on the choice of the clustering algorithm and the mechanism for selecting the number of clusters (or, alternatively, the cut-off value on merging clusters). This parameterization is particularly challenging for our corpus because we do not have a large dedicated development set. Consequently, the bulk of our evaluation is performed on the binary classification phase. However, for the purpose of comparing with prior work on coreference, we also perform a global evaluation, which measures the overall results after clustering.

For the binary evaluation, we use the area under the ROC curve (AUC) as an error metric (Bradley, 1997). AUC evaluates classifier performance without requiring the specification of a cutoff. This metric penalizes misorderings – cases in which the classifier ranks negative examples more highly than positive examples. Such ROC analysis is increasingly popular, and has been used in a variety of NLP tasks, including the detection of action items in emails (Bennett & Carbonell, 2007) and topic segmentation (Malioutov & Barzilay, 2006). Although binary evaluation is not typically used for coreference resolution, we believe it is an appropriate choice here, for the reasons noted above.

The global evaluation uses the constrained entity-alignment f-measure (CEAF) for evaluation (Luo, 2005). This metric avoids well-known problems with the earlier MUC evaluation metric (Vilain, Burger, Aberdeen, Connolly, & Hirschman, 1995). The clustering step is performed using two standard techniques from the literature, which we describe in Section 6.3. In future work we plan to explore techniques that perform coreference in a single joint step (e.g., Daumé III & Marcu, 2005). Then a global metric would be more appropriate to measure the contributions of our model directly.

6.1.2 Baselines

Conditional modality fusion (CMF) is compared with traditional approaches to modality combination for NLP tasks:

- **Early fusion**. The early fusion baseline includes all features in a single vector, ignoring modality. This is equivalent to standard maximum-entropy classification. Early fusion is implemented with a conditionally-trained log-linear classifier; it uses the same code as the CMF model, but always includes all features.

- **Late fusion**. The two late fusion baselines train separate classifiers for gesture and speech, then combine their posteriors. The modality-specific classifiers are conditionally-trained log-linear classifiers, and again use the same code as the CMF model. For simplicity, a parameter sweep identifies the interpolation weights that maximize performance on the test set. Thus, it is likely that these results somewhat overestimate the performance of these baseline models. Both additive and multiplicative combination are considered.

- **No fusion**. These baselines include the features from only a single modality, and again build a conditionally-trained log-linear classifier. Implementation uses the same code as the CMF model, but weights on features outside the target modality are forced to zero.

An important question is how our results compare with existing state-of-the-art coreference systems. The "no fusion, verbal features only" baseline provides a reasonable representation of prior work on coreference, by applying a maximum-entropy classifier to a set of typical textual features. A direct comparison with existing implemented systems would be ideal, but all such available systems use textual features that are inapplicable to our spoken-language dataset, such as punctuation, capitalization, and gazetteers.

6.1.3 Parameter Tuning

As the small size of the corpus did not permit dedicated test and training sets, results are computed using leave-one-out cross-validation, with one fold for each of the sixteen documents in the corpus. Parameter tuning was performed using cross validation within each training fold. This includes the selection of the regularization constant, which controls the trade-off between fitting the training data and learning a model that is simpler (and thus, potentially more general). In addition, binning of continuous features was performed within each cross-validation fold, using the method described in Section 5.2. Finally, as noted above, model weights are initialized to zero, enabling deterministic reproducibility of the experiments.

6.2 Results

Conditional modality fusion outperforms all other approaches by a statistically significant margin (Table 4). Compared with early fusion, the different-zero model for conditional modality fusion offers an absolute improvement of 1.17% in area under the ROC curve (AUC) – compare lines 1 and 4 in the table. A paired t-test shows that this result is statistically significant ($p < .01, t(15) = 3.73$). CMF obtains higher performance on fourteen of the sixteen cross-validation folds. Both additive and multiplicative late fusion perform on par with early fusion. The p-values of the significance tests for of all pairwise comparisons are shown in Table 5.

Early fusion with gesture features is superior to unimodal verbal classification by an absolute improvement of 1.64% AUC ($p < .01, t(15) = 4.45$) – compare lines 4 and 7 in Table 4. The additional 1.17% AUC provided by conditional modality fusion amounts to a relative 73% increase in the power of the gesture features. The results are relatively robust to variations in the regularization constant, as shown in Figure 3. This means that the

CMF different-different (DD)	Uses two different sets of weights for both verbal and gestural features, depending on the hidden variable (equation 6).
CMF different-zero (DZ)	Uses different weights on the verbal features depending on the hidden variable; if the hidden variable indicates non-salience, gesture weights are set to zero (equation 5).
CMF same-zero (SZ)	Uses the same weights on verbal features regardless of gesture salience; if the hidden variable indicates non-salience, gesture weights are set to zero (equation 4).
Early fusion (E)	Standard log-linear classifier. Uses the same weights on verbal and gestural features, regardless of hidden variable (equation 3).
Late fusion, multiplicative (LM)	Trains separate log-linear classifiers for gesture and verbal features. Combines posteriors through multiplication.
Late fusion, additive (LA)	Trains separate log-linear classifiers for gesture and verbal features. Combines posteriors through interpolation.
No fusion (VO, GO)	Uses only one modality for classification.

Table 3: Summary of systems compared in this evaluation

model	AUC
1. CMF different-zero	**.8226**
2. CMF different-different	.8174
3. CMF same-zero	.8084
4. Early fusion (same-same)	.8109
5. Late fusion, multiplicative	.8103
6. Late fusion, additive	.8068
7. No fusion (verbal features only)	.7945
8. No fusion (gesture features only)	.6732

Table 4: Coreference performance, in area under the ROC curve (AUC). The systems are described in Table 3

	DD	SZ	E	LM	LA	VO	GO
CMF different-zero (DZ)	.01	.01	.01	.01	.01	.01	.01
CMF different-different (DD)		.05	ns	ns	.05	.01	.01
CMF same-zero (SZ)			ns	ns	ns	.05	.01
Early fusion (E)				ns	ns	.01	.01
Late fusion, multiplicative (LM)					ns	.01	.01
Late fusion, additive (LA)						.01	.01
Verbal features only (VO)							.01
Gesture features only (GO)							

Table 5: P-values of the pairwise comparison between models. "ns" indicates that the difference in model performance is not significant at $p < .05$. The parentheses in the left column explain the abbreviations in the top line.

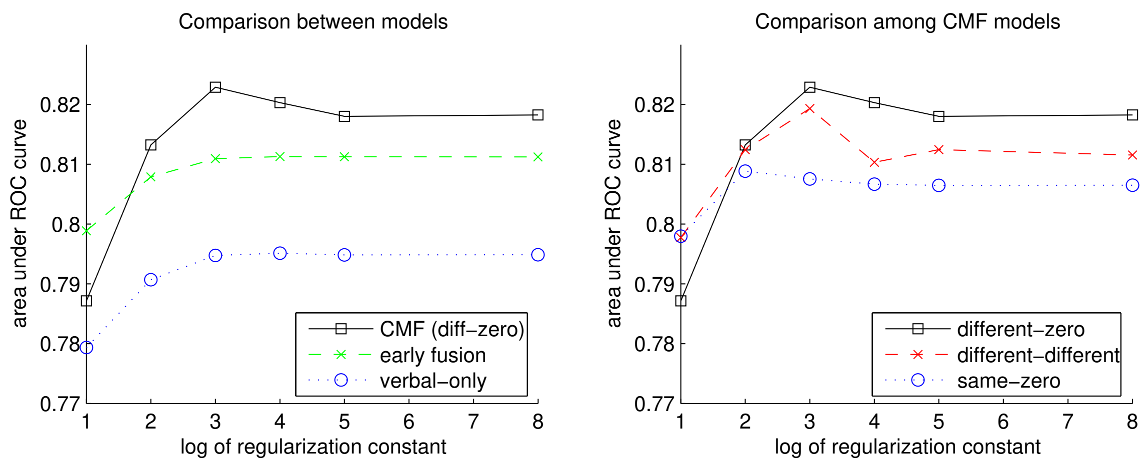


Figure 3: Results with regularization constant

performance gains obtained by conditional modality fusion are not highly dependent on finding the optimal regularization constant.

As noted in Section 4.3, conditional modality fusion optimizes a non-convex objective. We perform an additional evaluation to determine whether performance is sensitive to initialization. Randomizing the weights over five different iterations with our best-performing system, we observed a standard deviation of $1.09 * 10^{-3}$ in area under the ROC curve (AUC). In all other experiments the weights were initialized to zero, enabling the results to be reproduced deterministically.

6.3 Global Metric

Coreference is traditionally evaluated with a global error metric. Our research is directed specifically at the binary classification of coreference between pairs of noun phrases, so we

model	first-antecedent	best-antecedent
CMF (different-zero)	55.67	56.02
CMF (different-different)	54.71	56.20
CMF (same-zero)	53.91	55.32
Early fusion (same-same)	54.18	55.50
Late fusion, multiplicative	53.74	54.44
Late fusion, additive	53.56	55.94
No fusion (verbal features only)	53.47	55.15
No fusion (gesture features only)	44.68	44.85

Table 6: CEAF global evaluation scores, using best clustering threshold

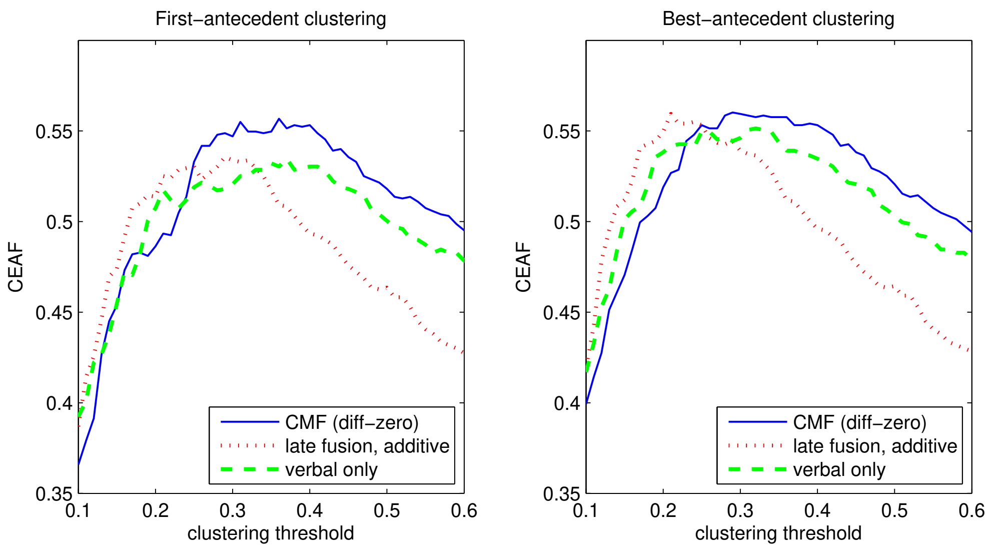


Figure 4: Global coreference performance, measured using CEAF scores, plotted against the threshold on clustering

have focused on evaluating that specific portion of the larger coreference problem. However, for the purpose of comparing with prior research on coreference, we present results using a more traditional global metric.

To perform a global evaluation, we must cluster the noun phrases in the document, using the pairwise coreference likelihoods as a similarity metric. We experiment with two clustering methods that are commonly used in the literature. The **first-antecedent** technique resolves NPs to the first antecedent whose similarity is above a predefined threshold (Soon et al., 2001). The **best-antecedent** technique resolves each noun phrase to the most compatible prior noun phrase, unless none is above the threshold (Ng & Cardie, 2002).

Figure 4 shows the global scores, plotted against the value of the clustering threshold. For clarity, only the best performing system from each class is shown: for conditional modality fusion, we plot the different-zero model; from the multimodal baselines, we plot the additive late fusion model (the combination of additive late fusion and the best-antecedent clustering method is the best performing multimodal baseline); from the unimodal baseline, we plot the verbal-features only baseline. Table 6 lists the performance of each method at its optimum clustering threshold. For comparison, Ng reports a CEAF score of 62.3 (Ng, 2007) on the ACE dataset, although the results are not directly comparable due to the differences in corpora.

As shown in these results, performance is sensitive to both the clustering method and the clustering threshold. Conditional modality fusion generally achieves the best results, and best-antecedent clustering generally outperforms the first-antecedent technique. Nonetheless, the advantage of conditional modality fusion is smaller here than with ROC analysis. We believe that ROC analysis demonstrates the advantage of conditional modality fusion more directly, while the global metric interposes a clustering step that obscures differences between the classification techniques. Nonetheless, the global metric may be a better overall measure of the quality of coreference for downstream applications such as search or summarization. In future work, we hope to investigate whether the conditional modality fusion approach can be applied to global models of coreference that do not require separate classification and clustering phases (e.g., Daumé III & Marcu, 2005).

6.4 Feature Analysis

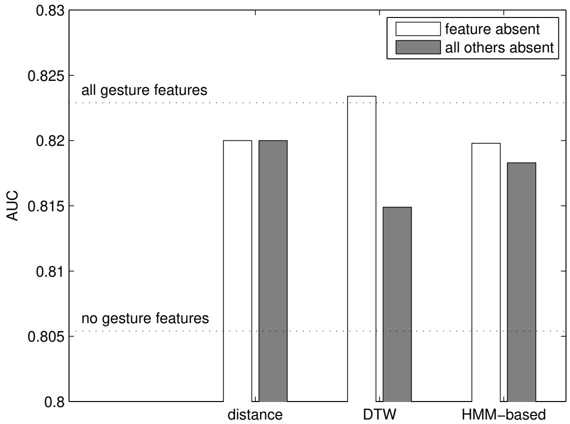
The machine learning approach that we have adopted permits a novel analysis in which we compare the linguistic contribution of our gesture features in the presence of other verbal features. Thus we can investigate which gesture features supply unique information over and above the verbal features. In addition, we analyze which types of verbal features correlate closely with gesture features, and which are independent. All statistical significance results are based on two-tailed, paired t-tests.

6.4.1 Gestural Similarity

Figure 5 shows the contribution of three classes of gestural similarity features: FOCUS-DISTANCE, DTW-AGREEMENT, and the two HMM-based features (SAME-CLUSTER and JS-DIV). The top dotted line in the graph shows performance of the different-zero model with the complete feature set, and the bottom line shows performance of this model without any gestural similarity features.[15]

Each feature group conveys useful information, as performance with any one feature group is always better than performance without gestural similarity features ($p < .01, t(15) = 3.86$ for DTW-AGREEMENT, the weakest of the three feature groups). The performance using only the FOCUS-DISTANCE is significantly better than when only the DTW-AGREEMENT feature is used ($p < .05, t(15) = 2.44$); other comparisons are not significant. We also find

15. Note that the baseline of “no gesture features” is higher than the “no fusion (verbal features only)” baseline from Table 4. Although the feature groups here are identical, the classifiers are different. The “no fusion (verbal features only)” baseline uses a standard log-linear classifier, while “no gesture features” uses conditional modality fusion, permitting two sets of weights for the verbal features, as shown in equation 5.



feature group	+	-
all gesture similarity features	.8226	.8054
FOCUS-DISTANCE	.8200	.8200
DTW-AGREEMENT	.8149	.8234
HMM-based	.8183	.8198

Figure 5: An analysis of the contributions of each set of gestural similarity features. The "plus" column on the left of the table shows results when only that feature set was present; the "minus" column shows results when it was removed. As before, the metric is area under the ROC curve (AUC).

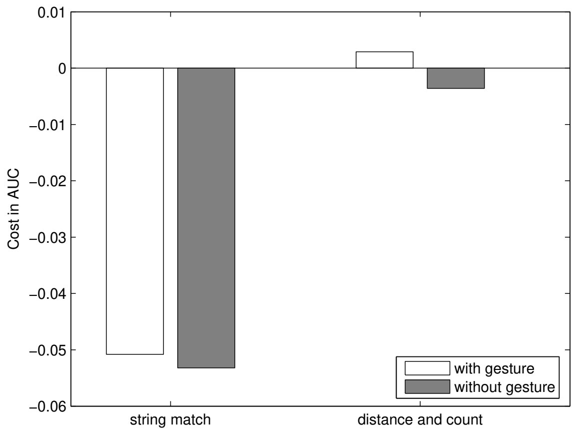
evidence of redundancy between the feature groups, as removing any individual feature group does not significantly impair performance if the other two feature groups remain.

6.4.2 Verbal and Gestural Overlap

Next, we assess the degree of overlap between gesture and verbal information. We hypothesize that gesture is complementary with certain verbal features, and redundant with others. For example, string match features such as EDIT-DISTANCE and EXACT-MATCH seem unlikely to convey the same information as gesture. To see why, consider the cases in which string match is likely to be helpful: fully-specified noun phrases such as "the red ball," rather than pronouns. Empirical research suggests that the majority of informative gestures occur during pronouns and other underspecified utterances, where string match is unlikely to be helpful (Kehler, 2000). Thus, we expect a low level of overlap between gesture and string match features.

Distributional features are another source of verbal information. They include the number of intervening sentences or noun phrases between the two candidate NPs, and the number of times each NP appears in the document. These features establish the context that may permit the resolution of references that are ambiguous by their surface form alone. For example, if a noun phrase occurred very recently, or very often, a pronominal reference may be sufficiently clear. Since gesture may also be used in such cases, we expect some redundancy between gestural similarity and distributional features.

These intuitions lead us to specific predictions about system performance. The presence of the gesture similarity features should mitigate the cost of removing the distributional features, if the gesture features reinforce some of the same information. However, the presence of the gesture features should have no effect on the cost of removing the string match features.



feature group	with gesture similarity features	without
no string match	.7721	.7522
no distance, count	.8258	.8018

Figure 6: The contribution of verbal features, with and without gesture similarity features. The graph shows the loss incurred by removing each verbal feature class, conditioned on the presence of gesture similarity features. The table shows the overall performance with each combination of feature groups.

These predictions are supported by the data (Figure 6). Removing the distributional features does not impair performance as long as the gesture features are present, but does impair performance if the gesture features are also removed – this difference is statistically significant ($p < .01, t(15) = 3.76$). This observation is consistent with our hypothesis that these feature groups convey similar information. In contrast, the cost of removing the string match features does not vary by a significant margin, regardless of whether the gesture features are present. This accords with the intuition that these feature groups convey independent information.

7. Evaluation on Keyframe Extraction

The previous sections describe an application of conditional modality fusion to natural language processing: by using gesture features only when they are meaningful, their contribution to coreference classification is enhanced. In this section, we show that conditional modality fusion also predicts which gestures are useful for human viewers. Specifically, we use the conditional modality fusion estimate of gesture salience to select keyframes from a video. We demonstrate that the keyframes selected by this method match those selected by human raters better than the keyframes selected by traditional text and image-based algorithms.

In Section 7.1, we explain the keyframe-based summarization task. We describe our basic modeling approach in Section 7.2. The evaluation setup is presented in Section 7.3. Section 7.4 gives the experimental results.

7.1 Why Keyframe-Based Summarization

Our goal is to produce a "comic book" summary of a video, in which a transcript is augmented with salient *keyframes* – still images that clarify the accompanying text. Keyframe-based summaries allow viewers to quickly review key points of a video presentation, without requiring the time and hardware necessary to view the actual video (Boreczky, Girgensohn, Golovchinsky, & Uchihashi, 2000). As we have argued above, textual transcriptions alone do not capture all relevant information, and a keyframe-based summary may provide the minimal visual information required to understand such a presentation. Appendix B contains an excerpt from a summary produced by our system.

As noted, gesture supplements speech with unique semantic information. Thus, keyframes showing salient gestures would be a valuable addition to the transcript text. Ideally, we would select keyframes that avoid redundancy between the visual and verbal modalities, while conveying all relevant information.

Existing techniques for keyframe extraction have usually focused on edited videos such as news broadcasts (e.g., Uchihashi, Foote, Girgensohn, & Boreczky, 1999; Boreczky et al., 2000; Zhu, Fan, Elmagarmid, & Wu, 2003). Such systems seek to detect large-scale changes in image features to identify different scenes, and then choose a representative example from each scene. This approach is poorly suited to unedited videos, such as a recording of a classroom lecture or business presentation. In such videos, the key visual information is not the variation in scenes or camera angles, but the visual communication provided by the gestures of the speaker. Our goal is to capture relevant keyframes by identifying salient gestures, using the model developed in the previous sections of this paper.

7.2 Modeling Approach

One possible approach is to formulate gesture extraction as a standard supervised learning task, using a corpus in which salient gestures are annotated. However, such annotation is expensive, and we prefer to avoid it. Instead we learn salience by bootstrapping from multimodal coreference resolution, using conditional modality fusion. By learning to predict the specific instances in which gesture helps, we can obtain a model of gesture salience. For example, we expect that a pointing gesture in the presence of an anaphoric expression would be found to be highly salient (as in Figure 1); a more ambiguous hand pose in the presence of a fully-specified noun phrase would not be salient. This approach will not identify *all* salient gestures, but will identify those that occur in the context of the selected language understanding task. In coreference resolution, only gestures that co-occur with noun phrases can be selected. As noun phrases are ubiquitous in language, this should still cover a usefully broad collection of gestures.

Using the model for coreference resolution introduced in Section 4, we obtain the probability distribution for the hidden variable, which controls whether the gesture features are included for coreference resolution. Our basic hypothesis is that instances in which gesture features are included with high likelihood are likely to correspond to salient gestures. The gestures rated salient by this method are used to select keyframes in the summary.

Models of coreference resolution and gesture salience are learned jointly, based on the "same-zero" model defined in Equation 4. After training, a set of weights $\mathbf{w}_h$ is obtained, allowing the estimation of gesture salience at each noun phrase. We sum over all possible values for y and h_2, obtaining $\sum_{y,h_2} \psi(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) = h_1 \mathbf{w}_h^T \mathbf{x}_{h_1}$. We find the potential for the case when the gesture is salient by setting $h_1 = 1$, yielding $\mathbf{w}_h^T \mathbf{x}_{h_1}$.[16] Our working assumption is that this potential is a reasonable proxy for the informativeness of a keyframe that displays the noun phrase's accompanying gesture.

The potential provides an ordering on all noun phrases in the document. We select keyframes from the midpoints of the top n noun phrases, where n is specified in advance by the annotator. Providing the system with the ground truth number of keyframes follows common practice from the textual summarization literature – summaries of different lengths are difficult to compare, as the summary duration is governed partially by the annotator's preference for brevity or completeness (Mani & Maybury, 1999). Each keyframe is given a caption that includes the relevant noun phrase and accompanying text, up to the noun phrase in the next keyframe. Portions of the output of the system are shown in Figure 1 and Appendix B.

7.3 Evaluation Setup

Our evaluation methodology is similar to the intrinsic evaluation developed for the Document Understanding Conference.[17] We assess the quality of the automatically extracted keyframes by comparing them to human-annotated ground truth.

7.3.1 Dataset

The dataset again consists of dialogues collected using the procedure described in Section 3. The same coreference annotations described in Section 3.2 are used. Additionally, nine of the sixteen videos are annotated for keyframes. Of these, three are used in developing our system and the baselines, and the remaining six are used for final evaluation (these are indicated by asterisks in the table in Appendix C). There is no explicit training on the keyframe annotations, but the development set was used for evaluation as the system was under construction.

The specification of the ground truth annotation required that the keyframes capture all static visual information that the annotator deems crucial to understanding the content of the video. The number of selected frames was left to their discretion; on average, 17.8 keyframes were selected per document, out of an average total of 4296 frames per document. Annotation was performed by two raters; on a subset of two videos annotated by both raters, the raw interrater agreement was 86%, yielding a kappa of .52.

One important difference between our dataset and standard sentence extraction datasets is that many frames may be nearly identical, due to the high frame rate of video. For this reason, rather than annotating individual frames, the annotators marked *regions* with identical visual information. These regions define equivalence classes, such that any frame from a given region would be equally acceptable. If a single keyframe were selected from every ground truth region, the result would be the minimal set of keyframes necessary for a reader to fully understand the discourse. On average, the 17.8 regions selected per document spanned 568 frames.

16. Note that if we consider the same noun phrase as the anaphor (x_{h_2}) and sum over all possible values of h_1, the resulting potential is identical.
17. http://duc.nist.gov

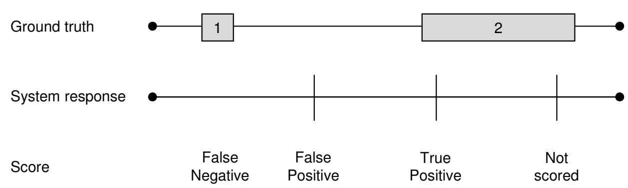


Figure 7: An example of the scoring setup.

7.3.2 Training Coreference Resolution

As described in Section 7.2, our approach to keyframe extraction is based on a model for gesture salience that is learned from labeled data on coreference resolution. The training phase is performed as leave-one-out cross-validation: a separate set of weights is learned for each presentation, using the other fifteen presentations as a training set. The learned weights are used to obtain the values of the hidden variable indicating gesture salience, as described in the previous subsection.

7.3.3 Evaluation Metric

Figure 7 illustrates the scoring setup. The top row in the figure represents the ground truth; the middle row represents the system response, with vertical lines indicating selected keyframes; the bottom row shows how the response is scored.

For all systems the number of keyframes is fixed to be equal to the number of regions in the ground truth annotation. If the system response includes a keyframe that is not within any ground truth region, a false positive is recorded. If the system response fails to include a keyframe from a region in the ground truth, a false negative is recorded; a true positive is recorded for the first frame that is selected from a given ground truth region, but additional frames from the same region are not scored. The system is thus still penalized for each redundant keyframe, because it has "wasted" one of a finite number of keyframes it is allowed to select. At the same time, such an error seems less grave than a true substitution error, in which a keyframe not containing relevant visual information is selected. We report the F-measure, which is the harmonic mean of recall and precision.

7.3.4 Baselines

We compare the performance of our system against three baselines, which we present in order of increasing competitiveness.

- **Random-keyframe:** Our simplest baseline selects n keyframes at random from throughout the document. This baseline is similar to the "random sentence" baselines that are common in the textual summarization literature (Mani & Maybury, 1999). The number of keyframes selected in this baseline is equal to the number of regions in the ground truth. This baseline expresses a lower bound on the performance that any reasonable system should achieve on this task. Our results report the average of 500 independent runs.

- **NP-salience:** The NP-SALIENCE system is based on frequency-based approaches to identifying salient NPs for the purpose of text summarization (Mani & Maybury, 1999). The salience heuristic prefers the most common representative tokens of the largest and most homogeneous coreference clusters.[18] The largest cluster is the one containing the most noun phrases; homogeneity is measured by the inverse of the number of unique surface forms. This provides a total ordering on NPs in the document; we select keyframes at the midpoint of the top n noun phrases, where n is the number of keyframe regions in the ground truth. In future work we hope to explore finding the best point *within* each noun phrase for keyframe selection.

- **Pose-clustering:** Our final baseline is based purely on visual features. It employs clustering to find a representative subset of frames with minimum mutual redundancy. Uchihashi et al. (1999), in a seminal paper on keyframe selection, perform clustering on all frames in the video, using the similarity of color histograms as a distance metric. Representative images from each cluster are then used as keyframes. More recent video summarization techniques have advanced the clustering algorithms (Liu & Kender, 2007) and the similarity metric (Zhu et al., 2003), but the basic approach of forming clusters based on visual similarity and choosing exemplar keyframes from these clusters is still used in much of the state-of-the-art research on this topic (see Lew, Sebe, Djeraba, & Jain, 2006, for a survey).

  In our dataset, there is a single fixed camera and no change in the video except for the movements of the speaker; thus, the color histograms are nearly constant throughout. Instead, we use the tracked coordinates of the speaker's hands and upper-body, normalize all values, and use the Euclidean distance metric. In this setting, clusters correspond to typical body poses, and segments correspond to holds in these poses. Following Uchihashi et al. (1999), the video is divided into segments in which cluster membership is constant, and keyframes are taken at the midpoints of segments. We use importance metric from this paper for ranking segments, and choose the top n, where n is the number of keyframes in the ground truth.

7.4 Results

Table 7 compares the performance of our method (GESTURE-SALIENCE) with the three baselines. Using paired t-tests, we find that **Gesture-salience** significantly outperforms all alternatives ($p < .05$ in all cases). The POSE-CLUSTERING and NP-SALIENCE systems are statistically equivalent; both are significantly better than the RANDOM-KEYFRAME baseline ($p < .05$).

The set of baselines against which our system is compared is necessarily incomplete, as there are many ways in which keyframes extraction could be performed. For example, prosodic features could be used to identify moments of particular interest in the dialogue (Sundaram & Chang, 2003). In addition, a combination of baselines including visual and linguistic features may also perform better than any individual baseline. However, developing more complicated baselines is somewhat beside the point. The evaluation demonstrates that a simple yet effective technique for selecting meaningful keyframes is obtained

18. Here, coreference clusters are based on manual annotations.

Model	F-Measure	Recall	Precision
GESTURE-SALIENCE	**.404**	**.383**	**.427**
Pose-clustering	.290	.290	.290
NP-salience	.239	.234	.245
Random-keyframe	.120	.119	.121

Table 7: Comparison of performance on keyframe selection task

as a byproduct of conditional modality fusion. This suggests that the estimates of gesture salience given by our model cohere with human perception.

Error analysis A manual inspection of the system output revealed that in many cases our system selected a noun phrase that was accompanied by a relevant gesture, but the specific keyframe was slightly off. Our method always chooses the keyframe at the midpoint of the accompanying noun phrase; often, the relevant gesture is brief, and does not necessarily overlap with the middle of the noun phrase. Thus, one promising approach to improving results would be to "look inside" each noun phrase, using local gesture features to attempt to identify the specific frame in which the gesture is most salient.

Other errors arise because some key gestures are not related to noun phrases. For example, suppose the speaker says "it shoots the ball up," and accompanies only the word "up" with a gesture indicating the ball's trajectory. This gesture might be important to understanding the speaker's meaning, but since it does not overlap with a noun phrase, the gesture will not be identified by our system. We believe that our results show that focusing on noun phrases is a good start for linguistically-motivated keyframe extraction, and that our unsupervised approach is successful at identifying which noun phrases require keyframes. As gesture is applied to other language tasks, we hope to model the salience of gesture at other phrase types, thus increasing the coverage of our approach to keyframe extraction.

8. Conclusions and Future Work

In summary, this work is motivated by the idea that gestures are best interpreted not as individual units with self-contained meaning, but in the context of other gestural and linguistic information. Previous NLP research on gesture has largely focused on building recognizers for gestures that characterize specific language phenomena: for example, detecting hand gestures that cue sentence boundaries (Chen et al., 2006), or body language that suggests topic shifts (Nakano et al., 2003). Such approaches treat gesture as a sort of visual punctuation. In contrast, we are interested in the semantics that gesture carries. We do not take a recognition-based approach because we believe it unlikely that the space of possible meaningful gestures could be delineated by any training set. Instead, we start from the hypothesis that *patterns* in gesture will correspond to patterns in meaning, so that the degree of similarity between two gestures predicts the semantic similarity of the associated speech. This approach is validated by our experimental results, which show substantial improvement in noun phrase coreference resolution using gesture features. This is one of the first results showing that automatically-extracted visual features significantly improve discourse analysis.

A second key finding is that a structured approach to multimodal integration is crucial to achieving the full benefits offered by gesture features for language understanding. Rather than building separate verbal and gesture interpretation units – or simply concatenating their features – we build a model whose structure encodes the role of each modality. In particular, the gesture modality supplements speech only intermittently, and therefore we represent gesture salience explicitly with a hidden variable. This approach, which we call *conditional modality fusion*, yields a 73% relative improvement in the contribution of the gesture features towards coreference resolution. This improvement is attained by modeling gesture salience with a hidden variable and ignoring gestures that are not salient.

Conditional modality fusion induces an estimate of gesture salience within the context of a specific linguistic task. To test the generality of the salience model, we transfer the derived estimates to a completely different task: keyframe extraction. Without any labeled data on the keyframe task, this simple algorithm outperforms competitive unimodal alternatives. This suggests that the model of gesture salience learned from coreference coheres with human perception of gesture salience.

The theme of generality in gesture salience is suggestive for future research. In principle, a general model of gesture salience could be applied to a range of discourse-related language processing tasks. For example, consider topic segmentation: in text, changes in the distribution of lexical items is a strong indicator of topic boundaries (Hearst, 1994). Assuming that salient gestures carry unique semantic content, changes in the distribution of features of salient gestures could be used in a similar way, supplementing a purely textual analysis.

Moreover, the combination of multiple language processing tasks in a single joint framework raises the possibility that gesture salience could be modeled more robustly as knowledge is transferred between tasks. We have argued against using an explicit universal taxonomy of gesture, favoring an approach focused on the relevance of gesture to specific language processing problems. However, such a joint framework would generalize the notion of salience in a bottom-up, data-driven way. Such research may be relevant from a purely linguistic standpoint: for example, investigating which types of language phenomena share coherent notions of gesture salience, and how gesture salience is expressed in visual and linguistic features. In this paper we have argued that structured models such as conditional modality fusion can be used to incorporate linguistic ideas about gesture in a principled way. We hope that future work will show that such models can also provide a new tool to study the linguistics of gesture.

Another possibility for future work is to investigate richer models of gesture salience. The structure we have explored here is minimal – a binary variable to indicate the salience of a gesture for coreference resolution. We see this as a first step towards more complex structural representations for gesture salience that may yield greater gains in performance. For example, it is likely that gesture salience observes some temporal regularity, suggesting a Markov state model. Other tasks may involve more structured dependencies among gestures, requiring models such as probabilistic context-free grammars.

Finally, we note that hand gesture is one of several modalities that accompany spoken language. Prosody has attracted the greatest amount of attention in natural language processing, but other coverbal modalities include facial expressions, body posture, and – in settings such as lectures or meetings – writing and diagrams. The relationship between these modalities is poorly understood; future research might explore the mapping of linguistic functions to modalities, and whether there are sets of modalities that are redundant or complementary.

Acknowledgments

The authors acknowledge the editor and anonymous reviewers for their helpful comments. We also thank our colleagues Aaron Adler, S. R. K. Branavan, Emma Brunskill, Sonya Cates, Erdong Chen, C. Mario Christoudias, Michael Collins, Pawan Deshpande, Lisa Guttentag, Igor Malioutov, Max Van Kleek, Michael Oltmans, Tom Ouyang, Christina Sauper, Tom Yeh, and Luke Zettlemoyer. The authors acknowledge the support of the National Science Foundation (Barzilay; CAREER grant IIS-0448168 and grant IIS-0415865) and the Microsoft Faculty Fellowship. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation or Microsoft.

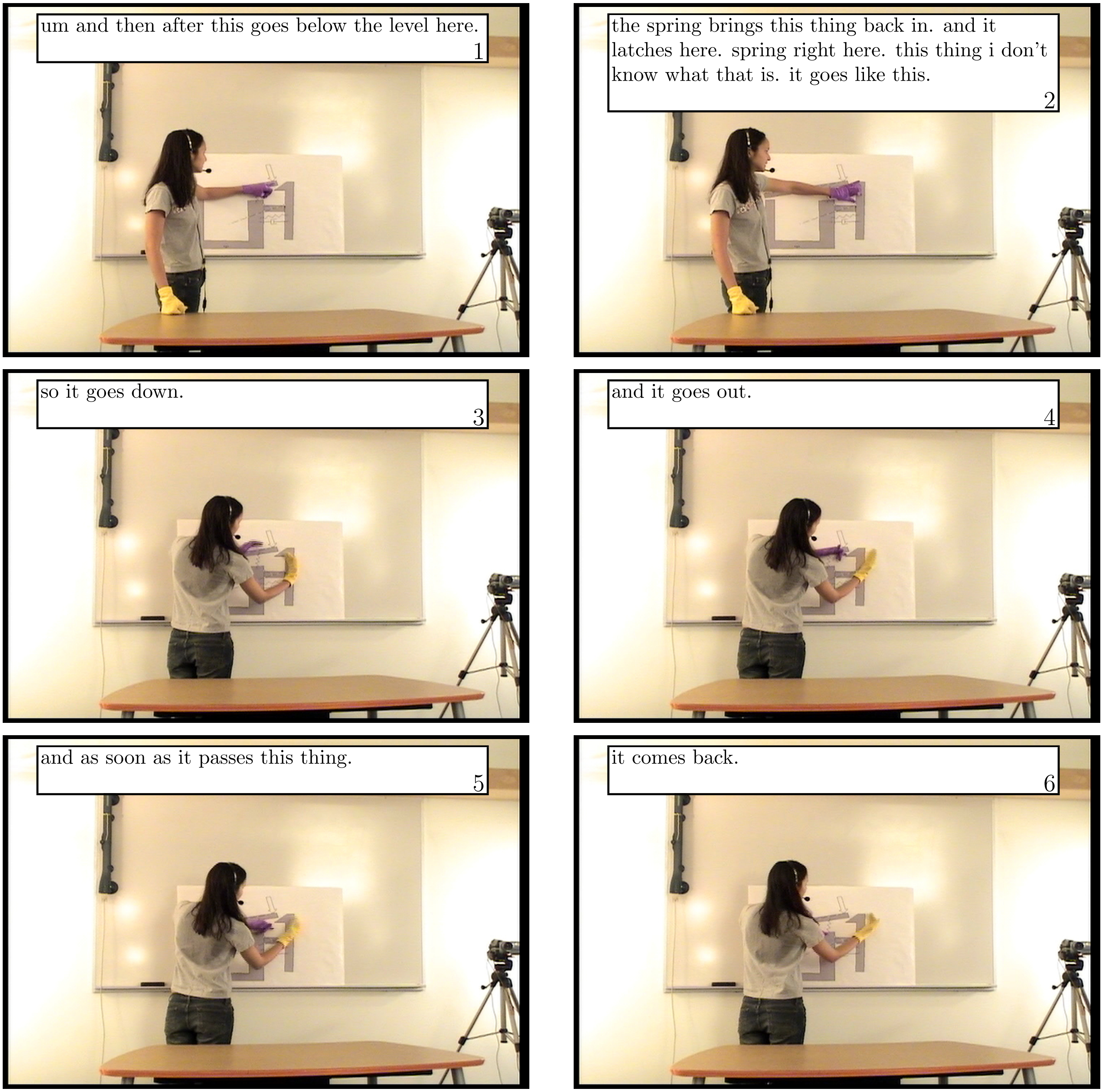
Appendix A. Example Transcript

```
1   ok so [(0) this object right here].
2   i'm going to attempt to explain what [(0) this] does to you.
3   [(1) this ball right here] is the only
4   [(1) this ball]
5   [(2) this spring]
6   [(3) this long arm]
7   and [(4) this] are [(5) the only objects that actually move].
8   no i take [(6) that] back.
9   [(7) this] rotates as well.
10   while [(8) these things] stay fixed.
11   what happens is [(1) the ball] comes down [(9) here].
12   and [(2) this spring] is active.
13   meaning that [(2) it's] going up and down.
14   because [(4) this] will come up.
15   jostle [(3) that].
16   and then go around.
17   so [(3) it'll] as [(4) this] raises [(3) it] up.
18   [(10) this hand] goes down.
19   and then [(10) it'll] spring back up.
20   [(1) the ball] typically goes up [(11) here].
21   bounces off [(12) here].
22   gets caught in like [(13) a groove].
23   [(7) this] is continually moving around in [(14) a circle]
24   then [(15) this] happened three times
25   i watched [(16) a video] and [(15) it] happened [(17) three times]
26   [(1) the ball] never went through [(18) there] or [(19) over here]
27   [(1) it] always would get down back to [(20) here]
28   and then down through [(9) here]
29   sometimes [(21) this thing] would hit [(1) it] harder
30   and [(1) it] would go higher up
31   and sometimes [(1) it] would just kind of loop over
32   no no [(1) it] only came down through [(9) here]
33   i have no idea why there's [(22) anchors] on [(23) here]
34   [(24) that] wasn't really made clear to me
35   and yeah [(25) that's] pretty much [(26) it]
36   [(1) it's] essentially [(1) a bouncy ball]
37   but [(1) it] just pretty much drops like [(27) dead weight]
38   when [(1) it] hits [(28) something]
39   and that was [(26) it]
40   [(16) it] was probably like [(16) a forty five second video] at most
41   and [(29) it] happened [(17) three times] in [(16) that video]
42   so [(16) it] moves relatively quickly
43   not so much lodged as just like [(1) it] would like come down [(13) here]
```

44 and as [(7) this] is moving [(7) it] would
45 just kind of like dump [(1) it] into [(20) here]
46 [(7) it's] more of something that's in [(30) the way]
47 than actually [(31) a transfer]
48 because if [(7) this] wasn't [(32) here]
49 [(1) it] would still fall down [(20) here] and then get in [(9) here]
50 that's it
51 i'm not actually sure what [(0) this] does
52 [(0) it] looks like [(0) it] looks just like
53 [(0) this] on [(33) the computer screen]
54 so so um [(0) it] basically looks like
55 [(34) kind of a so so game of pinball]
56 [(35) that] was [(36) my understanding of it]
57 i'm not sure what else [(37) it's] supposed to do
58 ok we're done guys with [(37) this one]

Appendix B. Example Keyframe Summary

This is an excerpt of a keyframe summary that was generated automatically, as described in Section 7.



Appendix C. Dataset Parameters

number	speaker	topic	gender	duration	# words	# NPs
1	1	pinball	F	3:00	455	95
2	1	candy dispenser	F	2:27	428	101
3*	2	latch	F	1:19	104	27
4	2	pinball	F	2:31	283	65
5*	3	pinball	F	3:00	325	69
6*	4	candy dispenser	M	3:00	404	100
7	4	pinball	M	3:00	421	109
8	5	pinball	M	3:00	362	89
9*	5	piston	M	3:00	313	69
10*	6	pinball	M	3:00	315	71
11	7	latchbox	M	2:20	347	72
12	7	pinball	M	3:00	221	51
13	8	pinball	F	2:23	192	47
14*	8	piston	F	0:47	48	8
15	9	pinball	M	2:30	378	87
16	9	candy dispenser	M	2:43	358	77
total				41:00	4954	1137

Corpus statistics for the dataset used in our experiments. Asterisks indicate videos that were used in the keyframe evaluation.

References

Adler, A., Eisenstein, J., Oltmans, M., Guttentag, L., & Davis, R. (2004). Building the design studio of the future. In *Proceedings of AAAI Workshop on Making Pen-Based Interaction Intelligent and Natural*, pp. 1–7.

Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., et al. (1995). The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *7*(1), 7–48.

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: an entity-based approach. In *Proceedings of the ACL*, pp. 141–148.

Bennett, P., & Carbonell, J. (2007). Combining Probability-Based Rankers for Action-Item Detection. In *Proceedings of HLT-NAACL*, pp. 324–331.

Biber, D. (1988). *Variation Across Speech and Language*. Cambridge University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blitzer, J., McDonald, R., & Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of EMNLP*, pp. 120–128.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*, pp. 92–100.

Boreczky, J., Girgensohn, A., Golovchinsky, G., & Uchihashi, S. (2000). An interactive comic book presentation for exploring video. In *Proceedings of CHI*, pp. 185–192.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the ACL*, pp. 155–162.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2005). The AMI meeting a corpus: a pre-announcement. In *Proceedings of the Workshop on Machine Learning for Multimodal Interaction*, pp. 28–39.

Chai, J. Y., Hong, P., Zhou, M. X., & Prasov, Z. (2004). Optimization in multimodal interpretation. In *Proceedings of the ACL*, pp. 1–8.

Chai, J. Y., & Qu, S. (2005). A salience driven approach to robust input interpretation in multimodal conversational systems. In *Proceedings of HLT-EMNLP*, pp. 217–224.

Chelba, C., & Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, *20*(4), 382–399.

Chen, L., Harper, M., & Huang, Z. (2006). Using maximum entropy (ME) model to incorporate gesture cues for sentence segmentation. In *Proceedings of ICMI*, pp. 185–192.

Chen, L., Liu, Y., Harper, M. P., & Shriberg, E. (2004). Multimodal model integration for sentence unit detection. In *Proceedings of ICMI*, pp. 121–128.

Chen, L., Rose, R. T., Parrill, F., Han, X., Tu, J., Huang, Z., Harper, M., Quek, F., McNeill, D., Tuttle, R., & Huang, T. (2005). VACE multimodal meeting corpus. In *Proceedings of the Workshop on Machine Learning for Multimodal Interaction*, pp. 40–51.

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., & Clow, J. (1997). Quickset: Multimodal interaction for distributed applications. In *Proceedings of ACM Multimedia*, pp. 31–40.

Darrell, T., & Pentland, A. (1993). Space-time gestures. In *Proceedings of CVPR*, pp. 335–340.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the ACL*, pp. 256–263.

Daumé III, H., & Marcu, D. (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of HLT-EMNLP*, pp. 97–104.

Deutscher, J., Blake, A., & Reid, I. (2000). Articulated body motion capture by annealed particle filtering. In *Proceedings of CVPR*, Vol. 2, pp. 126–133.

Eisenstein, J., Barzilay, R., & Davis, R. (2007). Turning lectures into comic books with linguistically salient gestures. In *Proceedings of AAAI*, pp. 877–882.

Eisenstein, J., & Davis, R. (2006). Gesture improves coreference resolution. In *Proceedings of HLT-NAACL, Companion Volume: Short Papers*, pp. 37–40.

Eisenstein, J., & Davis, R. (2007). Conditional modality fusion for coreference resolution. In *Proceedings of the ACL*, pp. 352–359.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of IJCAI*, Vol. 2, pp. 1022–1027.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing (Vol. 1)*, pp. 517–520.

Goodwin, M., & Goodwin, C. (1986). Gesture and co-participation in the activity of searching for a word. *Semiotica*, *62*, 51–75.

Grosz, B., Joshi, A. K., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.

Haghighi, A., & Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of the ACL*, pp. 848–855.

Harabagiu, S. M., Bunescu, R. C., & Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of NAACL*, pp. 1–8.

Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of the ACL*.

Hirschman, L., & Chinchor, N. (1998). MUC-7 coreference task definition. In *Proceedings of the Message Understanding Conference*.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing*. Prentice Hall.

Huang, X., Alleva, F., Hwang, M.-Y., & Rosenfeld, R. (1993). An overview of the Sphinx-II speech recognition system. In *Proceedings of ARPA Human Language Technology Workshop*, pp. 81–86.

Ji, H., Westbrook, D., & Grishman, R. (2005). Using semantic relations to refine coreference decisions. In *Proceedings of HLT-EMNLP*, pp. 17–24.

Johnston, M., & Bangalore, S. (2000). Finite-state multimodal parsing and understanding,. In *Proceedings of COLING-2000*, pp. 369–375.

Jordan, P., & Walker, M. (2005). Learning Content Selection Rules for Generating Object Descriptions in Dialogue. *Journal of Artificial Intelligence Research*, *24*, 157–194.

Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. (2005). Effective use of prosody in parsing conversational speech. In *Proceedings of HLT-EMNLP*, pp. 233–240.

Kameyama, M. (1986). A property-sharing constraint in Centering. In *Proceedings of the ACL*, pp. 200–206.

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI*, pp. 685–690.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of the utterance. In Key, M. R. (Ed.), *The relation between verbal and non-verbal communication*, pp. 207–227. Mouton.

Kettebekov, S., Yeasin, M., & Sharma, R. (2005). Prosody based audiovisual coanalysis for coverbal gesture recognition. *IEEE Transactions on Multimedia*, *7*(2), 234–242.

Kibble, R., & Power, R. (2004). Optimising referential coherence in text generation. *Computational Linguistics*, *30*(4), 401–416.

Kim, J., Schwarm, S. E., & Osterdorf, M. (2004). Detecting structural metadata with decision trees and transformation-based learning. In *Proceedings of HLT-NAACL'04*.

Koo, T., & Collins, M. (2005). Hidden-variable models for discriminative reranking. In *Proceedings of HLT-EMNLP*, pp. 507–514.

Kopp, S., Tepper, P., Ferriman, K., & Cassell, J. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In Nishida, T. (Ed.), *Conversational Informatics: An Engineering Approach*. Wiley.

Kudo, T., & Matsumoto, Y. (2001). Chunking with support vector machines. In *Proceedings of NAACL*, pp. 1–8.

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, *20*(4), 535–561.

Lascarides, A., & Stone, M. (2006). Formal semantics for iconic gesture. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL)*, pp. 64–71.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications and Applications*, *2*(1), 1–19.

Li, X., & Roth, D. (2001). Exploring evidence for shallow parsing. In *Proceedings of CoNLL*, pp. 1–7.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145–151.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*, 503–528.

Liu, T., & Kender, J. R. (2007). Computational approaches to temporal sampling of video sequences. *ACM Transactions on Multimedia Computing, Communications and Applications*, *3*(2), 7.

Liu, Y. (2004). *Structural Event Detection for Rich Transcription of Speech*. Ph.D. thesis, Purdue University.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP*, pp. 25–32.

Malioutov, I., & Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of the ACL*, pp. 25–32.

Mani, I., & Maybury, M. T. (Eds.). (1999). *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, USA.

McCallum, A., & Wellner, B. (2004). Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of NIPS*, pp. 905–912.

McNeill, D. (1992). *Hand and Mind*. The University of Chicago Press.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and communicative intention of the speaker. *Gesture*, *4*(2), 119–141.

Müller, C., Rapp, S., & Strube, M. (2002). Applying Co-Training to reference resolution. In *Proceedings of the ACL*, pp. 352–359.

Müller, C. (2007). Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the ACL*, pp. 816–823.

Nakano, Y., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of the ACL*, pp. 553–561.

Ng, V. (2007). Shallow semantics for coreference resolution. In *Proceedings of IJCAI*, pp. 1689–1694.

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the ACL*, pp. 104–111.

NIST (2003). The Rich Transcription Fall 2003 (RT-03F) Evaluation plan..

Passonneau, R. J. (1997). Applying reliability metrics to co-reference annotation. Tech. rep. CUCS-017-97, Columbia University.

Poddar, I., Sethi, Y., Ozyildiz, E., & Sharma, R. (1998). Toward natural gesture/speech HCI: A case study of weather narration. In *Proceedings of Perceptual User Interfaces*, pp. 1–6.

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004). Centering: a parametric theory and its instantiations. *Computational Linguistics*, *30*(3), 309–363.

Quattoni, A., Collins, M., & Darrell, T. (2004). Conditional random fields for object recognition. In *Proceedings of NIPS*, pp. 1097–1104.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., McCullough, K. E., & Ansari, R. (2002a). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction*, *9:3*, 171–193.

Quek, F., McNeill, D., Bryll, R., & Harper, M. (2002b). Gestural spatialization in natural discourse segmentation. In *Proceedings of International Conference on Spoken Language Processing*, pp. 189–192.

Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K. E., Furuyama, N., & Ansari, R. (2000). Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of CVPR*, Vol. 2, pp. 247–254.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Sarkar, A. (2001). Applying co-training methods to statistical parsing. In *Proceedings of NAACL*, pp. 1–8.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of NAACL*, pp. 134–141.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., & Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, *32*.

Sidner, C. L. (1979). Towards a computational theory of definite anaphora comprehension in english discourse. Tech. rep. AITR-537, Massachusetts Institute of Technology.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521–544.

Strube, M., & Hahn, U. (1999). Functional centering: grounding referential coherence in information structure. *Computational Linguistics*, *25*(3), 309–344.

Strube, M., Rapp, S., & Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of EMNLP*, pp. 312–319.

Strube, M., & Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the ACL*, pp. 168–175.

Sundaram, H., & Chang, S.-F. (2003). Video analysis and summarization at structural and semantic levels. In D. Feng, W. C. S., & Zhang, H. (Eds.), *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*, pp. 75–94. Springer Verlag.

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In Getoor, L., & Taskar, B. (Eds.), *Introduction to Statistical Relational Learning*, pp. 95–130. MIT Press.

Sutton, C., McCallum, A., & Rohanimanesh, K. (2007). Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, *8*, 693–723.

Toyama, K., & Horvitz, E. (2000). Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Proceedings of Asian Conference on Computer Vision (ACCV)*.

Uchihashi, S., Foote, J., Girgensohn, A., & Boreczky, J. (1999). Video manga: generating semantically meaningful video summaries. In *Proceedings of ACM MULTIMEDIA*, pp. 383–392.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of Message Understanding Conference*, pp. 45–52.

Walker, M., Joshi, A., & Prince, E. (Eds.). (1998). *Centering Theory in Discourse*. Clarendon Press, Oxford.

Walker, M. A. (1998). Centering, anaphora resolution, and discourse structure. In Marilyn A. Walker, A. K. J., & Prince, E. F. (Eds.), *Centering in Discourse*, pp. 401–435. Oxford University Press.

Wang, S., Quattoni, A., Morency, L.-P., Demirdjian, D., & Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In *Proceedings of CVPR*, Vol. 02, pp. 1521–1527.

Xiong, Y., & Quek, F. (2006). Hand Motion Gesture Frequency Properties and Multimodal Discourse Analysis. *International Journal of Computer Vision*, *69*(3), 353–371.

Yang, X., Su, J., & Tan, C. L. (2005). Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of the ACL*, pp. 165–172.

Yang, X., Zhou, G., Su, J., & Tan, C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of the ACL*, pp. 176–183.

Zhu, X., Fan, J., Elmagarmid, A., & Wu, X. (2003). Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Systems*, *9*(1), 31–53.