Conditional Modality Fusion for Coreference Resolution

Jacob Eisenstein and Randall Davis
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
{jacobe,davis}@csail.mit.edu

Abstract

Non-verbal modalities such as gesture can improve processing of spontaneous spoken language. For example, similar hand gestures tend to predict semantic similarity, so features that quantify gestural similarity can improve semantic tasks such as coreference resolution. However, not all hand movements are informative gestures; psychological research has shown that speakers are more likely to gesture meaningfully when their speech is ambiguous. Ideally, one would attend to gesture only in such circumstances, and ignore other hand movements. We present *conditional modality fusion*, which formalizes this intuition by treating the informativeness of gesture as a hidden variable to be learned jointly with the class label. Applied to coreference resolution, conditional modality fusion significantly outperforms both early and late modality fusion, which are current techniques for modality combination.

1 Introduction

Non-verbal modalities such as gesture and prosody can increase the robustness of NLP systems to the inevitable disfluency of spontaneous speech. For example, consider the following excerpt from a dialogue 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 the gestures shown in Figure 1 make the meaning more clear. However, non-verbal modalities are often noisy, and their interactions with speech are complex (McNeill, 1992). Gesture, for example, is sometimes communicative, but other times merely distracting. While people have little difficulty distinguishing between meaningful gestures and irrelevant hand motions (e.g., self-touching, adjusting glasses) (Goodwin and Goodwin, 1986), NLP systems may be confused by such seemingly random movements. Our goal is to include non-verbal features only in the specific cases when they are helpful and necessary.

We present a model that learns in an unsupervised fashion when non-verbal features are useful, allowing it to gate the contribution of those features. The relevance of the non-verbal features is treated as a hidden variable, which is learned jointly with the class label in a conditional model. We demonstrate that this improves performance on binary coreference resolution, the task of determining whether a noun phrases refers to a single semantic entity. Conditional modality fusion yields a relative increase of 73% in the contribution of hand-gesture features. The model is not specifically tailored to gesture-speech integration, and may also be applicable to other non-verbal modalities.

2 Related work

Most of the existing work on integrating non-verbal features relates to prosody. For example, Shriberg et al. (2000) explore the use of prosodic features for sentence and topic segmentation. The first modal-

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 352–359,
Prague, Czech Republic, June 2007.

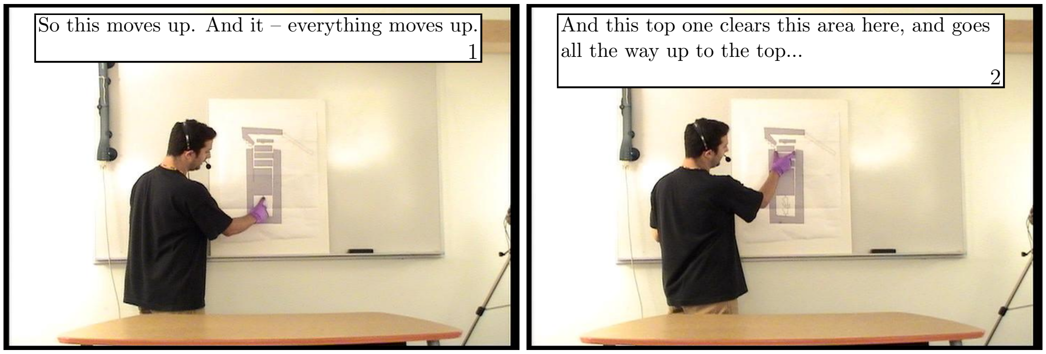

Figure 1: An example where gesture helps to disambiguate meaning.

ity combination technique that they consider trains a single classifier with all modalities combined into a single feature vector; this is sometimes called "early fusion." Shriberg et al. also consider training separate classifiers and combining their posteriors, either through weighted addition or multiplication; this is sometimes called "late fusion." Late fusion is also employed for gesture-speech combination in (Chen et al., 2004). Experiments in both (Shriberg et al., 2000) and (Kim et al., 2004) find no conclusive winner among early fusion, additive late fusion, and multiplicative late fusion.

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 the interpolation techniques described in (Shriberg et al., 2000), 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 is not gesturing meaningfully, 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.

3 Conditional modality fusion

The goal of our approach is to learn to weight the non-verbal features $\mathbf{x}_{nv}$ only when they are relevant. To do this, we introduce a hidden variable $m \in \{-1, 1\}$, which governs whether the non-verbal features are included. $p(m)$ is conditioned on a subset of features $\mathbf{x}_m$, which may belong to any modality; $p(m|\mathbf{x}_m)$ is learned jointly with the class label $p(y|\mathbf{x})$, with $y \in \{-1, 1\}$. For our coreference resolution model, y corresponds to whether a given pair of noun phrases refers to the same entity.

In a log-linear model, parameterized by weights $\mathbf{w}$, we have:

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

Here, ψ is a potential function representing the compatibility between the label y, the hidden variable m, 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 m. The denominator sums over both m and all possible labels y', yielding the conditional probability $p(y|\mathbf{x};\mathbf{w})$. The use of hidden variables in a conditionally-trained model follows (Quattoni et al., 2004).

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:

$$l(\mathbf{w}) = \sum_i \ln(p(y_i|\mathbf{x}_i; w)) \quad (1)$$

$$= \sum_i \ln \frac{\sum_m exp(\psi(y_i, m, \mathbf{x}_i; \mathbf{w}))}{\sum_{y',m} exp(\psi(y', m, \mathbf{x}_i; \mathbf{w}))} \quad (2)$$

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

The form of the potential function ψ is where our intuitions about the role of the hidden variable are formalized. Our goal is to include the non-verbal features $\mathbf{x}_{nv}$ only when they are relevant; consequently, the weight for these features should go to zero for some settings of the hidden variable m. In addition, *verbal* language is different when used in combination with meaningful non-verbal communication than when it is used unimodally (Kehler, 2000; Melinger and Levelt, 2004). Thus, we learn a different set of feature weights for each case: $\mathbf{w}_{v,1}$ when the non-verbal features are included, and $\mathbf{w}_{v,2}$ otherwise. The formal definition of the potential function for conditional modality fusion is:

$$\psi(y, m, \mathbf{x}; \mathbf{w}) \equiv$$

$$\begin{cases} y(\mathbf{w}_{v,1}^T \mathbf{x}_v + \mathbf{w}_{nv}^T \mathbf{x}_{nv}) + \mathbf{w}_m^T \mathbf{x}_m & m = 1 \\ y\mathbf{w}_{v,2}^T \mathbf{x}_v - \mathbf{w}_m^T \mathbf{x}_m & m = -1. \end{cases} \quad (3)$$

4 Application to coreference resolution

We apply conditional modality fusion to coreference resolution – the problem of partitioning the noun phrases in a document into clusters, where all members of a cluster refer to the same semantic entity. Coreference resolution on text datasets is well-studied (e.g., (Cardie and Wagstaff, 1999)). This prior work provides the departure point for our investigation of coreference resolution on spontaneous and unconstrained speech and gesture.

4.1 Form of the model

The form of the model used in this application is slightly different from that shown in Equation 3. When determining whether two noun phrases corefer, the features at each utterance must be considered. For example, if we are to compare the similarity of the gestures that accompany the two noun phrases, it should be the case that gesture is relevant during *both* time periods.

For this reason, we create two hidden variables, m_1 and m_2; they indicate the relevance of gesture over the first (antecedent) and second (anaphor) noun phrases, respectively. Since gesture similarity is only meaningful if the gesture is relevant during *both* NPs, the gesture features are included only if $m_1 = m_2 = 1$. Similarly, the linguistic feature weights $\mathbf{w}_{v,1}$ are used when $m_1 = m_2 = 1$; otherwise the weights $\mathbf{w}_{v,2}$ are used. This yields the model shown in Equation 4.

The vector of meta features $\mathbf{x}_{m_1}$ includes all single-phrase verbal and gesture features from Table 1, computed at the antecedent noun phrase; $\mathbf{x}_{m_2}$ includes the single-phrase verbal and gesture features, computed at the anaphoric noun phrase. The label-dependent verbal features $\mathbf{x}_v$ include both pairwise and single phrase verbal features from the table, while the label-dependent non-verbal features $\mathbf{x}_{nv}$ include only the pairwise gesture features. The single-phrase non-verbal features were not included because they were not thought to be informative as to whether the associated noun-phrase would participate in coreference relations.

4.2 Verbal features

We employ a set of verbal features that is similar to the features used by state-of-the-art coreference resolution systems that operate on text (e.g., (Cardie and Wagstaff, 1999)). Pairwise verbal features include: several string-match variants; distance features, measured in terms of the number of intervening noun phrases and sentences between the candidate NPs; and some syntactic features that can be computed from part of speech tags. Single-phrase verbal features describe the type of the noun phrase (definite, indefinite, demonstrative (e.g., *this ball*), or pronoun), the number of times it appeared in the document, and whether there were any adjecti-

$$\psi(y, m_1, m_2, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_{v,1}^T \mathbf{x}_v + \mathbf{w}_{nv}^T \mathbf{x}_{nv}) + m_1 \mathbf{w}_m^T \mathbf{x}_{m_1} + m_2 \mathbf{w}_m^T \mathbf{x}_{m_2}, & m_1 = m_2 = 1 \\ y \mathbf{w}_{v,2}^T \mathbf{x}_v + m_1 \mathbf{w}_m^T \mathbf{x}_{m_1} + m_2 \mathbf{w}_m^T \mathbf{x}_{m_2}, & \text{otherwise.} \end{cases} \quad (4)$$

val modifiers. The continuous-valued features were binned using a supervised technique (Fayyad and Irani, 1993).

Note that some features commonly used for coreference on the MUC and ACE corpora are not applicable here. For example, gazetteers listing names of nations or corporations are not relevant to our corpus, which focuses on discussions of mechanical devices (see section 5). Because we are working from transcripts rather than text, features dependent on punctuation and capitalization, such as apposition, are also not applicable.

4.3 Non-verbal features

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

The most straightforward measure of similarity is the Euclidean distance between the average hand position during each noun phrase – we call this the FOCUS-DISTANCE feature. 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 use a hidden Markov model (HMM) to perform a spatio-temporal clustering on hand position and velocity. The SAME-CLUSTER feature reports whether the hand positions during two noun phrases were usually grouped in the same cluster by the HMM. JS-DIV reports the Jensen-Shannon divergence, a continuous-valued feature used to measure the similarity in cluster assignment probabilities between the two gestures (Lin, 1991).

The gesture features described thus far 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 gesture trajectories, we use dynamic time warping (Huang et al., 2001), which gives a similarity metric for temporal data that is invariant to speed. This is reported in the DTW-DISTANCE feature.

All features are computed from hand and body pixel coordinates, which are obtained via computer vision; our vision system is similar to (Deutscher et al., 2000). The feature set currently supports only single-hand gestures, using the hand that is farthest from the body center. As with the verbal feature set, supervised binning was applied to the continuous-valued features.

4.4 Meta features

The role of the meta features is to determine whether the gesture features are relevant at a given point in time. To make this determination, both verbal and non-verbal features are applied; the only requirement is that they be computable at a single instant in time (unlike features that measure the similarity between two NPs or gestures).

Verbal meta features Meaningful gesture has been shown to be more frequent when the associated speech is ambiguous (Melinger and Levelt, 2004). Kehler finds that fully-specified noun phrases are less likely to receive multimodal support (Kehler, 2000). These findings lead us to expect that pro-

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

nouns should be likely to co-occur with meaningful gestures, while definite NPs and noun phrases that include adjectival modifiers should be unlikely to do so. To capture these intuitions, all single-phrase verbal features are included as meta features.

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 hands from the rest position, over the duration of the NP.

As noted in the previous section, a spatio-temporal clustering was performed on the hand positions and velocities, using an HMM. The REST-CLUSTER feature takes the value "true" iff the most frequently occupied cluster during the NP is the cluster closest to rest position. In addition, parameter tying in the HMM forces all clusters but one to represent static hold, with the remaining cluster accounting for the transition movements between holds. Only this last cluster is permitted to have an expected non-zero speed; if the hand is most frequently in this cluster during the NP, then the MOVEMENT-CLUSTER feature takes the value "true."

4.5 Implementation

The objective function (Equation 1) is optimized using a Java implementation of L-BFGS, a quasi-Newton numerical optimization technique (Liu and Nocedal, 1989). Standard L2-norm regularization is employed to prevent overfitting, with cross-validation to select the regularization constant. Although standard logistic regression optimizes a convex objective, the inclusion of the hidden variable renders our objective non-convex. Thus, convergence to a global minimum is not guaranteed.

5 Evaluation setup

Dataset Our dataset consists of sixteen short dialogues, in which participants explained the behavior of mechanical devices to a friend. There are nine different pairs of participants; each contributed two dialogues, with two thrown out due to recording errors. One participant, the "speaker," saw a short video describing the function of the device prior to the dialogue; the other participant was tested on comprehension of the device's behavior after the dialogue. The speaker was given a pre-printed diagram to aid in the discussion. For simplicity, only the speaker's utterances were included in these experiments.

The dialogues were limited to three minutes in duration, and most of the participants used the entire allotted time. "Markable" noun phrases – those that are permitted to participate in coreference relations – were annotated by the first author, in accordance with the MUC task definition (Hirschman and Chinchor, 1997). A total of 1141 "markable" NPs were transcribed, roughly half the size of the MUC6 development set, which includes 2072 markable NPs over 30 documents.

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 partitioning phase, in which the clusters of mutually-coreferring NPs are formed, maximizing some global criterion (Cardie and Wagstaff, 1999). Our model does not address the formation of noun-phrase clusters, but only the question of whether each pair of noun phrases in the document corefer. Consequently, we evaluate only the binary classification phase, and report results in terms of the area under the ROC curve (AUC). As the small size of the corpus did not permit dedicated test and development sets, results are computed using leave-one-out cross-validation, with one fold for each of the sixteen documents in the corpus.

Baselines Three types of baselines were compared to our conditional modality fusion (CMF) technique:

- **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 linear classifier; it uses the same code as the CMF model, but always includes all features.
- **Late fusion**. The late fusion baselines train separate classifiers for gesture and speech, and then combine their posteriors. The modality-specific classifiers are conditionally-trained linear models, 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. We report results for both additive and multiplicative combination of posteriors.
- **No fusion**. These baselines include the features from only a single modality, and again build a conditionally-trained linear classifier. Implementation uses the same code as the CMF model, but weights on features outside the target modality are forced to zero.

Although a comparison with existing state-of-the-art coreference systems would be ideal, all such available systems use verbal features that are inapplicable to our dataset, such as punctuation, capitalization, and gazetteers. The verbal features that we have included are a representative sample from the literature (e.g., (Cardie and Wagstaff, 1999)). The "no fusion, verbal features only" baseline thus provides a reasonable representation of prior work on coreference, by applying a maximum-entropy classifier to this set of typical verbal features.

Parameter tuning Continuous features are binned separately for each cross-validation fold, using only the training data. The regularization constant is selected by cross-validation within each training subset.

6 Results

Conditional modality fusion outperforms all other approaches by a statistically significant margin (Table 2). Compared with early fusion, CMF offers an absolute improvement of 1.20% in area under the ROC curve (AUC).[1] A paired t-test shows that this

[1] AUC quantifies the ranking accuracy of a classifier. If the AUC is 1, all positively-labeled examples are ranked higher than all negative-labeled ones.

model	AUC
Conditional modality fusion	**.8226**
Early fusion	.8109
Late fusion, multiplicative	.8103
Late fusion, additive	.8068
No fusion (verbal features only)	.7945
No fusion (gesture features only)	.6732

Table 2: Results, in terms of areas under the ROC curve

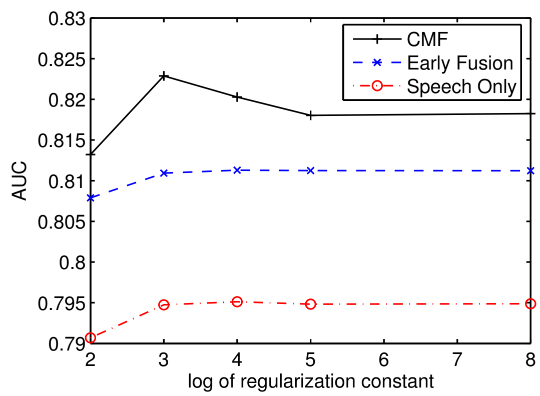

Figure 2: Conditional modality fusion is robust to variations in the regularization constant.

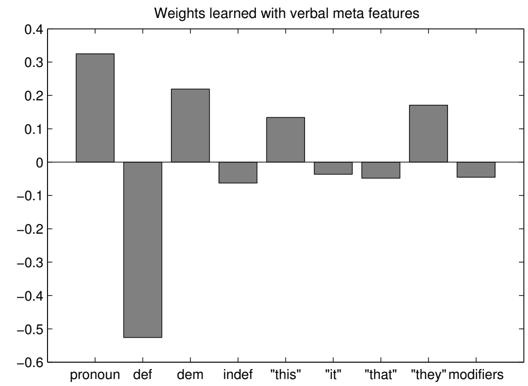

Figure 3: Weights for verbal meta features

result is statistically significant ($p < .002, t(15) = 3.73$). CMF obtains higher performance on fourteen of the sixteen test folds. Both additive and multiplicative late fusion perform on par with early fusion.

Early fusion with gesture features is superior to unimodal verbal classification by an absolute improvement of 1.64% AUC ($p < 4 * 10^{-4}, t(15) = 4.45$). Thus, while gesture features improve coreference resolution on this dataset, their effectiveness is increased by a relative 73% when conditional modality fusion is applied. Figure 2 shows how performance varies with the regularization constant.

7 Discussion

The feature weights learned by the system to determine coreference largely confirm our linguistic intuitions. Among the textual features, a large positive weight was assigned to the string match features, while a large negative weight was assigned to features such as number incompatibility (i.e., singular versus plural). The system also learned that gestures with similar hand positions and trajectories were likely to indicate coreferring noun phrases; all of our similarity metrics were correlated positively with coreference. A chi-squared analysis found that the EDIT DISTANCE was the most informative verbal feature. The most informative gesture feature was DTW-AGREEMENT feature, which measures the similarity between gesture trajectories.

As described in section 4, both textual and gestural features are used to determine whether the gesture is relevant. Among textual features, definite and indefinite noun phrases were assigned negative weights, suggesting gesture would not be useful to disambiguate coreference for such NPs. Pronouns were assigned positive weights, with "this" and the much less frequently used "they" receiving the strongest weights. "It" and "that" received lower weights; we observed that these pronouns were frequently used to refer to the immediately preceding noun phrase, so multimodal support was often unnecessary. Last, we note that NPs with adjectival modifiers were assigned negative weights, supporting the finding of (Kehler, 2000) that fully-specified NPs are less likely to receive multimodal support. A summary of the weights assigned to the verbal meta features is shown in Figure 3. Among gesture meta features, the weights learned by the system indicate that non-moving hand gestures away from the body are most likely to be informative in this dataset.

8 Future work

We have assumed that the relevance of gesture to semantics is dependent only on the currently available features, and not conditioned on prior history. In reality, meaningful gestures occur over contiguous blocks of time, rather than at randomly distributed instances. Indeed, the psychology literature describes a finite-state model of gesture, proceeding from "preparation," to "stroke," "hold," and then "retraction" (McNeill, 1992). These units are called *movement phases*. The relevance of various gesture features may be expected to depend on the movement phase. During strokes, the trajectory of the gesture may be the most relevant feature, while during holds, static features such as hand position and hand shape may dominate; during preparation and retraction, gesture features are likely to be irrelevant.

The identification of these movement phases should be independent of the specific problem of coreference resolution. Thus, additional labels for other linguistic phenomena (e.g., topic segmentation, disfluency) could be combined into the model. Ideally, each additional set of labels would transfer performance gains to the other labeling problems.

9 Conclusions

We have presented a new method for combining multiple modalities, which we feel is especially relevant to non-verbal modalities that are used to communicate only intermittently. Our model treats the relevance of the non-verbal modality as a hidden variable, learned jointly with the class labels. Applied to coreference resolution, this model yields a relative increase of 73% in the contribution of the gesture features. This gain is attained by identifying instances in which gesture features are especially relevant, and weighing their contribution more heavily. We next plan to investigate models with a temporal component, so that the behavior of the hidden variable is governed by a finite-state transducer.

Acknowledgments We thank Aaron Adler, Regina Barzilay, S. R. K. Branavan, Sonya Cates, Erdong Chen, Michael Collins, Lisa Guttentag, Michael Oltmans, and Tom Ouyang. This research is supported in part by MIT Project Oxygen.

References

Claire Cardie and Kiri Wagstaff. 1999. Noun phrase coreference as clustering. In *Proceedings of EMNLP*, pages 82–89.

Lei Chen, Yang Liu, Mary P. Harper, and Elizabeth Shriberg. 2004. Multimodal model integration for sentence unit detection. In *Proceedings of ICMI*, pages 121–128.

Jonathan Deutscher, Andrew Blake, and Ian Reid. 2000. Articulated body motion capture by annealed particle filtering. In *Proceedings of CVPR*, volume 2, pages 126–133.

Usama M. Fayyad and Keki B. Irani. 1993. Multi-interval discretization of continuousvalued attributes for classification learning. In *Proceedings of IJCAI-93*, volume 2, pages 1022–1027. Morgan Kaufmann.

M.H. Goodwin and C. Goodwin. 1986. Gesture and coparticipation in the activity of searching for a word. *Semiotica*, 62:51–75.

Lynette Hirschman and Nancy Chinchor. 1997. MUC-7 coreference task definition. In *Proceedings of the Message Understanding Conference*.

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. 2001. *Spoken Language Processing*. Prentice Hall.

Andrew Kehler. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI*, pages 685–690.

Joungbum Kim, Sarah E. Schwarm, and Mari Osterdorf. 2004. Detecting structural metadata with decision trees and transformation-based learning. In *Proceedings of HLT-NAACL'04*. ACL Press.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE transactions on information theory*, 37:145–151.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

David McNeill. 1992. *Hand and Mind*. The University of Chicago Press.

Alissa Melinger and Willem J. M. Levelt. 2004. Gesture and communicative intention of the speaker. *Gesture*, 4(2):119–141.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2004. Conditional random fields for object recognition. In *Neural Information Processing Systems*, pages 1097–1104.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tur, and Gokhan Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32.

Kentaro Toyama and Eric Horvitz. 2000. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Proceedings of ACCV '00, Fourth Asian Conference on Computer Vision*.