Gesture Improves 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

Coreference resolution, like many problems in natural language processing, has most often been explored using datasets of written text. While spontaneous spoken language poses well-known challenges, it also offers additional modalities that may help disambiguate some of the inherent disfluency. We explore features of hand gesture that are correlated with coreference. Combining these features with a traditional textual model yields a statistically significant improvement in overall performance.

1 Introduction

Although the natural language processing community has traditionally focused largely on text, face-to-face spoken language is ubiquitous, and offers the potential for breakthrough applications in domains such as meetings, lectures, and presentations. We believe that in face-to-face discourse, it is important to consider the possibility that non-verbal communication may offer features that are critical to language understanding. However, due to the long-standing emphasis on text datasets, there has been relatively little work on non-textual features in unconstrained natural language (prosody being the most notable exception).

Multimodal research in NLP has typically focused on dialogue systems for human-computer interaction (e.g., (Oviatt, 1999)); in contrast, we are interested in the applicability of multimodal features to unconstrained human-human dialogues. We believe that such features will play an essential role in bringing NLP applications such as automatic summarization and segmentation to multimedia documents, such as lectures and meetings.

More specifically, in this paper we explore the possibility of applying hand gesture features to the problem of coreference resolution, which is thought to be fundamental to these more ambitious applications (Baldwin and Morton, 1998). To motivate the need for multimodal features in coreference resolution, consider the following transcript:

> "[This circle (1)] is rotating clockwise and [this piece of wood (2)] is attached at [this point (3)] and [this point (4)] but [it (5)] can rotate. So as [the circle (6)] rotates, [this (7)] moves in and out. So [this whole thing (8)] is just going back and forth."

Even given a high degree of domain knowledge (e.g., that "circles" often "rotate" but "points" rarely do), determining the coreference in this excerpt seems difficult. The word "this" accompanied by a gesture is frequently used to introduce a new entity, so it is difficult to determine from the text alone whether "[this (7)]" refers to "[this piece of wood (2)]," or to an entirely different part of the diagram. In addition, "[this whole thing (8)]" could be anaphoric, or it might refer to a new entity, perhaps some superset of predefined parts.

The example text was drawn from a small corpus of dialogues, which has been annotated for coreference. Participants in the study had little difficulty understanding what was communicated. While this does not prove that human listeners are using gesture or other multimodal features, it suggests that these features merit further investigation. We extracted hand positions from the videos in the corpus, using computer vision. From the raw hand positions, we derived gesture features that were used to supplement traditional textual features for coreference resolution. For a description of the study's protocol, automatic hand tracking, and a fuller examination of the gesture features, see (Eisenstein and Davis, 2006). In this paper, we present results showing that these features yield a significant improvement in performance.

Appeared in *Proceedings of the Human Language Technology Conference of the NAACL, 2006, Companion Volume: Short Papers* , pp.37-40.

2 Implementation

A set of commonly-used linguistic features were selected for this problem (Table 1). The first five features apply to pairs of NPs; the next set of features are applied individually to both of the NPs that are candidates for coreference. Thus, we include two features each, e.g., **J is PRONOUN** and **I is PRONOUN**, indicating respectively whether the candidate anaphor and candidate antecedent are pronouns. We include separate features for each of the four most common pronouns: "this", "it", "that", and "they," yielding features such as **J="this"**.

2.1 Gesture Features

The gesture features shown in Table 1 are derived from the raw hand positions using a simple, deterministic system. Temporally, all features are computed at the midpoint of each candidate NP; for a further examination of the sensitivity to temporal offset, see (Eisenstein and Davis, 2006).

At most one hand is determined to be the "focus hand," according to the following heuristic: select the hand farthest from the body in the x-dimension, as long as the hand is not occluded and its y-position is not below the speaker's waist. If neither hand meets these criteria, than no hand is said to be in focus. Occluded hands are also not permitted to be in focus; the listener's perspective was very similar to that of the camera, so it seemed unlikely that the speaker would occlude a meaningful gesture. In addition, our system's estimates of the position of an occluded hand are unlikely to be accurate.

If focus hands can be identified during both mentions, the Euclidean distance between focus points is computed. The distance is binned, using the supervised method described in (Fayyad and Irani, 1993). An advantage of binning the continuous features is that we can create a special bin for missing data, which occurs whenever a focus hand cannot be identified.

If the same hand is in focus during both NPs, then the value of **WHICH HAND** is set to "same"; if a different hand is in focus then the value is set to "different"; if a focus hand cannot be identified in one or both NPs, then the value is set to "missing." This multi-valued feature is automatically converted into a set of boolean features, so that all features can be represented as binary variables.

2.2 Coreference Resolution Algorithm

(McCallum and Wellner, 2004) formulates coreference resolution as a Conditional Random Field, where mentions are nodes, and their similarities are represented as weighted edges. Edge weights range from $-\infty$ to ∞, with larger values indicating greater similarity. The optimal solution is obtained by partitioning the graph into cliques such that the sum of the weights on edges within cliques is maximized, and the sum of the weights on edges between cliques is minimized:

$$\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y}} \sum_{i,j,i \neq j} y_{i,j} s(x_i, x_j) \qquad (1)$$

In equation 1, $\mathbf{x}$ is a set of mentions and $\mathbf{y}$ is a coreference partitioning, such that $y_{i,j} = 1$ if mentions x_i and x_j corefer, and $y_{i,j} = -1$ otherwise. $s(x_i, x_j)$ is a similarity score computed on mentions x_i and x_j.

Computing the optimal partitioning $\hat{\mathbf{y}}$ is equivalent to the problem of correlation clustering, which is known to be NP-hard (Demaine and Immorlica, to appear). Demaine and Immorlica (to appear) propose an approximation using integer programming, which we are currently investigating. However, in this research we use average-link clustering, which hierarchically groups the mentions $\mathbf{x}$, and then forms clusters using a cutoff chosen to maximize the f-measure on the training set.

We experiment with both pipeline and joint models for computing $s(x_i, x_j)$. In the pipeline model, $s(x_i, x_j)$ is the posterior of a classifier trained on pairs of mentions. The advantage of this approach is that any arbitrary classifier can be used; the downside is that minimizing the error on all pairs of mentions may not be equivalent to minimizing the overall error of the induced clustering. For experiments with the pipeline model, we found best results by boosting shallow decision trees, using the Weka implementation (Witten and Frank, 1999).

Our joint model is based on McCallum and Wellner's (2004) adaptation of the voted perceptron to coreference resolution. Here, s is given by the product of a vector of weights λ with a set of boolean features $\phi(x_i, x_j)$ induced from the pair of noun phrases: $s(x_i, x_j) = \lambda\phi(x_i, x_j)$. The maximum likelihood weights can be approximated by a voted perceptron, where, in the iteration t of the perceptron training:

$$\lambda_t = \lambda_{t-1} + \sum_{i,j,i \neq j} \phi(x_i, x_j)(y^*_{i,j} - \hat{y}_{i,j}) \qquad (2)$$

In equation 2, $\mathbf{y}^*$ is the ground truth partitioning from the labeled data. $\hat{\mathbf{y}}$ is the partitioning that maximizes equation 1 given the set of weights λ_{t-1}. As before, average-link clustering with an adaptive cutoff is used to partition the graph. The weights are then averaged across all iterations of the perceptron, as in (Collins, 2002).

3 Evaluation

The results of our experiments are computed using mention-based CEAF scoring (Luo, 2005), and are reported in Table 2. Leave-one-out evaluation was used to form 16 cross-validation folds, one for each document in the corpus. Using a planned, one-tailed pairwise t-test, the gesture features improved performance significantly

MARKABLE DIST	The number of markables between the candidate NPs
EXACT MATCH	True if the candidate NPs have identical surface forms
STR MATCH	True if the candidate NPs match after removing articles
NONPRO MATCH	True if the candidate NPs are not pronouns and have identical surface forms
NUMBER MATCH	True if the candidate NPs agree in number
PRONOUN	True if the NP is a pronoun
DEF NP	True if the NP begins with a definite article, e.g. "the box"
DEM NP	True if the NP is not a pronoun and begins with the word "this"
INDEF NP	True if the NP begins an indefinite article, e.g. "a box"
pronouns	Individual features for each of the four most common pronouns: "this", "it", "that", and "they"
FOCUS DIST	Distance between the position of the in-focus hand during j and i (see text)
WHICH HAND	Whether the hand in focus during j is the same as in i (see text)

Table 1: The feature set

System	Feature set	F1
AdaBoost	Gesture + Speech	54.9
AdaBoost	Speech only	52.8
Voted Perceptron	Gesture + Speech	53.7
Voted Perceptron	Speech only	52.9
Baseline	EXACT MATCH only	50.2
Baseline	None corefer	41.5
Baseline	All corefer	18.8

Table 2: Results

for the boosted decision trees ($t(15) = 2.48, p < .02$), though not for the voted perceptron ($t(15) = 1.07, p = .15$).

In the "all corefer" baseline, all NPs are grouped into a single cluster; in the "none corefer", each NP gets its own cluster. In the "EXACT MATCH" baseline, two NPs corefer when their surface forms are identical. All experimental systems outperform all baselines by a statistically significant amount. There are few other reported results for coreference resolution on spontaneous, unconstrained speech; (Strube and Müller, 2003) similarly finds low overall scores for pronoun resolution on the Switchboard Corpus, albeit by a different scoring metric. Unfortunately, they do not compare performance to equivalent baselines.

For the AdaBoost method, 50 iterations of boosting are performed on shallow decision trees, with a maximum tree depth of three. For the voted perceptron, 50 training iterations were performed. The performance of the voted perceptron on this task was somewhat unstable, varying depending on the order in which the documents were presented. This may be because a small change in the weights can lead to a very different partitioning, which in turn affects the setting of the weights in the next perceptron iteration. For these results, the order of presentation of the documents was randomized, and the scores for the voted perceptron are the average of 10 different runs ($\sigma = 0.32\%$ with gestures, 0.40% without).

Although the AdaBoost method minimizes pairwise error rather than the overall error of the partitioning, its performance was superior to the voted perceptron. One possible explanation is that by boosting small decision trees, AdaBoost was able to take advantage of non-linear combinations of features. We tested the voted perceptron using all pairwise combinations of features, but this did not improve performance.

4 Discussion

If gesture features play a role in coreference resolution, then one might expect the probability of coreference to vary significantly when conditioned on features describing the gesture. As shown in Table 3, the prediction holds: the binned **FOCUS DIST** gesture feature has the fifth highest χ^2 value, and the relationship between coreference and all gesture features was significant ($\chi^2 = 727.8, dof = 4, p < .01$). Note also that although **FOCUS DIST** ranks fifth, three of the features above it are variants of a string-match feature, and so are highly redundant.

The **WHICH HAND** feature is less strongly correlated with coreference, but the conditional probabilities do correspond with intuition. If the NPs corefer, then the probability of using the same hand to gesture during both NPs is 59.9%; if not, then the likelihood is 52.8%. The probability of not observing a focus hand is 20.3% when the NPs corefer, 25.1% when they do not; in other words, gesture is more likely for both NPs of a coreferent pair than for the NPs of a non-coreferent pair. The relation between the **WHICH HAND** feature and coreference is also significantly different from the null hypothesis ($\chi^2 = 57.2, dof = 2, p < .01$).

Rank	Feature	χ^2
1.	EXACT MATCH	1777.9
2.	NONPRO MATCH	1357.5
3.	STR MATCH	1201.8
4.	J = "it"	732.8
5.	**FOCUS DIST**	727.8
6.	MARKABLE DIST	619.6
7.	J is PRONOUN	457.5
8.	NUMBER	367.9
9.	I = "it"	238.6
10.	I is PRONOUN	132.6
11.	J is INDEF NP	79.3
12.	**SAME FOCUS HAND**	57.2

Table 3: Top 12 Features By Chi-Squared

5 Related Work

Research on multimodality in the NLP community has usually focused on multimodal dialogue systems (e.g., (Oviatt, 1999)). These systems differ fundamentally from ours in that they address human-*computer* interaction, whereas we address human-*human* interaction. Multimodal dialogue systems tackle interesting and difficult challenges, but the grammar, vocabulary, and recognized gestures are often pre-specified, and dialogue is controlled at least in part by the computer. In our data, all of these things are unconstrained.

Prosody has been shown to improve performance on several NLP problems, such as topic and sentence segmentation (e.g., (Shriberg et al., 2000)). We are aware of no equivalent work showing statistically significant improvement on unconstrained speech using hand gesture features. (Nakano et al., 2003) shows that body posture predicts turn boundaries, but does not show that these features improve performance beyond a text-only system. (Chen et al., 2004) shows that gesture may improve sentence segmentation; however, in this study, the improvement afforded by gesture is not statistically significant, and evaluation was performed on a subset of their original corpus that was chosen to include only the three speakers who gestured most frequently. Still, this work provides a valuable starting point for the integration of gesture feature into NLP systems.

6 Conclusion

We have described how gesture features can be used to improve coreference resolution on a corpus of unconstrained speech. Hand position and hand choice correlate significantly with coreference, explaining this gain in performance. We believe this is the first example of hand gesture features improving performance by a statistically significant margin on unconstrained speech.

References

Breck Baldwin and Thomas Morton. 1998. Dynamic coreference-based summarization. In *Proc. of EMNLP*.

Lei Chen, Yang Liu, Mary P. Harper, and Elizabeth Shriberg. 2004. Multimodal model integration for sentence unit detection. In *Proceedings of International Conference on Multimodal Interfaces (ICMI'04)*. ACM Press.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

Erik D. Demaine and Nicole Immorlica. to appear. Correlation clustering in general weighted graphs. *Theoretical Computer Science*.

Jacob Eisenstein and Randall Davis. 2006. Gesture features for coreference resolution. In *Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

Usama M. Fayyad and Keki B. Irani. 1993. Multi-interval discretization of continuousvalued attributes for classification learning. In *Proceedings of IJCAI-93*.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. of HLT-EMNLP*, pages 25–32.

Andrew McCallum and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Neural Information Processing Systems*.

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of ACL'03*.

Sharon L. Oviatt. 1999. Mutual disambiguation of recognition errors in a multimodel architecture. In *Human Factors in Computing Systems (CHI'99)*, pages 576–583.

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

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL '03*, pages 168–175.

Ian H. Witten and Eibe Frank. 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.