Semantic Back-Pointers from Gesture

Jacob Eisenstein
MIT Computer Science and Artificial Intelligence Laboratory
77 Massachusetts Ave, MA 02139
jacobe@csail.mit.edu

1 Introduction

Although the natural-language processing community has dedicated much of its focus to text, face-to-face spoken language is ubiquitous, and offers the potential for breakthrough applications in domains such as meetings, lectures, and presentations. Because spontaneous spoken language is typically more disfluent and less structured than written text, it may be critical to identify features from additional modalities that can aid in 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 studied non-textual modality, e.g. (Shriberg et al., 2000)).

There are many non-verbal modalities that may contribute to face-to-face communication, including body posture, hand gesture, facial expression, prosody, and free-hand drawing. Hand gesture may be more expressive than any non-verbal modality besides drawing, since it serves as the foundation for sign languages in hearing-disabled communities. While non-deaf speakers rarely use any such systematized language as American Sign Language (ASL) while gesturing, the existence of ASL speaks to the potential of gesture for communicative expressivity.

Hand gesture relates to spoken language in several ways:

- Hand gesture communicates meaning. For example, (Kopp et al., 2006) describe a model of how hand gesture is used to convey spatial properties of its referents when speakers give navigational directions. This model both explains observed behavior of human speakers, and serves as the basis for an implemented embodied agent.
- Hand gesture communicates discourse structure. (Quek et al., 2002) and (McNeill, 1992) describe how the structure of discourse is mirrored by the the structure of the gestures, when speakers describe sequences of events in cartoon narratives.
- Hand gesture segments in unison with speech, suggesting possible applications to speech recognition and syntactic processing. (Morrel-Samuels and Krauss, 1992) show a strong correlation between the onset and duration of gestures, and their "lexical affiliates" – the phrase that is thought to relate semantically to the gesture. Also, (Chen et al., 2004) show that gesture features may improve sentence segmentation.

These examples are a subset of a broad literature on gesture that suggests that this modality could play an important role in improving the performance of NLP systems on spontaneous spoken language. However, the existence of significant relationships between gesture and speech does not prove that gesture will improve NLP; gesture features could be redundant with existing textual features, or they may be simply too noisy or speaker-dependant to be useful. To test this, my thesis research will identify specific, objective NLP tasks, and attempt to show that automatically-detected gestural features improve performance beyond what is attainable using textual features.

The relationship between gesture and meaning is particularly intriguing, since gesture seems to offer a unique, spatial representation of meaning to sup-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 215–218,
New York, June 2006.

plement verbal expression. However, the expression of meaning through gesture is likely to be highly variable and speaker dependent, as the set of possible mappings between meaning and gestural form is large, if not infinite. For this reason, I take the point of view that it is too difficult to attempt to decode individual gestures. A more feasible approach is to identify similarities between pairs or groups of gestures. If gestures do communicate semantics, then similar gestures should predict semantic similarity. Thus, gestures can help computers understand speech by providing a set of "back pointers" between moments that are semantically related. Using this model, my dissertation will explore measures of gesture similarity and applications of gesture similarity to NLP.

A set of semantic "back pointers" decoded from gestural features could be relevant to a number of NLP benchmark problems. I will investigate two: coreference resolution and disfluency detection. In coreference resolution, we seek to identify whether two noun phrases refer to the same semantic entity. A similarity in the gestural features observed during two different noun phrases might suggest a similarity in meaning. This problem has the advantage of permitting a quantitative evaluation of the relationship between gesture and semantics, without requiring the construction of a domain ontology.

Restarts are disfluencies that occur when a speaker begins an utterance, and then stops and starts over again. It is thought that the gesture may return to its state at the beginning of the utterance, providing a back-pointer to the restart insertion point (Esposito et al., 2001). If so, then a similar training procedure and set of gestural features can be used for both coreference resolution and restart correction. Both of these problems have objective, quantifiable success measures, and both may play an important role in bringing to spontaneous spoken language useful NLP applications such as summarization, segmentation, and question answering.

2 Current Status

My initial work involved hand annotation of gesture, using the system proposed in (McNeill, 1992). It was thought that hand annotation would identify relevant features to be detected by computer vision systems. However, in (Eisenstein and Davis, 2004), we found that the gesture phrase type (e.g., deictic, iconic, beat) could be predicted accurately by lexical information alone, without regard to hand movement. This suggests that this level of annotation inherently captures a synthesis of gesture and speech, rather than gesture alone. This conclusion was strengthened by (Eisenstein and Davis, 2005), where we found that hand-annotated gesture features correlate well with sentence boundaries, but that the gesture features were almost completely redundant with information in the lexical features, and did not improve overall performance.

The corpus used in my initial research was not suitable for automatic extraction of gesture features by computer vision, so a new corpus was gathered, using a better-defined experimental protocol and higher quality video and audio recording (Adler et al., 2004). An articulated upper body tracker, largely based on the work of (Deutscher et al., 2000), was used to identify hand and arm positions, using color and motion cues. All future work will be based on this new corpus, which contains six videos each from nine pairs of speakers. Each video is roughly two to three minutes in length.

Each speaker was presented with three different experimental conditions regarding how information in the corpus was to be presented: a) a pre-printed diagram was provided, b) the speaker was allowed to draw a diagram using a tracked marker, c) no presentational aids were allowed. The first condition was designed to be relevant to presentations involving pre-created presentation materials, such as Powerpoint slides. The second condition was intended to be similar to classroom lectures or design presentations. The third condition was aimed more at direct one-on-one interaction.

My preliminary work has involved data from the first condition, in which speakers gestured at preprinted diagrams. An empirical study on this part of the corpus has identified several gesture features that are relevant to coreference resolution (Eisenstein and Davis, 2006a). In particular, gesture similarity can be measured by hand position and the choice of the hand which makes the gesture; these similarities correlate with the likelihood of coreference. In addition, the likelihood of a gestural hold – where the hand rests in place for a period of

time – acts as a meta-feature, indicating that gestural cues are likely to be particularly important to disambiguate the meaning of the associated noun phrase. In (Eisenstein and Davis, 2006b), these features are combined with traditional textual features for coreference resolution, with encouraging results. The hand position gesture feature was found to be the fifth most informative feature by Chi-squared analysis, and the inclusion of gesture features yielded a statistically significant increase in performance over the textual features.

3 Future Directions

The work on coreference can be considered preliminary, because it is focused on a subset of our corpus in which speakers use pre-printed diagrams as an explanatory aide. This changes their gestures (Eisenstein and Davis, 2003), increasing the proportion of *deictic* gestures, in which hand position is the most important feature (McNeill, 1992). Hand position is assumed to be less useful in characterizing the similarity of *iconic* gestures, which express meaning through motion or handshape. Using the subsection of the corpus in which no explanatory aids were provided, I will investigate how to assess the similarity of such dynamic gestures, in the hope that coreference resolution can still benefit from gestural cues in this more general case.

Disfluency repair is another plausible domain in which gesture might improve performance. There are at least two ways in which gesture could be relevant to disfluency repair. Using the semantic backpointer model, restart repairs could be identified if there is a strong gestural similarity between the original start point and the restart. Alternatively, gesture could play a pragmatic function, if there are characteristic gestures that indicate restarts or other repairs. In one case, we are looking for a similarity between the disfluency and the repair point; in the other case, we are looking for similarities across all disfluencies, or across all repair points. It is hoped that this research will not only improve processing of spoken natural language, but also enhance our understanding of how speakers use gesture to structure their discourse.

4 Related Work

The bulk of research on multimodality in the NLP community relates to multimodal dialogue systems (e.g., (Johnston and Bangalore, 2000)). This research differs fundamentally from mine in that it addresses human-*computer* interaction, whereas I am studying human-*human* interaction. Multimodal dialogue systems tackle many interesting challenges, but the grammar, vocabulary, and recognized gestures are often pre-specified, and dialogue is controlled at least in part by the computer. In my data, all of these things are unconstrained.

Another important area of research is the generation of multimodal communication in animated agents (e.g., (Cassell et al., 2001; Kopp et al., 2006; Nakano et al., 2003)). While the models developed in these papers are interesting and often well-motivated by the psychological literature, it remains to be seen whether they are both broad and precise enough to apply to gesture recognition.

There is a substantial body of empirical work describing relationships between non-verbal and linguistic phenomena, much of which suggests that gesture could be used to improve the detection of such phenomena. (Quek et al., 2002) describe examples in which gesture correlates with topic shifts in the discourse structure, raising the possibility that topic segmentation and summarization could be aided by gesture features; Cassell et al. (2001) make a similar argument using body posture. (Nakano et al., 2003) describes how head gestures and eye gaze relate to turn taking and dialogue grounding. All of the studies listed in this paragraph identify relevant correlations between non-verbal communication and linguistic phenomena, but none construct a predictive system that uses the non-verbal modalities to improve performance beyond a text-only system.

Prosody has been shown to improve performance on several NLP problems, such as topic and sentence segmentation (e.g., (Shriberg et al., 2000; Kim et al., 2004)). The prosody literature demonstrates that non-verbal features can improve performance on a wide variety of NLP tasks. However, it also warns that performance is often quite sensitive, both to the representation of prosodic features, and how they are integrated with other linguistic features.

The literature on prosody would suggest parallels for gesture features, but little such work has been reported. (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.

5 Summary

Spontaneous spoken language poses difficult problems for natural language processing, but these difficulties may be offset by the availability of additional communicative modalities. Using a model of hand gesture as providing a set of semantic back-pointers to previous utterances, I am exploring whether gesture can improve performance on quantitative NLP benchmark tasks. Preliminary results on coreference resolution are encouraging.

References

Aaron Adler, Jacob Eisenstein, Michael Oltmans, Lisa Guttentag, and Randall Davis. 2004. Building the design studio of the future. In *Making Pen-Based Interaction Intelligent and Natural*, pages 1–7, Menlo Park, California, October 21-24. AAAI Press.

Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proc. of ACL*, pages 106–115.

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.

Jonathan Deutscher, Andrew Blake, and Ian Reid. 2000. Articulated body motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133.

Jacob Eisenstein and Randall Davis. 2003. Natural gesture in descriptive monologues. In *UIST'03 Supplemental Proceedings*, pages 69–70. ACM Press.

Jacob Eisenstein and Randall Davis. 2004. Visual and linguistic information in gesture classification. In *Proceedings of International Conference on Multimodal Interfaces(ICMI'04)*. ACM Press.

Jacob Eisenstein and Randall Davis. 2005. Gestural cues for sentence segmentation. Technical Report AIM-2005-014, MIT AI Memo.

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

Jacob Eisenstein and Randall Davis. 2006b. Gesture improves coreference resolution. In *Proceedings of NAACL*.

Anna Esposito, Karl E. McCullough, and Francis Quek. 2001. Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses. In *Proceedings of IEEE Workshop on Cues in Communication*.

Michael Johnston and Srinivas Bangalore. 2000. Finite-state multimodal parsing and understanding,. In *Proceedings of COLING-2000*, pages 369–375.

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.

Stefan Kopp, Paul Tepper, Kim Ferriman, and Justine Cassell. 2006. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Spatial Cognition and Computation*, In preparation.

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

P. Morrel-Samuels and R. M. Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:615–623.

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

Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, pages 171–193.

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