Multimodal Alignment by Optimization

Jacob Eisenstein JACOBE@MIT.EDU
Chris Mario Christoudias CMCH@MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory, 200 Technology Square, Cambridge MA, 02139 USA

1. Introduction

In face to face communication, speakers frequently use gesture to supplement speech (Chovil, 1992), using the additional modality to provide unique, non-redundant information (McNeill, 1992). The study of multimodal user interfaces is an attempt to give machines access to this same diversity of modalities. One of the simplest and most direct ways in which gesture can supplement verbal communication is by grounding references, usually through deixis. For example, it is impossible to extract the semantic content of the verbal utterance "I'll take this one" without an accompanying pointing gesture indicating the thing that is desired. This paper describes a novel technique for aligning gestures and speech. We evaluate our approach on a corpus of spoken explanations that does not require a fixed grammar or vocabulary.

1.1 Related Work

Previous multimodal user interfaces have taken two approaches to aligning gesture and spoken words. Several systems, including "Put-That-There" (Bolt, 1980) and ICONIC (Koons et al., 1993), simply choose the gesture nearest the relevant word. We evaluated this approach on our corpus and found that it is effective for short commands, but not for longer, more grammatically complex utterances.

Other researchers have used heavy-duty linguistic machinery such as unification (Johnston et al., 1997) and finite state transducers (Johnston & Bangalore, 2000). These systems have the drawback of depending on top-down grammatical parsing, which is not robust to disfluency, and require the user to learn their grammar and lexicon.

Our goal was to achieve comparable performance with a system that could be applied to *any* spoken language utterance. We drew inspiration from the field of anaphora resolution, which attempts to ground pronomial references (e.g, "it", "that") with noun phrases that appear earlier in the utterance. In particular, preference-based anaphora resolvers attempt to find the optimal set of anaphoric bindings based on soft linguistic constraints (e.g., gender and number agreement). We used a similar approach to gesture-speech alignment.

2. Preference-Based Alignment

In keeping with our lightweight NLP approach, we identify keywords that are likely to require gestural referents for resolution.[1] Our goal is to produce a set of alignments that match at least some of the keywords with one or more gestures. These alignments are guided by a set of linguistically-motivated preferences:

- The relevant gesture is usually close in time to the keyword.
- The gesture usually precedes the keyword.
- Multiple gestures are rarely aligned with the same keyword.
- Multiple keywords are rarely aligned with the same gesture.
- Some types of gestures (e.g. deictics) are more likely to be aligned with the given keywords than others.
- Some keyword/gesture combinations are particularly likely, e.g. "here" and a deictic hold.

We assign *penalties* to sets of bindings that violate these preferences, using a set of parameterizable penalty functions that can be applied to any set of bindings. Given a set of references and gestures, we can then try to find the set of bindings with the minimal penalty. This minimization is performed using greedy hill climbing.

The penalty functions themselves are parameterizable: we still must decide how much to penalize each of the above preferences. Choosing appropriate penalty parameters can be viewed as another optimization problem, which can be performed offline on training data. We tried several techniques for learning the penalty parameters: setting them by hand, gradient descent, simulated annealing, and a genetic algorithm. The genetic algorithm was the most successful, and the results we describe are based on parameters learned by this technique.

[1] We believe, however, that our alignment approach could be applied to more sophisticated entities such as noun phrases or semantic objects.

3. Evaluation

To evaluate our system, we collected a corpus of 26 monologues. Speakers were instructed to describe the behavior of a mechanical device as they would to another person, and were allowed to gesture at a predrawn diagram of the device. Gesture, grammar, and vocabulary were all completely unconstrained.

Each monologue was transcribed by hand; no gesture recognition was performed. Transcriptions included timestamps, gesture type, and other information, and subdivided each gesture into constituent *movement phrases*. Monologues ranged in size from minimums of fifteen seconds, 23 words, and six gesture phrases, to maximums of 90 seconds, 150 words, and 38 gesture phrases.

Since the penalty parameters have to be learned, the data was divided randomly into training and test sets. The results presented here are averages over ten different runs. As a baseline for comparison, we measured the performance of simply choosing the gesture nearest to every keyword. This baseline is identical to the alignment technique described in several implemented systems (Bolt, 1980; Koons et al., 1993; Sharma et al., 2000).

Performance is described in terms of accuracy and precision, because false negatives may not equal false positives: neither our system's output nor ground truth are guaranteed to contain exactly one aligned gesture for every keyword. As shown in Table 1, our system outperforms the baseline with respect to both recall and precision. One benefit of our penalty-based appoach is that it allows us to easily trade off between recall and precision. Reducing the penalties for unassigned gestures and keywords will cause the system to create fewer alignments, increasing precision and decreasing recall. This could be useful in a system where mistaken gesture/speech alignments are particularly undesirable. By increasing these same penalties, the opposite effect can also be achieved.

The performance of both systems decreases as monologue length increases. The baseline recall falls to 75% for the top quartile of monologues (by number of keywords). Similarly, our system's recall falls to 90% for the top quartile of monologues from the test set. Since longer monologues appear to be more grammatically complex and disfluent, this suggests that a naive approach will be effective for short commands, but not for longer utterances.

	Baseline	**Training**	**Test**
Recall	84.2%	94.6%	95.1%
σ	n/a	1.2%	5.1%
Precision	82.8%	94.5%	94.5%
σ	n/a	1.2%	5.0%

Table 1. Performance of our system versus a baseline

4. Conclusion

Very few comparable systems have been evaluated using a corpus of unconstrained speech and gesture. The only known exception is (Quek et al., 2002), which includes an informal evaluation on unconstrained human-to-human communication, but no quantitative results. Our system is shown to be robust to spoken English, even with a high level of disfluency.

Alignment is only one component of a comprehensive system for recognizing and understanding multimodal communication. While our evaluation indicates that our approach achieves what appears to be a high level of accuracy, the true test will be whether our system can actually support semantic information extraction from multimodal data. Only the construction of a comprehensive end-to-end system will reveal whether our relatively simple approach is sufficient, or whether more powerful – but also more brittle – linguistic tools will be required.

References

Bolt, R. A. (1980). Put-That-There: Voice and gesture at the graphics interface. *Proceedings of the 7th annual conference on Computer graphics and interactive techniques* (pp. 262–270). Seattle, Washington.

Chovil, N. (1992). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction, 25*, 163–194.

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

Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A., & Smith, I. (1997). Unification-based multimodal integration. *ACL-1997* (pp. 281–288). Somerset, New Jersey: Association for Computational Linguistics.

Koons, D. B., Sparrell, C. J., & Thorisson, K. R. (1993). Integrating simultaneous input from speech, gaze, and hand gestures. 257–276.

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

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

Sharma, R., Cai, J., Chakravarthy, S., Poddar, I., & Sethi, Y. (2000). Exploiting speech/gesture co-occurence for improving continuous gesture recognition in weather narration. *Proc. International Conference on Face and Gesture Recognition*.

Wu, L., Oviatt, S. L., & Cohen, P. R. (1999). Multimodal integration - a statistical view. *IEEE Transactions on Multimedia, 1*, 334–341.