Gesture Features for Sentence Segmentation

Jacob Eisenstein and Randall Davis

MIT Computer Science and Artificial Intelligence Laboratory,
32 Vassar Street, Cambridge MA 02139 USA,
{jacobe+davis} @csail.mit.edu

One of the ways in which gesture supplements communication is by helping to identify the "meta-data" that comprises the organizational structure of the discourse. One such type of meta-data is sentence unit boundaries;the detection of sentence boundaries in informal, spontaneous speech is a difficult problem. In this abstract, we explore whether gestural cues can improve sentence boundary detection.

We have hand-annotated a corpus of 26 short videos of spontaneous speech and gesture (see [1] for a more complete account of this research). We employ the movement phase and gesture phrase taxonomies as summarized by McNeill [2]. These gesture features correlate well with sentence boundaries in this corpus. Table 1 shows the probabilities of various gesture features, conditioned on the presence of sentence boundary events.

Feature	Value	$p(.\|S_{boundary})$	$p(.\|\neg S_{boundary})$
Gesture Unit Boundary	TRUE	.039	.0069
Gesture Phrase	BOUNDARY	.21	.088
	DEICTIC	.30	.43
	ICONIC	.46	.46
	BEAT	.031	.028
Movement Phase	BOUNDARY	.27	.13
	PREPARE	.066	.057
	STROKE	.31	.46
	HOLD	.26	.30
	RETRACT	.090	.052

Table 1. Bayesian model of gesture data

Given this relationship between gesture and sentence boundary events, we would expect the inclusion of gesture feature to improve sentence segmentation over a system using purely linguistic features. However, as shown in Figure 1, gesture features yielded only marginal improvements. The graph on the left shows that without using pause duration, gesture yields small but statistically insignificant improvements over purely lexical features, regardless of the choice of classifier or multimodal integration technique. The graph on the right shows that this pattern holds when the pause feature is included; the overall performance for both the baseline and multimodal systems improve.

In Proceedings *of 6th European Gesture Workshop,* pp.1-2. 2005.

We performed a multivariate linear regression (Table 2) between the true sentence boundary events and the predictions given by lexical, pause, and gesture models. The correlation between the gesture model posterior and the true sentence boundary events is statistically significant at $r = .17$ ($p < .01, df = 2103$). However, there is a great deal of overlap between the gesture features and those accounted for by the other models. Consequently, the predictive power of the gesture model *residual* – the part not accounted for by the lexical and pause models – is only $r = .05$. This is still statistically significant since the sample size is very large ($p < .03, df = 2103$), but unlikely to improve sentence segmentation unless the features captured by the lexical and pause models are very noisy.

Feature	r_{model}	$r_{residual}$	$r^2_{residual}/r^2_{model}$
Lexical	.42	.36	.74
Pause	.29	.16	.34
Gesture	.17	.05	.087

Table 2. Regression analysis of each feature type

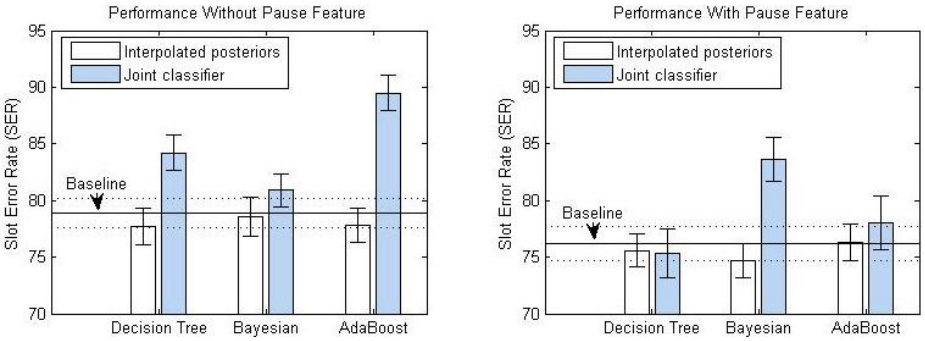

Fig. 1. Performance of multimodal sentence unit boundary detection. A lower Slot Error Rate indicates better performance. Error bars are 95% confidence intervals.

References

1. Eisenstein, J., Davis, R.: Gestural cues for sentence segmentation. Technical report, MIT AI Memo (2005)
2. McNeill, D.: Hand and Mind. The University of Chicago Press (1992)