JOINT SEGMENTATION AND CLASSIFICATION OF DIALOG ACTS IN MULTIPRTY MEETINGS

Matthias Zimmermann[1], Elizabeth Shriberg[1,2], Andreas Stolcke[1,2]

[1]International Computer Science Institute (ICSI), Berkeley, USA, [2]SRI International, Menlo Park, USA
{zimmerma, ees, stolcke}@icsi.berkeley.edu

Abstract

We investigate a scheme for joint segmentation and classification of dialog acts (DAs) based on a combination of a hidden-event language model (HELMs) and a maximum entropy classifier (MaxEnt). The MaxEnt takes into account dependencies between the duration of a pause and its word context (the identity of the surrounding words).

Task

The task considered in this work is to segment a sequence of words into individual DAs and assign one of five following mutually exclusive DA labels:

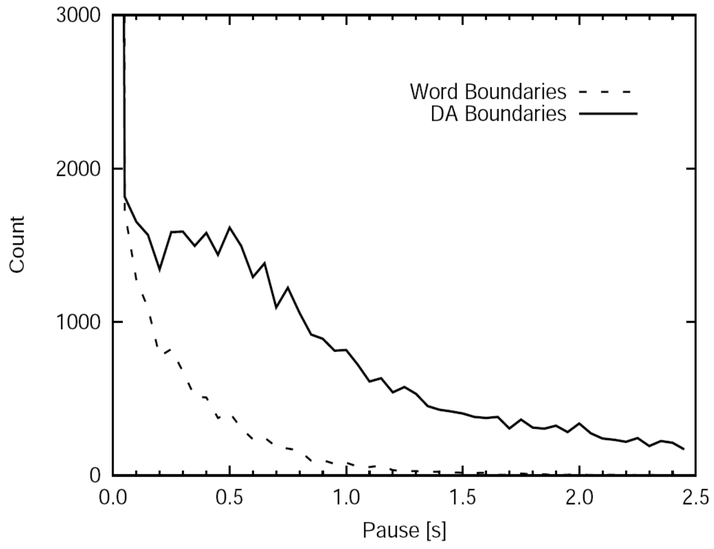
Backchannel	B	*uhhuh / right / yeah*
Disruption	D	*it's just it's / i mean you'd y-*
Floor grabber	F	*um / so / yeah*
Question	Q	*right / what does p. stand for anyway*
Statement	S	*yeah / so here's the thing*

Note, that we use Disruptions for all DAs that have been marked to be incomplete. All DAs that are related to the management of the floor are labeled as Floor grabbers.

For the evaluation of the system we use four error metrics. **NIST** and **DSER** evaluate the segmentation performance while the **Strict** and the **DER** metrics assess the quality of the joint segmentation and classification:

Ref:	`S|Q.Q.Q.Q|S.S.S|B|S.S`	
Sys:	`S|Q|S|Q.Q|D.D.D|F.F|S`	Error
NIST:	`.c.f.f...c.....c.m.f.`	100%
DSER:	`c|  e  |  c  |e| e`	60%
Strict:	`c.e.e.e.e.e.e.e.e.e.e`	91%
DER:	`c|  e  |  e  |e| e`	80%

e=correct, e=error, f=false alarm, and m=miss



Method

In previous work [1] the pause duration was modeled independently from words using decision trees. This paper applies a MaxEnt classifier to jointly model a pause duration and its word context to account for dependencies between a pause duration and its surrounding words, similar to [2]. For example, the average pause duration between the words *um um* is 0.935s while the average pause length between *you know* is only 0.006s. See the Figure above for the average (word independent) relationship between pause duration and having a DA boundary vs. a non-boundary event.

The features used by the **MaxEnt** classification scheme are the binned durations of the pauses between two words (10 bins from 0.0 to 3.0s) and up to four neighboring words, as well as word bigrams and pause/word bigram combinations. The labels consist of the 5 DA types and a non-boundary category to mark the boundary type after each word. For the combination of the output of the MaxEnt classifier with the HELM the approach described in [3] is used. This combination also matches the experimental setup of [1].

Experiments

The same experimental setup as in [1,4] for the joint segmentation and classification of the DAs was used.

Cond.	System	NIST	DSER	Strict	DER
Ref	[1]	34.5%	40.8%	64.4%	54.4%
	No Context	35.5%	39.7%	65.0%	55.9%
	With Context	34.8%	36.8%	62.8%	51.0%
STT	[1]	45.5%	49.4%	75.4%	64.3%
	No Context	49.7%	48.0%	76.2%	65.1%
	With Context	44.6%	47.4%	73.6%	62.6%

Reference conditions assume the true words while STT Manual uses the words generated by a speech-to-text (STT) system using a manual segmentation of the audio input. STT Auto refers to a setup where the segmentation of the audio input is done automatically.

Conclusion

We extended our previous work [4] to jointly model prosodic pause features with the word context in a MaxEnt classifier, in contrast to [1] that modeled prosodic features separately from word identities. Resulting error rate reductions between 2.2% and 4.9% (absolute) confirm the validity of the presented approach.

[1] J. Ang et al. “Automatic dialog act segmentation and classification in multiparty meetings”, ICASSP, 2005

[2] J. Huang and G. Zweig, “Maximum entropy model for punctuation annotation from speech”, ICSLP, 2002

[3] A. Stolcke and E. Shriberg, “Automatic detection of sentence boundaries and disfluencies based on recognized words”, ICSLP, 1998

[4] M. Zimmermann et al. “Toward joint segmentation and classification of dialog acts in multi-party meetings”, MLMI, 2005

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication), by DARPA Contract NBCHD030010 through the SRI CALO project (approved for public release, distribution unlimited), NSF Awards IIS-0121396 and IRI-9619921, and the Swiss National Science Foundation through the research network IM2.