Automatic Detection of Discourse Structure for Speech Recognition and Understanding

Daniel Jurafsky *(University of Colorado)*, **Rebecca Bates** *(Boston University)*,
Noah Coccaro *(University of Colorado)*, **Rachel Martin** *(Johns Hopkins University)*,
Marie Meteer *(BBN)*, **Klaus Ries** *(CMU/Universität Karlsruhe)*,
Elizabeth Shriberg *(SRI)*, **Andreas Stolcke** *(SRI)*,
Paul Taylor *(University of Edinburgh)*, **Carol Van Ess-Dykema** *(DoD)*

We describe a new approach for statistical modeling and detection of discourse structure for natural conversational speech. Our model is based on 42 'Dialog Acts' (DAs), (question, answer, backchannel, agreement, disagreement, apology, etc). We labeled 1155 conversations from the Switchboard (SWBD) database (Godfrey *et al.* 1992) of human-to-human telephone conversations with these 42 types and trained a Dialog Act detector based on three distinct knowledge sources: sequences of words which characterize a dialog act, prosodic features which characterize a dialog act, and a statistical Discourse Grammar. Our combined detector, although still in preliminary stages, already achieves a 65% Dialog Act detection rate based on acoustic waveforms, and 72% accuracy based on word transcripts. Using this detector to switch among the 42 Dialog-Act-Specific trigram LMs also gave us an encouraging but not statistically significant reduction in SWBD word error.

1 Introduction

The ability to model and automatically detect discourse structure is essential as we address problems like *understanding spontaneous dialog* (a meeting summarizer needs to know who said what to who), *building human-computer dialog systems* (a conversational agent needs to know whether it just got asked a question or ordered to do something), and *transcription of conversational speech* (utterances with different discourse function also have very different words). This paper describes our preliminary work (as part of the 1997 Summer Workshop on Innovative Techniques in LVCSR) on automatically detecting discourse structure for speech recognition and understanding tasks.

Table 1 shows a sample of the kind of discourse structure we are modeling and detecting. Besides the usefulness of discourse structure detection for speech understanding, discourse structure can be directly relevant for speech recognition tasks. For example in the state-of-the-art HTK recognizer we used, the word **do** has an error rate of 72%. But **do** is in almost every **Yes-No-Question**; if we could detect **Yes-No-Question**s (for example by looking for utterances with rising intonation) we could increase the probability of **do** and hence decrease the error rate.

There are many excellent previous attempts to build predictive, stochastic models of dialog structure (Kita *et al.* 1996; Mast *et al.* 1996; Nagata and Morimoto 1994; Reithinger *et al.* 1996; Suhm and Waibel 1994; Taylor *et al.* 1998; Woszczyna and Waibel 1994; Yamaoka and Iida 1991), and our effort is in many ways inspired by

Spkr	Dialog Act	Utterance
A	**Wh-Question**	What kind do you have now?
B	**Statement**	*Uh, we have a, a Mazda nine twenty nine and a Ford Crown Victoria and a little two seater CRX.*
A	**Acknowledge-Answer**	Oh, okay.
B	**Opinion**	*Uh, it's rather difficult to, to project what kind of, uh, -*
A	**Statement**	we'd, look, always look into, uh, consumer reports to see what kind of, uh, report, or, uh, repair records that the various cars have –
B	**Turn-Exit**	*So, uh, -*
A	**Yes-No-Quest**	And did you find that you like the foreign cars better than the domestic?
B	**Answer-Yes**	*Uh, yeah,*
B	**Statement**	*We've been extremely pleased with our Mazdas.*
A	**Backchannel-Quest**	Oh, really?
B	**Answer-Yes**	*Yeah.*

Table 1: *A fragment of a labeled switchboard conversation.*

this work, and indeed our group overlaps in personnel with some of these projects. Our project extends these earlier efforts particularly in its scale (our models were trained on 1155 dialog-annotated conversations comprising 205,000 utterances and 1.4 million words; an order of magnitude larger than any previous system) and in focusing on longer, less task-oriented dialogs.

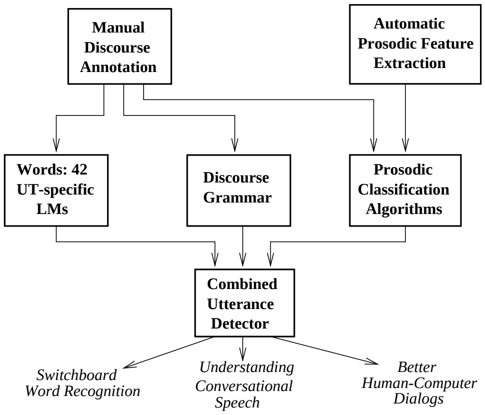
Manual Discourse Annotation
Automatic Prosodic Feature Extraction
Words: 42 UT-specific LMs
Discourse Grammar
Prosodic Classification Algorithms
Combined Utterance Detector
Switchboard Word Recognition
Understanding Conversational Speech
Better Human-Computer Dialogs

Figure 1: *Outline of Paper and Project*

Figure 1 outlines the major stages of our work and of this paper. We begin by discussing how we manually annotated 1155 conversations with hand-labeled discourse-tags. We then describe the 3 knowledge sources for dialog act detection (word-sequences, discourse grammar, and prosody), show how these knowledge sources can be combined, and finally apply the detector to help improve word recognition of SWBD.

2 Manual Discourse Tagging

In order to tag the 1155 SWBD conversations, we first designed the SWBD-DAMSL tagset (Jurafsky *et al.* 1997b) by augmenting the DAMSL tag-set (Core and Allen 1997). Our SWBD-DAMSL tagset consists of approximately 60 groupable labels in orthogonal dimensions which the labelers combined to produce 220 unique tags for the 205,000 SWBD utterances. The SWBD conversations had already been hand-segmented into utterances by the Linguistic Data Consortium ((Meteer *et al.* 1995); an utterance roughly corresponds to a sentence). Each utterance thus received exactly one of these 220 tags. The average conversation consisted of 144 turns, 271 utterances, and took 28 minutes for trained CU Boulder grad students to label. The labeling agreement was 84% (κ = .80; (Carletta 1996)). We then clustered these 220 tags into 42 final tags. All of our experiments were run with this 42-tag tagset. Table 2 shows the most common tags. [1]

Tag	**Example**	**Count**	**%**
Statement	*Me, I'm in the legal department.*	72,824	36%
Backchannel	*Uh-huh.*	37,096	19%
Opinion	*I think it's great*	25,197	13%
Agree/Accept	*That's exactly it.*	10,820	5%
Abandoned/Turn-Exit	*So, -/*	10,569	5%
Appreciation	*I can imagine.*	4,633	2%
Yes-No-Question	*Do you have to have any special training*	4,624	2%
Non-verbal	*<Laughter>,<Throat_clearing>*	3,548	2%
Yes answers	*Yes.*	2,934	1%
Conventional-closing	*Well, it's been nice talking to you.*	2,486	1%
Uninterpretable	*But, uh, yeah*	2,158	1%
Wh-Question	*Well, how old are you?*	1,911	1%
No answers	*No.*	1,340	1%
Response Ack	*Oh, okay.*	1,277	1%
Hedge	*I don't know if I'm making any sense or not.*	1,182	1%
Declarative Question	*So you can afford to get a house?*	1,174	1%
Other	*Well give me a break, you know.*	1,074	1%
Backchannel-Question	*Is that right?*	1,019	1%

Table 2: *18 most frequent tags (of 42)*

3 Dialog Act Detection

The goal of our dialog act (DA) detection algorithms is to automatically assign the correct tag from our 42 DA set to each of the presegmented utterance wavefiles. We achieved a 65% detection accuracy, based on automatic word recognition and prosodic analysis. This compares with a baseline of 35% if we simply chose the most frequent dialog act each time. Human labelers were able to do significantly better (84%). However, note that the human labeling was based purely on word transcripts. Using actual, rather than recognized words, our DA detection algorithm achieved 72% accuracy, so we can expect substantially improved automatic detection simple as a result of continually improving recognition accuracy.

Our algorithm is based on combining three sources of knowledge:

[1] For many of our experiments we combined the **Statement** and **Opinion** classes; these two classes together comprise 49% of the utterances, but a full **83%** of the words in the corpus. As we will see, this limits the affect our tagging had on word-related metrics like word error.

Prosodic Information: Using prosodic features such as pitch and speaking rate to choose DA. For example based on the literature we predicted that **Yes-No-Questions** would be detectable from their final F0 rise.

Words and Word Grammar: Pick the most likely DA given the word string. For example, **88.4%** of the trigrams **“<start> do you”** occur in **Yes-No-Questions**.

Discourse Grammar: Pick the Dialog Act which is most likely given the surrounding DAs. For example a **Command** will be **Agreed** to with probability **.23**, a **Yes-No-Question** will receive a **Yes** answer with probability **.30**.

The utterance detection algorithm we describe is based on hand-segmented utterance boundaries. That is, both our training and test sets were segmented by hand into turns and utterances.

3.1 Prosodic Dialog Act Detection

Extending earlier work by others on the use of prosodic knowledge for dialog act prediction (Mast *et al.* 1996; Taylor *et al.* 1997; Terry *et al.* 1994; Waibel 1988), we automatically extracted prosodic features for each utterance, performed various normalizing and postprocessing, and trained CART decision trees to predict the dialog act of an utterance. Our goal was to discriminate classes that were particularly confusable given only words, and also to understand prosodic feature usage so as to build future prosodic detectors. We used no word information in extracting or computing features, other than the location of utterance boundaries which were assumed for all knowledge sources. Features included duration, pause, F0, energy (RMS and signal-to-noise-ratio), and speaking rate (using a signal processing measure ‘enrate’ (Morgan *et al.* 1997)) measures. Because the distribution of DAs was highly skewed,we downsampled our data to uniform priors (train and test) to train more discriminate trees.

We built trees to detect specifically confusable dialog acts including **Yes-No-Questions**, and **Abandoned/Turn Exits**. For **Yes-No-Question**s, for example, a word-based detector for all 42 types only achieved 32% accuracy. But **Yes-No-Questions** are strongly prosodically marked: they generally have a rising F0 contour. Table 3 shows our accuracy for a single prosodic tree for distinguishing **Yes-No-Question** from all other dialog acts. We achieved an accuracy (# of correct classifications / all data) of 70.3% (where chance is 50%).

Test	Count	Accuracy	Perp	Entropy$_e$	Efficiency
Prior		50.0	2	0.693	
Cond. HLD	618	70.3	1.80	0.589	15.0%

Table 3: *Results of Yes-No-Question Detection Tree*

This **Yes-No-Question** tree relied mainly on the F0 rise, but also on other features; nearly all the main feature types played a role in the trees. For details of the decision trees, see Shriberg *et al.* (submitted) and Jurafsky *et al.* (1997a). In a number of focussed analyses assuming uniform DA priors, prosody alone allowed classification significantly above chance. In addition, although space does not permit discussion, adding prosody to word information significantly improved classification for the majority of the analyses.

3.2 Word-sequence-based Dialog Act Detection

Word-based DA detection is based on separate trigram language models for each of the 42 dialog acts, (i.e. one LM for **Statements**, another for **Yes-No-Questions**, another for **Backchannels**, etc). and choosing the dialog act that assigns the highest likelihood to the word string (Garner *et al.* 1996; Peskin *et al.* 1996). The resulting LMs were quite distinct from each other, and had a significantly lower perplexity (66.9) on the test set than the baseline LM (76.8) indicating that the 42 LMs do in fact capture the lexical distinctions among the 42 dialog acts (see Table 6).

We then used the 42 language models to choose the most likely DA given the word string, by maximizing over likelihoods of the utterance-words given the utterance (e.g. $P(\text{I lived in Chicago}|\text{Statement}))$ for each utterance in a conversation and for each DA. Table 5 in §3.4 shows that by using the bigram Discourse Grammar described below we achieve 64.6% utterance detection accuracy using the likelihoods computed via the 2500-best word strings from each utterance (where 35% is chance). (Using the correct (reference, i.e. cheating) word strings, we achieved 70.6% accuracy).

3.3 Discourse Grammar

Our discourse grammar is a backoff N-gram (Katz 1987) with Witten-Bell discounting (Witten and Bell 1991) which predicts the sequence of dialog acts given the previous types; the use of N-gram discourse grammars was motivated by previous work by Kita *et al.* (1996); Mast *et al.* (1996); Nagata and Morimoto (1994); Suhm and Waibel (1994); Taylor *et al.* (1997); Taylor *et al.* (1998); Woszczyna and Waibel (1994); Yamaoka and Iida (1991). For example, in the sample conversation in Table 1, the grammar gives the probability of the utterance in Channel A being an **Acknowledge-Answer** given that the previous utterance was a **Statement** on Channel B and before that was a **Wh-Question** on Channel A.

	N-gram model			
n	0	1	2	3
perplexity	42.0	9.0	5.1	4.8

Table 4: *DA Perplexity (conditioned on turns).*

As Table 4 shows, the discourse grammar does in fact progressively reduce the perplexity of the utterance detection task as a larger dialog act history is added. We also explored alternative models for discourse grammar, including maximum entropy models and cache models. See Jurafsky *et al.* (1997a) for further details.

3.4 The Combined Dialog Act Detector

We then ran a number of different experiments combining our three knowledge sources (words, prosody, discourse) for DA detection. The prosodic component of these combined detection results is still preliminary, because we only had a very preliminary prosodic detection tree at this point (distinguishing **Statements**, **Questions**, **Backchannels**, **Agreements**, and **Abandoned** from each other and from other DAs), and also because we are still studying the optimal way to combine different prosodic classifiers. See Stolcke *et al.* (submitted) and Jurafsky *et al.* (1997a) for the mathematical foundation of our combinations; Table 5 simply shows our final detection results.

Discourse	Accuracy (%)		
Grammar	Prosody only	Rec. Words only	Combined
None	38.9	42.8	56.5
Unigram	48.3	61.9	62.6
Bigram	50.2	64.6	**65.0**

Table 5: *Combined utterance detection accuracies.*

Using the recognized words together with the bigram discourse grammar accounts for the bulk of our accuracy, although we expect more help from the prosody as we train more trees.

4 Word Recognition Experiments

We applied our detection algorithm to the SWBD word-recognition task by using a mixture of the 42 DA-specific LMs to rescore each test-set utterance, and using the combined detector to set the mixture weights. Table 6 shows word error and perplexities obtained for the DA-conditioned mixture LM. Also shown are the results for the baseline LM, and for the 'cheating' LM, conditioned on the true DA labels. WER is reduced by only 0.3% over the baseline, a non-significant change ($0.3 > p > 0.2$).

Model	WER (%)	Perplexity
Baseline	**41.2**	76.8
Mixture LM	**40.9**	66.9
Cheating LM	40.3	66.8

Table 6: *Non-significant reduction in SWBD word error.*

It is encouraging that the perplexity of the DA-conditioned mixture model is virtually the same as that of the cheating LM. But the cheating experiment shows that even perfect knowledge of the dialog acts can only be expected to give about a 1 percent reduction in WER. This is mainly because **Statements** (non-opinion plus opinion) account for 83% of the words in our corpus (since e.g. backchannels and answers tend to be short). Table 7 shows, however, that using utterance-specific language models can significantly improve WER for some dialog acts, and hence this approach could prove useful for tasks with a different distribution of utterance types.

5 Conclusions

We have described a new approach for statistical modeling and detection of discourse structure for natural conversational speech. Our algorithm has possibilities for reducing word error in speech recognition. Although the skewed dialog act distribution limited our maximum word error improvement for the Switchboard task, improvements for WER of individual dialog acts suggests that the algorithm has

Dialog Act	WER	Oracle WER	Improvement with Oracle
Answer No	29.4	11.8	-17.6%
Backchannel	25.9	18.6	-7.3%
Backchannel Questions	15.2	9.1	-6.1%
Abandoned/Turn-Exit	48.9	45.2	-3.7%
Wh-Questions	38.4	34.9	-3.5%
Yes-No-Questions	55.5	52.3	-3.2%
Statement	42.0	41.5	-0.5%

Table 7: *Cheating Error Rates on Specific Dialog Acts*

potential to improve recognition on other tasks (like conversational agents) where questions and other non-statements are more common. Furthermore, by combining our three knowledge sources, we achieved significant improvements in our ability to automatically detect dialog acts, which will help address tasks like understanding spontaneous dialog and building human-computer dialog systems.

Acknowledgments

This project was supported by the generosity of many: the 1997 Workshop on Innovative Techniques in LVCSR, the Center for Speech and Language Processing at Johns Hopkins University, and the NSF (via NSF IRI-9619921 and IRI-9314967). Thanks to the students who did the labeling: (1) the discourse labelers at Boulder: Debra Biasca (who managed the Boulder labelers), Marion Bond, Traci Curl, Anu Erringer, Michelle Gregory, Lori Heintzelman, Taimi Metzler, and Amma Oduro and (2) the intonation labelers at Edinburgh: Helen Wright, Kurt Dusterhoff, Rob Clark, Cassie Mayo and Matthew Bull. Many thanks to Susann LuperFoy, Nigel Ward, James Allen, Julia Hirschberg, and Marilyn Walker for advice on the design of the SWBD-DAMSL tag-set, to Mitch Weintraub and Chuck Wooters for many helpful comments, to Bill Byrne, Harriet Nock, and Joe Picone for baselines, test data, and recognition environment, to Fred Jelinek for his advice and encouragement and to Kimberly Shiring for everything else.

References

CARLETTA, JEAN. 1996. Assessing agreement on classification tasks: The Kappa statistic. Computational Linguistics 22.249–254.

CORE, MARK G., and JAMES ALLEN. 1997. Coding dialogs with the DAMSL annotation scheme. AAAI Fall Symposium on Communicative Action in Humans and Machines, MIT, Cambridge, MA.

GARNER, P. N., S. R. BROWNING, R. K. MOORE, and R. J. RUSSELL. 1996. A theory of word frequencies and its application to dialogue move recognition. ICSLP-96, 1880–1883, Philadephia.

GODFREY, J., E. HOLLIMAN, and J. MCDANIEL. 1992. SWITCHBOARD: Telephone speech corpus for research and development. Proceedings of ICASSP-92, 517–520, San Francisco.

JURAFSKY, DANIEL, REBECCA BATES, NOAH COCCARO, RACHEL MARTIN, MARIE METEER, KLAUS RIES, ELIZABETH SHRIBERG, ANDREAS STOLCKE, PAUL TAYLOR, and CAROL VAN ESS-DYKEMA. 1997a. Switchboard discourse language modeling project report. Technical report, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD.

JURAFSKY, DANIEL, ELIZABETH SHRIBERG, and DEBRA BIASCA, 1997b. Switchboard-DAMSL Labeling Project Coder's Manual. http://stripe.colorado.edu/˜jurafsky/ manual.august1.html.

KATZ, SLAVA M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recogniser. IEEE Trans. ASSP 35.400–401.

KITA, KENJI, YOSHIKAZU FUKUI, MASAAKI NAGATA, and TSUYOSHI MORIMOTO. 1996. Automatic acquisition of probabilistic dialogue models. ICSLP-96, 196–199, Philadephia.

MAST, M., R. KOMPE, ST. HARBECK, A. KIESSLING, H. NIEMANN, , and E. NÖTH. 1996. Dialog act classification with the help of prosody. ICSLP-96, 1728–1731, Philadephia.

METEER, MARIE, and OTHERS. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus. Linguistic Data Consortium. Revised June 1995 by Ann Taylor. ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps.gz.

MORGAN, NELSON, ERIC FOSLER, and NIKKI MIRGHAFORI. 1997. Speech recognition using on-line estimation of speaking rate. EUROSPEECH-97, Rhodes, Greece.

NAGATA, MASAAKI, and TSUYOSHI MORIMOTO. 1994. First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance. Speech Communication 15.193–203.

PESKIN, BARBARA, SEAN CONNOLLY, LARRY GILLICK, STEPHEN LOWE, DON MCALLASTER, VENKI NAGESHA, PAUL VAN MULBREGT, and STEVEN WEGMANN. 1996. Improvements in switchboard recognition and topic identification. ICASSP-96, volume 1, 303–306.

REITHINGER, NORBERT, RALF ENGEL, MICHAEL KIPP, and MARTIN KLESEN. 1996. Predicting dialogue acts for a speech-to-speech translation system. ICSLP-96, 654–657, Philadephia.

SHRIBERG, ELIZABETH, REBECCA BATES, PAUL TAYLOR, ANDREAS STOLCKE, DANIEL JURAFSKY, KLAUS RIES, NOAH COCCARO, RACHEL MARTIN, MARIE METEER, and CAROL VAN ESS-DYKEMA. submitted. Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech .

STOLCKE, ANDREAS, ELIZABETH SHRIBERG, REBECCA BATES, NOAH COCCARO, DANIEL JURAFSKY, RACHEL MARTIN, MARIE METEER, KLAUS RIES, PAUL TAYLOR, and CAROL VAN ESS-DYKEMA. submitted. Dialog act modeling for conversational speech. AAAI Spring Symposium on Applying Machine Learning to Discourse Processing .

SUHM, B., and A. WAIBEL. 1994. Toward better language models for spontaneous speech. ICSLP-94, 831–834.

TAYLOR, PAUL, SIMON KING, STEPHEN ISARD, HELEN WRIGHT, and JACQUELINE KOWTKO. 1997. Using intonation to constrain language models in speech recognition. EUROSPEECH-97, 2763–2766, Rhodes, Greece.

TAYLOR, PAUL A., S. KING, S. D. ISARD, and H. WRIGHT. 1998. Intonation and dialogue context as constraints for speech recognition. Submitted to Language and Speech .

TERRY, MARK, RANDALL SPARKS, and PATRICK OBENCHAIN. 1994. Automated query identification in English dialogue. ICSLP-94, 891–894.

WAIBEL, ALEX. 1988. Prosody and Speech Recognition. San Mateo, CA.: Morgan Kaufmann.

WITTEN, I. H., and T. C. BELL. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. IEEE Trans. Information Theory 37.1085–1094.

WOSZCZYNA, M., and A. WAIBEL. 1994. Inferring linguistic structure in spoken language. ICSLP-94, 847–850, Yokohama, Japan.

YAMAOKA, TAKAYUKI, and HITOSHI IIDA. 1991. Dialogue interpretation model and its application to next utterance prediction for spoken language processing. EUROSPEECH-91, 849–852, Genova, Italy.