Dialog Act Modeling for Conversational Speech

Andreas Stolcke and Elizabeth Shriberg, SRI International
Rebecca Bates, Boston University **Noah Coccaro and Daniel Jurafsky,** University of Colorado at Boulder
Rachel Martin, Johns Hopkins University **Marie Meteer,** BBN Technologies **Klaus Ries,** Carnegie Mellow University
Paul Taylor, University of Edinburgh **Carol Van Ess-Dykema,** Department of Defense

Abstract

We describe an integrated approach for statistical modeling of discourse structure for natural conversational speech. Our model is based on 42 'dialog acts' (e.g., Statement, Question, Backchannel, Agreement, Disagreement, Apology), which were hand-labeled in 1155 conversations from the Switchboard corpus of spontaneous human-to-human telephone speech. We developed several models and algorithms to automatically detect dialog acts from transcribed or automatically recognized words and from prosodic properties of the speech signal, and by using a statistical discourse grammar. All of these components were probabilistic in nature and estimated from data, employing a variety of techniques (hidden Markov models, N-gram language models, maximum entropy estimation, decision tree classifiers, and neural networks). In preliminary studies, we achieved a dialog act labeling accuracy of 65% based on recognized words and prosody, and an accuracy of 72% based on word transcripts. Since humans achieve 84% on this task (with chance performance at 35%) we find these results encouraging.

Introduction

The ability to model and automatically detect discourse structure is essential as we address problems such as *understanding spontaneous dialog* (a meeting summarizer needs to know who said what to whom), *building human-computer dialog systems* (a conversational agent needs to know whether it just got asked a question or ordered to do something), and simple *transcription of conversational speech* (utterances with different discourse function also have very different words). This paper describes an effort to automate the annotation of natural dialog at the level of *dialog acts* (DAs), a shallow first level of analysis that is essential to the tasks mentioned. Table 1 shows a sample of the kind of discourse structure we are modeling and detecting. Each utterance is categorized into one of several utterance types according to syntactic and pragmatic criteria.

Our approach was to build statistical models for various aspects of dialog acts, such as their lexical realizations, prosodic characteristics, and sequence distribution, and to integrate these into a probabilistic DA detector. There are many excellent previous attempts to build predictive, stochastic models of dialog structure (Kita *et al.* 1996; Mast *et al.* 1996; Nagata & Morimoto 1994; Reithinger *et al.* 1996; Suhm & Waibel 1994; Taylor *et al.* 1997; Woszczyna & Waibel 1994; Yamaoka & Iida 1991), and our effort is in many ways inspired by this work. Our project extends this earlier work, particularly in its scale; our models were trained on an order of magnitude more data than any previous system. In addition, whereas previous work has largely dealt with constrained, task-oriented dialog, our focus is on unconstrained, spontaneous conversation. Finally, we believe our approach to model integration, in particular our use of automatically recognized words, to be novel. A more complete account of this work can be found in Jurafsky *et al.* (1997).

The Dialog Act Labeling Task

The data consisted of a substantial portion of the waveforms and corresponding transcripts from the Switchboard corpus of conversational telephone speech (Godfrey, Holliman, & McDaniel 1992) distributed by the Linguistic Data Consortium (LDC). The raw Switchboard data is not segmented in a linguistically consistent way; we therefore made use of a version that had been hand-segmented at the utterance level (Meteer & others 1995). Automatic segmentation of spontaneous speech is an open research problem in its own right (Mast *et al.* 1996; Stolcke & Shriberg 1996), but we decided not to confound the DA detection task with the additional problems introduced by automatic segmentation.

We chose to follow a recent standard for shallow discourse structure annotation, the Dialog Act Markup in Several Layers (DAMSL) tag set, which was recently designed by the natural-language processing community (Core & Allen 1997). We began with this markup system and modified it in several ways to make it more useful for our corpus. The tag set distinguishes 42 mutually exclusive utterance types; Table 2 shows the 10 most frequent categories with examples and rela-

Table 1: A fragment of a labeled switchboard conversation.

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.*

tive frequencies. A detailed description of the labeling system can be found in Jurafsky, Shriberg, & Biasca (1997).

Note that our tag set incorporates both traditional sociolinguistic and discourse-theoretic rhetorical relations/adjacency-pairs as well as some more-form-based labels. Furthermore, the tag set is structured so as to allow labelers to annotate a Switchboard conversation in about 30 minutes, and without having to listen to each utterance. Without these constraints the tag set might have included some finer distinctions, but we felt that this drawback was balanced by the ability to cover a large amount of data.

Labeling was carried out in a three-month period by eight linguistics graduate students at CU Boulder. Inter-labeler agreement was 84%, resulting in a Kappa statistic of 0.80. The Kappa statistic measures agreement normalized for chance; values of 0.8 or higher are considered considered high reliability (Carletta 1996).

A total of 1155 Switchboard conversations were labeled, comprising 205,000 utterances and 1.4 million words. The data was partitioned into a training set of 1115 conversations (1.4M words, 198K utterances), used for estimating the various components of our model, and a test set of 19 conversations (29K words, 4K utterances). Remaining conversations were set aside as future test sets.

Hidden Markov Modeling of Dialog

Our goal is to perform DA classification using a probabilistic framework, giving us a principled approach for combining multiple knowledge sources (using the laws of probability), as well as the ability to derive model parameters automatically from a corpus, using statistical inference techniques.

Given all available evidence E about a conversation, the goal is to find the DA sequence U that has the highest posterior probability $P(U|E)$ given that evidence. Applying Bayes' Rule we get

$$\begin{aligned} U^* &= \operatorname*{argmax}_U P(U|E) \\ &= \operatorname*{argmax}_U \frac{P(U)P(E|U)}{P(E)} \\ &= \operatorname*{argmax}_U P(U)P(E|U) \end{aligned} \quad (1)$$

Here $P(U)$ represents the prior probability of a DA sequence, and $P(E|U)$ is the likelihood of U given the evidence. The likelihood is usually much more straightforward to model than the posterior itself. This has to do with the fact that our models are generative or causal in nature, i.e., they describe how the evidence is produced by the underlying DA sequence U.

Estimating $P(U)$ requires building a probabilistic discourse grammar, i.e., a statistical model of DA sequences. We did so using familiar techniques from language modeling for speech recognition, although the sequenced objects in this case are DA labels rather than words.

Dialog act likelihoods

The computation of likelihoods $P(E|U)$ depends on the types of evidence used. In our experiments we used the following sources of evidence, either alone or in combination:

Transcribed words: The likelihoods used in Eq. 1 are $P(W|U)$, where W refers to the true (hand-transcribed) words spoken in a conversation.

Recognized words: The evidence consists of recognizer acoustics A, and we seek to compute $P(A|U)$. As described later, this involves considering multiple alternative recognized word sequences.

Prosodic features: Evidence is given by the acoustic features F capturing various aspects of pitch, duration, energy, etc., of the speech signal; the associated likelihoods are $P(F|U)$.

To make both the modeling and the search for the best DA sequence feasible, we further require that our likelihood models are *decomposable by utterance*. This means that the likelihood given a complete conversation can be factored into likelihoods given the individ-

Table 2: The 10 most frequent (out of 42) dialog act labels.

Tag	Abbrev	Example	%
Statement-non-opinion	sd	*Me, I'm in the legal department.*	36%
Acknowledge (Backchannel)	b	*Uh-huh.*	19%
Statement-opinion	sv	*I think it's great*	13%
Agree/Accept	aa	*That's exactly it.*	5%
Abandoned or Turn-Exit	%	*So, -/*	5%
Appreciation	ba	*I can imagine.*	2%
Yes-No-Question	qy	*Do you have to have any special training?*	2%
Non-verbal	x	*<Laughter>,<Throat_clearing>*	2%
Yes answers	ny	*Yes.*	1%
Conventional-closing	fc	*Well, it's been nice talking to you.*	1%

ual utterances. We use U_i for the ith DA label in the sequence U, i.e., $U = (U_1, \ldots, U_i, \ldots, U_n)$, where n is the number of utterances in a conversation. In addition, we use E_i for that portion of the evidence that corresponds to the ith utterance, e.g., the words or the prosody of the ith utterance. Decomposability of the likelihood means that

$$P(E|U) = P(E_1|U_1) \cdot \ldots \cdot P(E_n|U_n)$$

Applied to the three types of evidence introduced earlier, it is clear that this assumption is not strictly true. For example, speakers might tend to reuse words found earlier in the conversation, violating the independence of the $P(W_i|U_i)$. Similarly, speakers might adjust their pitch or volume over time, e.g., to the conversation partner, violating the independence of the $P(F_i|U_i)$. As in other areas of statistical modeling, we count on the fact that these violations are small compared to the properties actually modeled, namely, the dependence of E_i on U_i.

Markov modeling

Returning to the prior of DA sequences $P(U)$, it is convenient to make certain independence assumptions here, too. In particular, we assume that the prior distribution of U is Markovian, i.e., that each U_i depends only on a fixed number k of preceding DA labels:

$$P(U_i|U_1, \ldots, U_{i-1}) = P(U_i|U_{i-k}, \ldots, U_{i-1})$$

(k is the order of the Markov process describing U). The N-gram based discourse grammars we used have this property. As described later, $k = 1$ is a very good choice, i.e., conditioning on the DA types more than one removed from the current one does not improve the quality of the model by much.

The importance of the Markov assumption for the discourse grammar is that we can now view the whole system of discourse grammar and local utterance-based likelihoods as a kth-order *hidden Markov model* (HMM) (Rabiner & Juang 1986). The HMM states correspond to DAs, observations correspond to utterances, transition probabilities are given by the discourse grammar, and observation probabilities are given by the local likelihoods $P(E_i|U_i)$. This allows us to use efficient dynamic programming algorithms to compute the relevant aspects of the model, such as

- the most probable DA sequence (the Viterbi algorithm)
- the posterior probability of various DAs for a given utterance, after considering all the evidence (the forward-backward algorithm)

Dialog act decoding

The Viterbi algorithm for HMMs finds the globally most probable state sequence. When applied to a discourse model with locally decomposable likelihoods and Markovian discourse grammar, it will therefore find precisely the DA sequence with the highest posterior probability:

$$U^* = \operatorname*{argmax}_U P(U|E)$$

The combination of likelihood and prior modeling, HMMs, and Viterbi decoding is fundamentally the same as the standard probabilistic approaches to speech recognition (Bahl, Jelinek, & Mercer 1983) and tagging (Church 1988). It maximizes the probability of getting the *entire* DA sequence correct, but it does not necessarily find the DA sequence that has the most DA labels correct (Stolcke, Konig, & Weintraub 1997). To minimize the overall utterance labeling error, we need to maximize the probability of getting each DA label correct individually, i.e., we need to maximize $P(U_i|E)$ for each $i = 1, \ldots, n$. We can compute the per-utterance posterior DA probabilities by summing:

$$P(u|E) = \sum_{U_i = u} P(U|E)$$

where the summation is over all sequences U whose ith element matches the label in question. The summation is efficiently carried out by the forward-backward algorithm for HMMs.

For 0th-order (unigram) discourse grammars, Viterbi decoding and forward/backward decoding always yield the same results. However, for higher-order discourse grammars we found that forward-backward

Table 3: Perplexities of dialog acts with and without turn information.

Discourse grammar	$P(U)$	$P(U,T)$	$P(U\|T)$
None	42	84	42
Unigram	11.0	18.5	9.0
Bigram	7.9	10.4	5.1
Trigram	7.5	9.8	4.8

decoding consistently gives slightly (up to 1% absolute) better accuracies, as expected. Therefore, we used this method throughout.

Discourse Grammars

The statistical discourse grammar models the prior probabilities $P(U)$ of DA sequences. In the case of conversations for which the identities of the speakers are known (as in Switchboard), the discourse grammar should also model turn-taking behavior. A straightforward approach is to model sequences of pairs (U_i, T_i) where U_i is the DA label and T_i represents the speaker. We are not trying to model speaker idiosyncrasies, so conversants are arbitrarily identified as **A** or **B**, and the model is made symmetric with respect to the choice of sides (e.g., by replicating the training sequences with sides switched). Our discourse grammars thus had a vocabulary of $42 \times 2 = 84$ labels, plus tags for the beginning and end of conversations.

N-gram discourse models

A computationally convenient type of discourse grammar is an N-gram model based on DA tags, as it allows efficient decoding in the HMM framework. We trained standard backoff N-gram models (Katz 1987), using the frequency smoothing approach of Witten & Bell (1991). Models of various orders were compared by their perplexities, i.e., the average number of choices the model predicts for each tag, conditioned on the preceding tags.

Table 3 shows perplexities for three types of models: $P(U)$, the DAs alone; $P(U,T)$, the combined DA/speaker ID sequence; and $P(U|T)$, the DAs conditioned on known speaker IDs (appropriate for the Switchboard task). As expected, we see an improvement (decreasing perplexities) for increasing N-gram order. However, the incremental gain of a trigram is small, and higher-order models did not prove useful. Comparing $P(U)$ and $P(U|T)$, we see that speaker identity adds substantial information, especially for higher-order models.

Other discourse models

We also investigated non-N-gram discourse models, based on various language modeling techniques known from speech recognition. One motivation for alternative models is that N-grams enforce a one-dimensional representation on DA sequences, whereas we saw above that the event space is really a multidimensional event (DA label and speaker labels). Another motivation is that N-grams fail to model long-distance dependencies, such as the fact that speakers may tend to repeat certain DAs or patterns throughout the conversation.

The first alternative approach was a standard *cache model* (Kuhn & de Mori 1990), which boosts the probabilities of previously observed unigrams and bigrams, on the theory that tokens tend to repeat themselves over longer distances. However, this does not seem to be true for DA sequences in our corpus, as the cache model showed no improvement over the standard N-gram.

Second, we built a discourse grammar that incorporated constraints on DA sequences in a non-hierarchical way, using *maximum entropy* (ME) estimation (Rosenfeld 1996). The model was designed so that the current DA label was constrained by features such as unigram statistics, the previous DA and the DA once removed, DAs occurring within a window in the past, and whether the previous utterance was by the same speaker. We found, however, that an ME model using N-gram constraints performed only slightly better than a corresponding backoff N-gram, and that adding the additional constraints did not improve relative to the trigram model. We conclude that DA sequences are mostly characterized by local interactions, and thus modeled well by low-order N-gram statistics.

Dialog Act Detection Using Words

DA classification using words is based on the observation that different DAs use distinctive word strings. For example, 92.4% of the "uh_huh"-s occur in **Backchannels**, and 88.4% of the trigrams "<start> do you" occur in **Yes-No-Questions**.

Detection from true words

Assuming that the true (hand-transcribed) words of utterances are given as evidence, we can compute word-based likelihoods $P(W|U)$ in a straightforward way, by computing a statistical language model for each of the 42 DAs. All DAs of a particular type found in the training corpus were pooled and a DA-specific trigram model was built using standard techniques (Katz-backoff with Witten-Bell discounting).

Detection from recognized words

For fully automatic DA detection, the above approach is only a partial solution, since we are not yet able to recognize words in spontaneous speech with perfect accuracy. We modify the likelihood approach to work with the acoustic information A (waveforms) available to a speech recognizer. We compute $P(A|U)$ by decomposing it into an acoustic likelihood $P(A|W)$ and a word-based likelihood $P(W|U)$, and summing over

Table 4: DA detection accuracies (in %) from transcribed and recognized words (chance = 35%).

Discourse Grammar	True	Recognized
None	54.3	42.8
Unigram	68.1	61.9
Bigram	70.6	64.6
Trigram	71.9	64.9

all word sequences:

$$P(A|U) = \sum_W P(A|W,U)P(W|U)$$
$$= \sum_W P(A|W)P(W|U)$$

The second line is justified under the assumption that the recognizer acoustics (typically, cepstral coefficients) are invariant to DA type once the words are fixed.[1]

The acoustic likelihoods $P(A|W)$ correspond to the acoustic scores the recognizer outputs for every hypothesized word sequence W. The summation over all W must be approximated; we did so by summing over the 2500 best hypotheses.

Results

Table 4 shows DA detection accuracies obtained by combining the word- and recognizer-based likelihoods with the N-gram discourse grammars described earlier. The best accuracy obtained from transcribed words, 72%, is encouraging given a comparable human performance of 84%. We observe about a 7% absolute reduction when using recognizer words; this is remarkable considering that the speech recognizer used had a word error rate of 41% on the test set.

Dialog Act Detection Using Prosody

We also investigated prosodic information, i.e., information independent of the words as well as the standard recognizer acoustics. Prosody is important for DA recognition for two reasons. One the one hand, as we saw earlier, word-based detection suffers from recognition errors. Second, some utterances are inherently ambiguous based on words alone. For example, some **Yes-No-Questions** have identical word sequences as **Statements**, but can often be distinguished by their final F0 rise.

Prosodic features

Prosodic DA classification was based on a large set of features computed automatically from the waveform, without reference to word or phone information.

[1]This is another approximation in our modeling. For example, a word pronunciation may change as a result of different emphasis placed on a word.

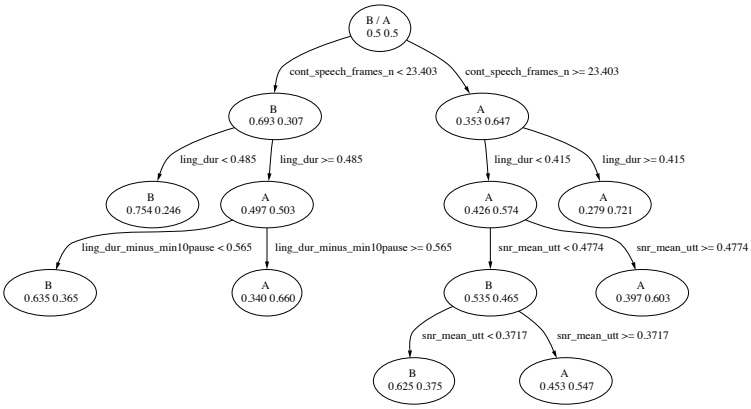

Figure 1: Decision tree for the classification of Backchannels (B) and Agreements (A). Each node is labeled with the majority class for that node, as well as the posterior probabilities of the two classes.

The features can be broadly grouped as referring to duration (e.g., utterance duration, with and without pauses), pauses (e.g., total and mean of non-speech regions exceeding 100 ms), pitch (e.g., mean and range of F0 over utterance, slope of F0 regression line), energy (e.g., mean and range of RMS energy, same for signal-to-noise ratio), speaking rate (based on the "enrate" measure of Morgan, Fosler, & Mirghafori (1997)), and gender (of both speaker and listener). Where appropriate, we included both raw features and values normalized by utterance and/or conversation. We also included features that are output by the prosodic event detector of Taylor *et al.* (1997) (e.g., the number of pitch accents in the utterance). A complete discussion of the features used can be found in Shriberg *et al.* (1997).

Prosodic decision trees

For our prosodic classifiers, we used CART-style decision trees (Breiman *et al.* 1983). Decision trees allow combination of discrete and continuous features, and can be inspected to gain an understanding of the role of different features and feature combinations.

To illustrate one area in which prosody could aid our classification task, we applied trees to distinctions known to be ambiguous from words alone. One frequent example in our corpus was the distinction between **Backchannels** and **Agreements** (see Table 2), which share terms such as "Right" and "Yeah." As shown in Figure 1, a prosodic tree trained on this distinction revealed that agreements have consistently longer durations and greater energy (as reflected by the SNR measure) than do backchannels.

The HMM framework requires that we compute prosodic likelihoods of the form $P(F_i|U_i)$ for each utterance U_i and associated prosodic feature values F_i. We have the apparent difficulty that decision trees give

Table 5: DA detection using prosody (chance = 35%).

Discourse Grammar	Accuracy (%)
None	38.9
Unigram	48.3
Bigram	50.2

estimates for the posterior probabilities, $P(U_i|F_i)$. The problem can be overcome by applying Bayes' Rule locally:

$$P(F_i|U_i) = P(F_i)\frac{P(U_i|F_i)}{P(U_i)} \propto \frac{P(U_i|F_i)}{P(U_i)}$$

A quantity proportional to the required likelihood can therefore be obtained by either dividing the posterior tree probability by the prior $P(U_i)$, or by training the tree on a uniform prior distribution of DA types. We chose the second approach, downsampling our training data to equate DA proportions.

Results

As a preliminary experiment to test the integration of prosody with other knowledge sources, we trained a single tree to discriminate among the five most frequent dialog acts (Statement, Backchannel, Opinion, Agreement, and Abandoned, totaling 78% of the data) and an "Other" category. The probability in the "Other" category was split uniformly among all the types in that category. Results for this "Top-5" tree are shown in Table 5. As shown, the tree performs significantly better than chance, but not as well as the word-based methods (see Table 4).

Neural network classifiers

Although we chose to use decision trees as prosodic classifiers for their relative ease of inspection, we might have used any suitable probabilistic classifier, i.e., any model that estimates the posterior probabilities of DAs given the prosodic features. We conducted preliminary experiments to assess how neural networks compare to decision trees for the type of data studied here. Neural networks are worth investigating since they offer potential advantages over decision trees. They can learn decision surfaces that lie at an angle to the axes of the input feature space, unlike standard CART trees which always split continuous features on one dimension at a time. The response function of neural networks is continuous (smooth) at the decision boundaries, allowing them to avoid hard decisions and the complete fragmentation of data associated with decision tree questions. Most important, neural networks with hidden units can learn new features that combine multiple input features. Results from preliminary experiments on the Top-5 classification task showed that a softmax network (Bridle 1990) without hidden units resulted in a slight improvement over a decision tree on the same task. The fact that hidden units did not afford an advantage here indicates that complex combinations of features (as far as the network could learn them) do not better predict DAs for the task than linear combinations of our input features. This further justifies our choice of decision trees for this task, although we should not discount other approaches in future studies.

Using Multiple Knowledge Sources

As mentioned earlier, we expect improved performance from combining word and prosodic information. Combining these knowledge sources requires estimating a combined likelihood $P(A_i, F_i|U_i)$ for each utterance. The simplest approach is to assume that the two types of acoustic observations (recognizer acoustics and prosodic features) are approximately conditionally independent once U_i is given:

$$\begin{aligned} P(A_i, F_i|U_i) &= P(A_i|U_i)P(F_i|A_i, U_i) \\ &\approx P(A_i|U_i)P(F_i|U_i) \end{aligned}$$

Since the recognizer acoustics are modeled by way of their dependence on words, it is particularly important to avoid using prosodic features that are directly correlated with word-identities, or features that are also modeled by the discourse grammars, such as utterance position relative to turn changes.

Results

For the one experiment we conducted using this approach, we combined the acoustic N-best likelihoods from our experiment with recognized words with the Top-5 tree classifier mentioned earlier. Results are summarized in Table 6.

Table 6: Combined utterance detection accuracies (chance = 35%).

Discourse Grammar	Accuracy (%)		
	Prosody	Recognizer	Combined
None	38.9	42.8	56.5
Unigram	48.3	61.9	62.6
Bigram	50.2	64.6	65.0

As shown, the combined classifier presents a slight improvement over the recognizer-based classifier. The experiment without discourse grammar indicates that the combined evidence is considerably stronger than either knowledge source alone, yet this improvement seems to be made largely redundant by the use of priors and the discourse grammar. For example, the ambiguity between Yes-No-Questions and statements where prosody is expected to help can also be removed by examining the context of the utterance (e.g., noticing that the following utterance is a yes/no answer).

Table 7: Accuracy (in %) for individual and combined models for three subtasks, using uniform priors (chance = 50%).

Knowledge Source	True words	Recog. words
Questions/Statements		
prosody only	75.97	75.97
words only	85.85	75.43
words+prosody	87.58	79.76
Agreements/Backchannels		
prosody only	72.88	72.88
words only	80.99	78.22
words+prosody	84.74	81.70

Focussed classifications

To gain a better understanding of the potential for prosodic DA detection independent of the effects of discourse grammar and the skewed DA distribution in Switchboard, we also examined several binary DA classification tasks. The choice of tasks was motivated by an analysis of confusions committed by a purely word-based DA detector, which tends to mistake Questions for Statements, and Backchannels for Agreements (and vice versa). We tested a prosodic classifier, and word-based classifier (with both transcribed and recognized words), and a combined classifier on these three tasks, downsampling the DA distribution to equate the class sizes in each case. Chance performance in all three experiments is therefore 50%. Results are summarized in Table 7.

As shown, the combined classifier was consistently more accurate than the classifier using words alone. Although the gain in accuracy was not statistically significant for the small recognizer test set because of a lack of power, replication for a larger test set showed the gain to be highly significant for both subtasks by a Sign test, $p < .001$ and $p < .0001$, respectively. Across these as well as additional subtasks, the relative advantage adding prosody was larger for recognized than for true words, suggesting that prosody is particularly helpful when word information is not perfect.

Feature Usage

Feature analyses, conducted by systematically leaving out feature types and rebuilding trees, revealed that although canonical features (such as F0 for question detection) were important, other less obvious features (e.g., duration and speaking rate for the same task) were also heavily used. Gender features were not used, suggesting that feature normalizations (especially F0) were appropriate, and that gender-independent modeling is feasible for these tasks. Overall, there was a high degree of correlation among features such that if certain features were removed, others could compensate to retain accuracy. Nevertheless, the features allowing best classification were dependent on the subtask, suggesting that a prosodic classifier should use as many different feature types as possible for optimal coverage across tasks.

Conclusions

We have developed an integrated probabilistic approach to dialog act classification on a large spontaneous speech corpus. The approach combines models for lexical and prosodic realizations of DAs, as well as a statistical discourse grammar. All components of the model are automatically trained, and are thus applicable to other domains for which labeled data is available. Detection accuracies achieved so far are highly encouraging, relative to the inherent difficulty of the task as measured by human labeler performance. We investigated several modeling alternatives for the components of the model (backoff N-grams and maximum entropy models for discourse grammars, decision trees and neural networks for prosodic classification). We found performance largely independent of these choices, indicating on the one hand that our current system does about as well as possible given current modeling techniques and the inherent difficulty of the task and our limited representation of it. On the other hand, to improve performance we will have to revisit our independence assumptions, as well as examine additional knowledge sources.

Future Work

For discourse and dialog modeling, we plan to try alternative approaches to encode the temporal sequencing of utterances. For example, we are currently not modeling the fact that utterances by the two speakers may actually overlap (e.g., backchannels interrupt an ongoing utterance). In addition, we should model more of the non-local aspects of discourse structure, despite our negative results so far. For example, a context-free discourse grammar could potentially account for the nested structures proposed in Grosz & Sidner (1986).

Word-based DA discrimination has obvious parallels to topic spotting and message classification, and we should explore techniques developed in that paradigm, such as keyword-based detectors (Rose, Chang, & Lippmann 1991). For prosodic DA detection, we are studying the use of multiple trees, both to cascade classifiers trained on subtasks, and to combine parallel classifiers using a disjoint subset of features, which we believe will increase robustness.

The integration of knowledge sources is especially promising, since we are currently making fairly severe independence assumptions here. Therefore our eventual goal would be a DA classifier that directly integrates discourse grammar, word information, and prosody. For example, it should be feasible to train a prosodic decision tree that takes the discourse context as one of its inputs. Such a model would subsume the discourse grammar, and is potentially able to capture

interactions between the context and prosody of the current utterance, which are currently assumed independent (given the current DA). A further idea along these lines is to make word knowledge directly available to a posterior probability estimator, allowing it to model correlations of words and prosody.

Acknowledgments

This research was carried out as part of the 1997 Workshop on Innovative Techniques in LVCSR at the Center for Speech and Language Processing at Johns Hopkins University. Additional support came from the NSF via grants IRI-9619921 and IRI-9314967. Special thanks go to the discourse labelers at CU Boulder and the intonation labelers at the University of Edinburgh.

References

Bahl, L. R.; Jelinek, F.; and Mercer, R. L. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 5(2):179–190.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1983. *Classification and Regression Trees*. Pacific Grove, California: Wadsworth & Brooks.

Bridle, J. 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Soulie, F., and J.Herault., eds., *Neurocomputing: Algorithms, Architectures and Applications*. Berlin: Springer. 227–236.

Carletta, J. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22(2):249–254.

Church, K. W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, 136–143.

Core, M., and Allen, J. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Godfrey, J.; Holliman, E.; and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP*, 517–520.

Grosz, B., and Sidner, C. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

Jurafsky, D.; Bates, R.; Coccaro, N.; Martin, R.; Meteer, M.; Ries, K.; Shriberg, E.; Stolcke, A.; Taylor, P.; and Van Ess-Dykema, C. 1997. Switchboard discourse language modeling project report. Technical report, Center for Speech and Language Processing, Johns Hopkins University, Baltimore.

Jurafsky, D.; Shriberg, E.; and Biasca, D. 1997. Switchboard-DAMSL Labeling Project Coder's Manual. http://stripe.colorado.edu/~jurafsky/manual.august1.html.

Katz, S. M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Trans. Acoustics, Speech and Signal Processing* 35(3):400–401.

Kita, K.; Fukui, Y.; Nagata, M.; and Morimoto, T. 1996. Automatic acquisition of probabilistic dialogue models. In *Proc. ICSLP*, 196–199.

Kuhn, R., and de Mori, R. 1990. A cache-based natural language model for speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence* 12(6):570–583.

Mast, M.; Kompe, R.; Harbeck, S.; Kießling, A.; Niemann, H.; ; and Nöth, E. 1996. Dialog act classification with the help of prosody. In *Proc. ICSLP*, 1728–1731.

Meteer, M., et al. 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, N.; Fosler, E.; and Mirghafori, N. 1997. Speech recognition using on-line estimation of speaking rate. In *Proc. EUROSPEECH*.

Nagata, M., and Morimoto, T. 1994. First steps toward statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication* 15:193–203.

Rabiner, L. R., and Juang, B. H. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 3(1):4–16.

Reithinger, N.; Engel, R.; Kipp, M.; and Klesen, M. 1996. Predicting dialogue acts for a speech-to-speech translation system. In *Proc. ICSLP*, 654–657.

Rose, R. C.; Chang, E. I.; and Lippmann, R. P. 1991. Techniques for information retrieval from voice messages. In *Proc. ICASSP*, 317–320.

Rosenfeld, R. 1996. A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language* 10:187–228.

Shriberg, E.; Bates, R.; Taylor, P.; Stolcke, A.; Jurafsky, D.; Ries, K.; Coccaro, N.; Martin, R.; Meteer, M.; and Van Ess-Dykema, C. 1997. Can prosody aid the automatic classification of dialog acts in conversational speech? Submitted.

Stolcke, A., and Shriberg, E. 1996. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP*, 1005–1008.

Stolcke, A.; Konig, Y.; and Weintraub, M. 1997. Explicit word error minimization in N-best list rescoring. In *Proc. EUROSPEECH*, volume 1, 163–166.

Suhm, B., and Waibel, A. 1994. Toward better language models for spontaneous speech. In *Proc. ICLSP*, 831–834.

Taylor, P.; King, S.; Isard, S.; Wright, H.; and Kowtko, J. 1997. Using intonation to constrain language models in speech recognition. In *Proc. EUROSPEECH*, 2763–2766.

Witten, I. H., and Bell, T. C. 1991. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Information Theory* 37(4):1085–1094.

Woszczyna, M., and Waibel, A. 1994. Inferring linguistic structure in spoken language. In *Proc. ICSLP*, 847–850.

Yamaoka, T., and Iida, H. 1991. Dialogue interpretation model and its application to next utterance prediction for spoken language processing. In *Proc. EUROSPEECH*, 849–852.