Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 23 (2009) 407–422
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Combining lexical, syntactic and prosodic cues for improved online dialog act tagging

Vivek Kumar Rangarajan Sridhar [a,*,1], Srinivas Bangalore [b], Shrikanth Narayanan [a]

[a] *Ming Hsieh Department of Electrical Engineering, University of Southern California, 3740 McClilntock Avenue, Room EEB430, Los Angeles, CA 90089-2564, United States*
[b] *AT&T Labs – Research 180 Park Avenue, Florham Park, NJ 07932, United States*

Received 23 January 2008; received in revised form 6 October 2008; accepted 12 December 2008
Available online 25 December 2008

Abstract

Prosody is an important cue for identifying dialog acts. In this paper, we show that modeling the sequence of acoustic–prosodic values as *n*-gram features with a maximum entropy model for dialog act (DA) tagging can perform better than conventional approaches that use coarse representation of the prosodic contour through summative statistics of the prosodic contour. The proposed scheme for exploiting prosody results in an absolute improvement of 8.7% over the use of most other widely used representations of acoustic correlates of prosody. The proposed scheme is discriminative and exploits context in the form of lexical, syntactic and prosodic cues from preceding discourse segments. Such a decoding scheme facilitates online DA tagging and offers robustness in the decoding process, unlike greedy decoding schemes that can potentially propagate errors. Our approach is different from traditional DA systems that use the entire conversation for offline dialog act decoding with the aid of a discourse model. In contrast, we use only *static* features and approximate the previous dialog act tags in terms of lexical, syntactic and prosodic information extracted from previous utterances. Experiments on the Switchboard-DAMSL corpus, using only lexical, syntactic and prosodic cues from three previous utterances, yield a DA tagging accuracy of 72% compared to the best case scenario with accurate knowledge of previous DA tags (oracle), which results in 74% accuracy.

Keywords: Dialog act tagging; Prosodic cues; Acoustic correlates of prosody; Maximum entropy modeling; Discourse context

1. Introduction

In both human-to-human and human–computer speech communication, identifying whether an utterance is a statement, question, greeting, etc., is integral to producing, sustaining and understanding natural dialogs.

[*] Corresponding author. Tel.: +1 213 740 3477.
E-mail addresses: vrangara@usc.edu (V.K. Rangarajan Sridhar), srini@research.att.com (S. Bangalore), shri@sipi.usc.edu (S. Narayanan).
[1] He is now with BBN Technologies, Cambridge, MA 02138 (vsridhar@bbn.com).

0885-2308/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.12.001

Dialog act tags (Austin, 1962) are labels that are used to represent these surface level communicative acts in a conversation or dialog. While they may not provide a deep understanding of discourse structure, dialog acts (DAs) can serve as intermediate representations that can be useful in several speech and language processing applications. For example, in human–machine dialogs, constraining automatic speech recognition hypotheses by using a model of likely DAs to be expected at a dialog turn has been shown to improve the recognition accuracy (Stolcke et al., 2000; Taylor et al., 2000). Dialog acts have also found to be useful in spoken language understanding (Shriberg et al., 1998) and, more recently, in the annotation of archived conversations and meetings (Ang et al., 2005; Zimmermann et al., 2005), which in turn can help improve speech summarization (Murray et al., 2006) and retrieval. Incorporating DAs in speech-to-speech (s2s) translation (Lavie et al., 1996; Reithinger et al., 1996) was useful in the resolution of ambiguous communication.

Conceptually, the process of designing an automatic DA prediction system can be seen as comprising two steps:

- Identifying the lexical, syntactic and acoustic cues that are most useful in distinguishing among the various DAs.
- Combining the multiple cues in an algorithmic framework to implement their accurate recognition.

Methods for automatic cue-based identification of dialog acts typically exploit multiple knowledge sources in the form of lexical (Jurafsky et al., 1998; Stolcke et al., 2000), syntactic (Bangalore et al., 2006), prosodic (Shriberg et al., 1998; Taylor et al., 2000) and discourse structure (Jurafsky et al., 1997) cues. These cues have been modeled using a variety of methods including Hidden Markov models (Jurafsky et al., 1998), neural networks (Ries, 1999), fuzzy systems (Wu et al., 2002) and maximum entropy models (Bangalore et al., 2006; Rangarajan Sridhar et al., 2007a). Conventional dialog act tagging systems rely on the words and syntax of utterances (Hirschberg and Litman, 1993). However, in most applications that require transcriptions from an automatic speech recognizer, the lexical information obtained is typically noisy due to recognition errors. Moreover, some utterances are inherently ambiguous based on just lexical information. For example, an utterance such as “okay” can be used in the context of a statement, question or acknowledgment (Gravano et al., 2007).

While lexical information is a strong cue to DA identity, the prosodic information contained in the speech signal can provide a rich source of complementary information. In languages such as English and Spanish, discourse functions are characterized by distinct intonation patterns (Bolinger, 1978; Cruttenden, 1989). These intonation patterns can either be final fundamental frequency (f0) contour movements or characteristic global shapes of the pitch contour. For example, *yes–no* questions in English typically show a rising f0 contour at the end and *wh*- questions typically show a final falling pitch. Modeling the intonation pattern can thus be useful in discriminating sentence types. Previous work on exploiting intonation for DA tagging has mainly been through the use of representative statistics of the raw or normalized pitch contour, duration and energy such as mean, standard deviation, slope, etc. (Stolcke et al., 2000; Shriberg et al., 1998). However, these acoustic correlates of prosody provide only a coarse summary of the macroscopic prosodic contour and hence may not exploit the prosodic profile fully. In this work, we model the prosodic contour by extracting *n*-gram features from the acoustic–prosodic sequence. This *n*-gram feature representation is shown to yield better dialog act recognition accuracy compared to other methods that use summative statistics of acoustic–prosodic features. Further details of prosodic representations are provided in Section 6.

We also present a discriminatively trained maximum entropy modeling framework using the *n*-gram prosodic features that is suitable for online classification of DAs. Traditional DA taggers typically combine the lexical and prosodic features in a HMM framework with a Markovian discourse grammar (Stolcke et al., 2000; Jurafsky et al., 1998). The HMM representation facilitates optimal decoding through the Viterbi algorithm. However, such an approach limits DA classification to offline processing, as it uses the entire conversation during decoding. Even though this drawback can be overcome by using a greedy decoding approach, the resultant decoding is sensitive to noisy input and may cause error propagation. In contrast, our approach uses contextual features captured in the form of only lexical and prosodic cues from previous utterances. Such a scheme is computationally inexpensive and facilitates robust online decoding that can be performed alongside automatic speech recognition. We evaluate our proposed approach through experiments on the Maptask (Carletta et al., 1997) and Switchboard-DAMSL (Jurafsky et al., 1998) corpora.

2. Maximum entropy model for dialog act tagging

We use a maximum entropy sequence tagging model for the purpose of automatic DA tagging. We model the prediction problem as a classification task: given a sequence of utterances u_i in a dialog $U = u_1, u_2, \ldots, u_n$ and a dialog act vocabulary $(d_i \epsilon \mathscr{D}, | \mathscr{D} | = K)$, we need to predict the best dialog act sequence $D^* = d_1, d_2, \ldots, d_n$

$$D^* = \underset{D}{\operatorname{argmax}}\, P(D|U) = \underset{d_1,\ldots,d_n}{\operatorname{argmax}}\, P(d_1, \ldots, d_n | u_1, \ldots, u_n). \tag{1}$$

We approximate the sequence level global classification problem, using conditional independence assumptions, to a product of local classification problems as shown in Eq. (3). The classifier is then used to assign to each word a dialog act label conditioned on a vector of local contextual feature vectors comprising the lexical, syntactic and acoustic information

$$D^* = \underset{D}{\operatorname{argmax}}\, P(D|U), \tag{2}$$

$$\approx \underset{D}{\operatorname{argmax}} \prod_{i=1}^{n} P(d_i | \mathbf{\Phi}(u_{i-k}, \cdots, u_{i+l})) \tag{3}$$

$$= \underset{D}{\operatorname{argmax}} \prod_{i=1}^{n} P(d_i | \mathbf{\Phi}(W_{i-k}, \cdots, W_{i+l}, S_{i-k}, \cdots, S_{i+l}, A_{i-k}, \cdots, A_{i+l})) \tag{4}$$

where W_i is the word sequence, S_i is the syntactic feature sequence and A_i, the acoustic–prosodic observation belonging to utterances u_i. The variables l and k denote the right and left context, respectively. $\mathbf{\Phi}(W_{i-k}, \ldots, W_{i+l}, S_{i-k}, \ldots, S_{i+l}, A_{i-k}, \ldots, A_{i+l})$ is shortened to $\mathbf{\Phi}$ in the rest of the section.

To estimate the conditional distribution $P(d \mid \mathbf{\Phi})$ we use the general technique of choosing the maximum entropy (maxent) distribution that estimates the average of each feature over the training data (Berger et al., 1996). This can be written in terms of the Gibbs distribution parameterized with weights $\boldsymbol{\lambda}_m$, where m ranges over the label set and K is the size of the dialog act vocabulary. Hence,

$$P(d|\mathbf{\Phi}) = \frac{e^{\boldsymbol{\lambda}_d \cdot \mathbf{\Phi}}}{\sum_{m=1}^{K} e^{\boldsymbol{\lambda}_m \cdot \mathbf{\Phi}}} \tag{5}$$

To find the global maximum of the concave function in Eq. (5), we use Sequential L1-Regularized Maxent algorithm (SL1-Max) (Dudik et al., 2004). Compared to Iterative Scaling (IS) and gradient descent procedures, this algorithm results in faster convergence and provides L1-regularization as well as efficient heuristics to estimate the regularization meta-parameters. We use the machine learning toolkit LLAMA (Haffner, 2006) to estimate the conditional distribution using maxent. LLAMA encodes multiclass maxent as binary maxent to increase the training speed and to scale the method to large data sets. We use here K one-versus-other binary classifiers. Each output label d is projected onto a bit string, with components $b_j(d)$. The probability of each component is estimated independently:

$$P(b_j(d)|\mathbf{\Phi}) = 1 - P(\bar{b}_j(d)|\mathbf{\Phi}) = \frac{e^{\boldsymbol{\lambda}_j \cdot \mathbf{\Phi}}}{e^{\boldsymbol{\lambda}_j \cdot \mathbf{\Phi}}} + e^{\boldsymbol{\lambda}_{\bar{j}} \cdot \mathbf{\Phi}} = \frac{1}{1 + e^{-(\boldsymbol{\lambda}_j - \boldsymbol{\lambda}_{\bar{j}}) \cdot \mathbf{\Phi}}}, \tag{6}$$

where $\boldsymbol{\lambda}_{\bar{j}}$ is the parameter vector for $\bar{b}_j(d)$.

Assuming the bit vector components to be independent, we have,

$$P(d|\mathbf{\Phi}) = \prod_{j=1}^{K} P(b_j(d)|\mathbf{\Phi}). \tag{7}$$

Therefore, we can decouple the likelihoods and train the classifiers independently. In this work, we use the simplest and most commonly studied code, consisting of K one-versus-others binary components. The independence assumption of the bit vector components states that the output labels or classes are independent.

3. Data

The Maptask (Carletta et al., 1997) and Switchboard-DAMSL (Jurafsky et al., 1998) corpora have been extensively used for dialog act tagging studies. Maptask (Carletta et al., 1997) is a cooperative dialog task involving two participants. The two speakers, *instruction giver* and *instruction follower* are engaged in a dialog with the goal of reproducing the *instruction giver's* route on the *instruction follower's* map. The original dataset was slightly modified for the experiments of the present study. The raw move information was augmented with the speaker information while the non-verbal content (e.g., laughs, background noise) was removed. The Maptask tagging scheme has 12 unique moves; augmented with speaker information this results in 24 labels. The 12 moves in the corpus are: instruct, explain, check, align, query-yn, query-w, acknowledge, reply-y, reply-n, clarify and ready. The corpus consists of 128 dialogs and 26181 utterances. The inter-labeler agreement measured using the kappa statistic (κ) is 0.83. We used ten-fold cross validation for testing.

The Switchboard-DAMSL (SWBD-DAMSL) corpus consists of 1155 dialogs and 218,898 utterances from the Switchboard corpus of telephone conversations, tagged with discourse labels from a shallow discourse tag set. The original tag set of 375 unique tags was clustered to obtain 42 dialog act tags that distinguish mutually exclusive utterance types (Jurafsky et al., 1998). The inter-labeler agreement for this 42-label tag set is 84% ($\kappa = 0.80$), with the labeling performed at the utterance level. In our experiments, we used a set of 173 dialogs, selected at random for testing. The test set consisted of 29869 discourse segments. The experiments were performed on the 42 tag vocabulary as well as a simplified tag set consisting of 7 tags. We grouped the 42 tags into 7 disjoint classes, based on the frequency of the classes and grouped the remaining classes into an”other” category constituting less than 3% of the entire data. This grouping is similar to that presented in Shriberg et al. (1998). Such a simplified grouping is more generic and hence useful in speech applications that require only a coarse level of DA representation. It can also offer insights into common misclassifications encountered in the DA system. Fig. 1 shows the distribution of the simplified tag set in the Switchboard-DAMSL corpus. Statements are the most frequent (more than 50%) tags, followed by acknowledgements, abandoned or incomplete utterances and agreements. Questions and appreciations account for roughly 6% and 4% of the total utterances. In the next section, we describe the maximum entropy modeling framework that is used for automatic DA identification in the rest of the paper.

4. Features for dialog act classification

In this section, we describe the lexical, syntactic and prosodic cues used with the proposed maximum entropy modeling framework for DA tagging. The lexical, syntactic and prosodic cues extracted from the utterance text and speech signal are encoded as *n*-gram features and used as input to the maximum entropy

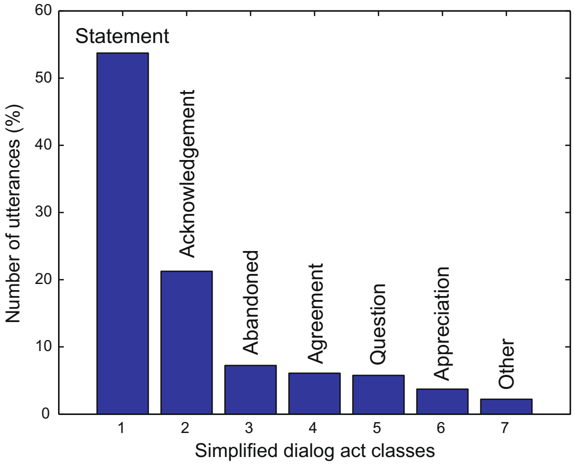
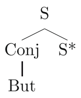
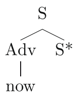
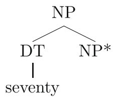
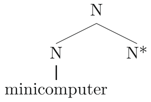
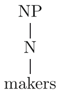
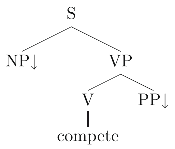
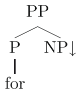
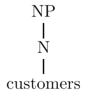


Fig. 1. The distribution of utterances by dialog act tag category in the Switchboard-DAMSL corpus.

Table 1
Illustration of POS tags and supertags generated for a sample utterance.

Words	But	now	seventy	minicomputer	makers	compete	for	customers
POS tags	CC	RB	NN	NN	NNS	VBP	IN	NN
Supertags								

framework. We only use features that are derived from the local context of the text being tagged, referred to as *static* features here on. One would have to perform a Viterbi search if the preceding prediction context (dialog act history) were to be used. Using static features is especially suitable for performing dialog act tagging in lockstep with automatic speech recognition, as the prediction can be performed incrementally instead of waiting for the entire utterance or dialog to be decoded. This is explained in more detail in Section 8.

4.1. Lexical and syntactic features

The lexical features used in our modeling framework are simply the words in a given utterance. We tag the utterances with part-of-speech tags using the AT&T POS tagger. The POS inventory is the same as the Penn treebank which includes 47 POS tags: 22 open class categories, 14 closed class categories and 11 punctuation labels.

In addition to the POS tags, we also annotate the utterance with Supertags (Bangalore and Joshi, 1999). Supertags encapsulate predicate-argument information in a local structure. They are the elementary trees of Tree-Adjoining Grammars (TAGs) (Joshi and Schabes, 1996). Similar to part-of-speech tags, supertags are associated with each word of an utterance, but provide much richer information than part-of-speech tags, as illustrated in the example in Table 1. Supertags can be composed with each other using substitution and adjunction operations (Joshi and Schabes, 1996) to derive the predicate-argument structure of an utterance.

There are two methods for creating a set of supertags. One approach is through the creation of a wide coverage English grammar in the lexicalized tree-adjoining grammar formalism, called XTAG (XTAG, 2001), wherein supertags are the resulting elementary structures. An alternate method for creating supertags is to employ rules that decompose the annotated parse of a sentence in Penn Treebank into its elementary trees (Chen and Vijay-Shanker, 2000; Xia et al., 2000). This second method for extracting supertags results in a larger set of supertags. For the experiments presented in this paper, we employ a set of 4726 supertags extracted from the Penn Treebank.

In addition to the lexical and syntactic cues, we also use categorical prosody labels predicted from our previously developed maximum entropy automatic prosody labeler (Rangarajan Sridhar et al., 2006; Rangarajan Sridhar et al., 2007b) to tag the utterances with prosodic labels. The prosody labeler uses lexical (words) and syntactic (parts-of-speech tags and supertags) information to predict binary pitch accent (**accent, none**) and boundary tone (**btone, none**) labels for each word (see Fig. 2). Our prosody labeler was trained on the entire

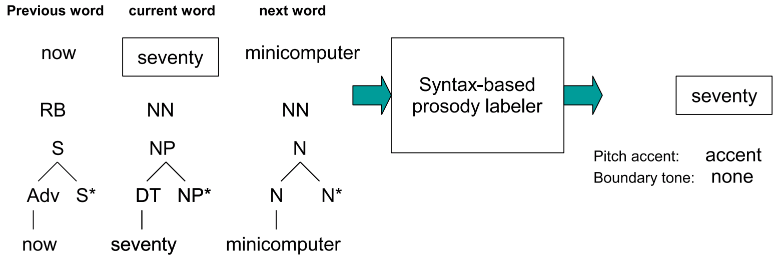


Fig. 2. Illustration of syntax-based prosody predicted by the prosody labeler. The prosody labeler uses lexical and syntactic context from surrounding words to predict the prosody labels for the current word.

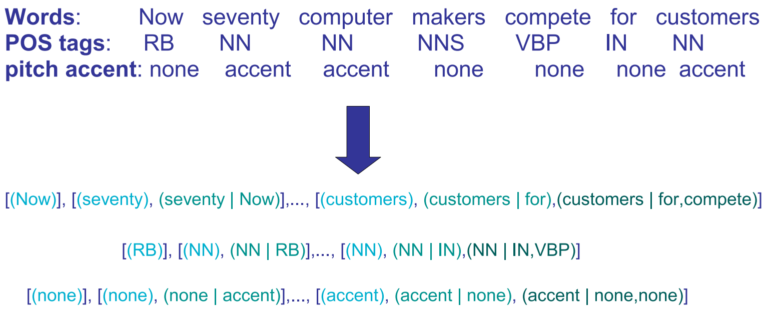

Words: Now seventy computer makers compete for customers
POS tags: RB NN NN NNS VBP IN NN
pitch accent: none accent accent none none none accent
[(Now)], [(seventy), (seventy | Now)],..., [(customers), (customers | for),(customers | for,compete)]
[(RB)], [(NN), (NN | RB)],..., [(NN), (NN | IN),(NN | IN,VBP)]
[(none)], [(none), (none | accent)],..., [(accent), (accent | none), (accent | none,none)]

Fig. 3. Illustration of *n*-gram feature encoding of lexical, syntactic and syntax-based prosody cues. The *n*-gram features represent the feature input space of the maximum entropy classifier. "|" denotes feature input conditioned on the history.

Boston University Radio News corpus. Even though the domain is not the same as that of our test corpora, we expect that the syntactic information in the form of POS tags and supertags can offer good generalization to circumvent the disparity in the domains. Moreover, we expect the syntax-based prosody labeler to offer additional discriminatory evidence beyond the lexical and syntactic features, as the mapping between prosody and syntax is non-linear.

4.2. Acoustic–prosodic features

Exploiting utterance level intonation characteristics in DA tagging presumes the capability for automatic segmentation of the input dialog into discourse segments. Many studies have addressed the problem of automatically detecting the utterance boundaries in a dialog using lexical and prosodic cues (Shriberg et al., 2000; Liu et al., 2006; Ang et al., 2005; Mrozinski et al., 2006). However, we do not attempt to address the problem of utterance segmentation in this paper, and assume that we have access to utterance segmentation marked either automatically or by human labelers. We compute the pitch (f0), RMS energy (e) of the utterance over 10 msec frame intervals. The pitch values in the unvoiced segments were smoothed using linear interpolation. Both the energy and the pitch were normalized with speaker specific mean and variance (*z*-norm).

5. Dialog act classification using true transcripts

We first perform DA tagging experiments on clean transcribed data. While this is typically not available in automated applications, it is a preliminary step and can offer valuable insights into common classification mistakes committed by the classifier when trained on lexical information alone. The lexical cues we use are word trigrams from the current utterance; parts-of-speech, supertagged utterances constitute the syntactic cues, and prosody tagged utterances comprise prosodic cues. In addition, we use the speaker identity information (speaker A or B for the particular dialog since our data were from 2 person interactions). The lexical and syntactic cues are encoded as *n*-gram features and used as input to the maximum entropy classifier. The feature encoding is illustrated in Fig. 3.

Table 2
Dialog act tagging accuracies (in %) for lexical and syntactic cues obtained from true transcripts with the maximum entropy model. Only the current utterance was used to derive the *n*-gram features.

Cues used (current utterance)	Maptask	SWBD-DAMSL	
	12 moves	42 tags	7 tags
Chance (majority tag)	15.6	39.9	54.4
Lexical	65.7	69.7	81.9
Lexical + syntactic	66.1	70.0	82.4
Lexical + syntactic + Syntax-based prosody	66.6	70.4	82.5

Table 3
Examples of misclassifications due to lexical ambiguity from the Switchboard-DAMSL corpus.

Utterance	Reference tag	Hypothesized tag
Yeah	Agreement	Acknowledgement
Right	Agreement	Acknowledgement
You just needed a majority	Question	Statement
Someone had to figure out what was going on	Question	Statement

Results of DA tagging using lexical and syntactic features from reference transcripts are presented in Table 2. Analysis of the confusion matrix obtained from the 7 way classification of dialog acts in the SWBD-DAMSL corpus indicates that the most common misclassifications are: agreements as acknowledgments; questions as statements and, abandoned utterances as statements. 58% of the total agreements in the test set are misclassified as acknowledgments and 61% of the questions are wrongly classified as statements. The misclassifications predominantly occur due to the ambiguity in lexical choice for these discourse functions. Table 3 shows an example of misclassification from the Switchboard-DAMSL corpus while using only lexical information. In the next section, we demonstrate how one can model the intonation characteristics associated with DA types for improved classification.

6. Dialog act classification using acoustic prosody

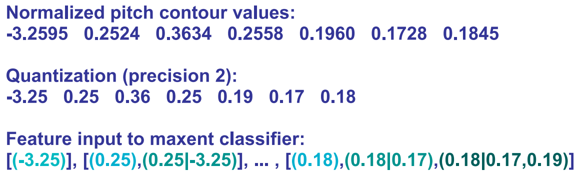
Given that most dialog act classification tasks are typically used as downstream applications that operate on speech input, in this section, we present a maximum entropy framework to model the acoustic–prosodic features in dialog act tagging. As mentioned earlier in Section 4.2, we do not attempt to address the problem of utterance segmentation in this paper. The experiments are performed only on the SWBD-DAMSL corpus since the Maptask corpus is not accompanied by utterance level segmentation. The utterance level segmentation for the SWBD-DAMSL annotations was obtained from the Mississippi State resegmentation of the Switchboard corpus (Hamaker et al., 1998). These segmentations were checked for inconsistencies and cleaned up further. The pitch and energy contour were extracted as explained in Section 4.2.

6.1. Related work

Before describing our proposed prosodic representation for DA tagging, we present a brief overview of previous work that has used prosodic cues for dialog act classification. The use of prosodic cues in DA classification is contingent on two main factors: the type of prosodic representation (categorical or continuous) and the framework used to integrate the prosodic representation with lexical and syntactic cues. Three main representations of the intonation contour have been used in previous work:

(i) Raw/normalized acoustic correlates of intonation such as pitch contour, duration and energy, or transformations thereof (Stolcke et al., 2000; Shriberg et al., 1998).
(ii) Discrete categorical representations of prosody through pitch accents and boundary tones (Black and Campbell, 1995; Reithinger et al., 1996).
(iii) Parametric representations of pitch contour (Yoshimura et al., 1996; Taylor et al., 2000).

Stolcke et al. (2000) used prosodic decision trees to model the raw/normalized acoustic correlates of prosody. They used correlates of duration, pause, pitch, energy and speaking rate as features in the classification. A HMM-based generative model was used for classification. The likelihoods due to multiple knowledge sources were decoupled and a prosodic decision tree classifier was used to estimate the likelihood (obtained from the posterior probability through Bayes rule) of the dialog acts during training. On the Switchboard-DAMSL dataset, they reported a dialog act labeling accuracy of 38.9% using prosody alone (chance being 35%). Using the reference word transcripts and preceding discourse context in an *n*-gram modeling framework, they obtained 71% accuracy. The combined use of prosody, discourse context and lexical cues from



Normalized pitch contour values:
-3.2595 0.2524 0.3634 0.2558 0.1960 0.1728 0.1845

Quantization (precision 2):
-3.25 0.25 0.36 0.25 0.19 0.17 0.18

Feature input to maxent classifier:
[(-3.25)], [(0.25),(0.25|-3.25)], ... , [(0.18),(0.18|0.17),(0.18|0.17,0.19)]

Fig. 4. Illustration of the quantized feature input to the maxent classifier. "|" denotes feature input conditioned on the value of the preceding element in the acoustic–prosodic sequence.

erroneous recognition output resulted in an accuracy of 65%. Ries (1999) and Fernandez and Picard (2002) have also used raw acoustic correlates of prosody for DA classification with neural networks and support vector machines, respectively.

Discrete categorical representations can be effective in characterizing pitch excursions associated with sentence types (Pierrehumbert, 1980). Reithinger et al. (1996) and Black and Campbell (1995) have used symbolic representation of prosodic events as additional features in dialog act tagging for S2S translation and text-to-speech synthesis, respectively. However, automatic detection of detailed categorical representations is still a topic of ongoing research.

Parametric approaches that are data-driven provide a configurational description of the macroscopic intonation contour. Yoshimura et al. (1996) proposed clustering of utterances based on vector-quantized f0 patterns and regression fits. Taylor et al. (2000) have demonstrated the use of parametric representations of the pitch contour for dialog act modeling in speech recognition. On a subset of the Maptask corpus (DCIEM Maptask corpus), they achieved an accuracy of 42% using intonation alone. Using both intonation and dialog history their system correctly identified dialog acts 64% of the time. The drawback of such an approach is that it requires segmentation of the prosodic contour into intonational events, which is not easy to obtain automatically. In the next section, we propose an *n*-gram feature representation of the prosodic contour that is subsequently used within the maxent framework for DA tagging. We also compare the proposed maximum entropy intonation model with the summative statistics acoustic correlates representation used in previous work (Shriberg et al., 1998).

6.2. Maximum entropy intonation model

We quantize the continuous acoustic–prosodic values by binning, and extract *n*-gram features from the resulting sequence. Such a representation scheme differs from the approach commonly used in DA tagging, where representative statistics of the prosodic contour are computed (Shriberg et al., 1998). The *n*-gram features derived from the pitch and energy contour are modeled using the maxent framework described in Section 2. For this case, Eq. (3) becomes,

$$D^* \approx \operatorname*{argmax}_D \prod_{i=1}^{n} p(d_i|\mathbf{\Phi}(A,i)) = \operatorname*{argmax}_D \prod_{i=1}^{n} p(d_i|a_i), \tag{8}$$

where $a_i = \{a_i^1, \ldots, a_i^{k_{u_i}}\}$ is the acoustic–prosodic feature sequence for utterance u_i and the variable k_{u_i} is the number of frames used in the analysis, see Fig. 4.

We fixed the analysis window to the last 100 frames (k_{u_i}) of the discourse segment corresponding to 1 second. The length of the window was empirically determined through optimization on a held-out set. The normalized prosodic contour was uniformly quantized into 10 bins and bigram features[2] were extracted from the sequence of frame level acoustic–prosodic values. Even though the quantization is lossy, it reduces the 'vocabulary' of the acoustic–prosodic features, and hence offers better estimates of the conditional probabilities. In order to test the sensitivity of the proposed framework to errors in utterance segmentation, we also varied the end points of the actual boundary by ±20 frames. There was no significant degradation in performance for this window. However, the performance dropped for incorrect segmentation beyond ±20 frames. Thus, the

[2] Higher order *n*-grams did not result in any significant improvement.

proposed model can also offer some robustness to errors in utterance segmentation. The results of the maxent intonation model are presented in Table 5.

6.3. Comparison with acoustic correlates of prosody

Acoustic correlates of prosody refer to simple transformations of pitch, intensity and duration extracted from the fundamental frequency (f0) contour, energy contour and segmental duration derived from automatic alignment, respectively. Such features have been demonstrated to be beneficial in disfluency detection (Liu et al., 2003), topic segmentation (Hirschberg and Nakatani, 1998), sentence boundary detection (Liu et al., 2006) and dialog act detection (Shriberg et al., 1998). The derived features are also normalized through a variety of speaker and utterance specific normalization techniques to account for the variability across speakers. The major drawback of such a representation is that it is lossy and is not consistent with the suprasegmental theory of prosody that advocates a sequential or continuous model of acoustic correlates over longer durations (O'Connor and Arnold, 1973).

The primary motivation for this experiment is to compare the *n*-gram feature representation of the prosodic contour with previous approaches that have used acoustic correlates of prosody (Shriberg et al., 1998). We extracted a set of 28 features from the pitch and energy contour of each utterance. These included duration of utterance, statistics of the pitch contour (e.g., mean and range of f0 over utterance, slope of f0 regression line) and energy contour (e.g., mean and range of rms energy). The features are directly borrowed from (Shriberg et al., 1998) and a decision tree classifier (J48 in WEKA toolkit (Witten and Frank, 2005)) was trained on the prosodic features for DA classification. The features that were used are summarized in Table 4.

In order to compare the *n*-gram feature representation (presented in Section 6.2) with that of using acoustic correlates, we also fit a decision tree to the *n*-gram features. The results are presented in Table 5. Results indicate that the *n*-gram feature representation performs better than using acoustic correlates, and offers an

Table 4
Acoustic correlates used in the experiment, organized by duration, pitch and energy categories.

Features used	Description
ling_dur	Duration of utterance
f0_mean_good_utt	Mean of f0 values above f0_min
f0_mean_n	Difference between mean f0 of utterance and mean f0 of convside for f0 values > f0_min
f0_mean_ratio	Ratio of f0 mean in utterance to f0 mean in convside
f0_mean_zcv	f0 mean in utterance normalized by mean and std dev of f0 values in convside
f0_sd_good_utt	Std dev of f0 values in utterance
f0_sd_n	Log ratio of std dev of f0 values in utterance and in convside
f0_max_n	Log ratio of max f0 values in utterance and in convside
f0_max_utt	Maximum f0 value in utterance (no smoothing)
max_f0_smooth	Maximum f0 value in smoothed f0 contour
f0_min_utt	Minimum value of f0 in utterance (no smoothing)
utt_grad	Linear regression slope over all points over utterance
pen_grad	Linear regression slope over penultimate 200 ms of utterance
end_grad	Linear regression slope over final 200 ms of utterance
end_f0_mean	Mean f0 in final 200 ms region
pen_f0_mean	Mean f0 in penultimate 200 ms region
abs_f0_diff	Difference between mean f0 of end and penultimate regions
rel_f0_diff	Ratio of f0 of final and penultimate regions
norm_end_f0_mean	Mean f0 in final region normalized by mean and std deviation in convside
norm_pen_f0_mean	Mean f0 in penultimate region normalized by mean and std deviation in convside
norm_f0_diff	Difference between mean f0 of final and penultimate regions, normalized by mean and std dev of f0 from convside
utt_nrg_mean	Mean RMS energy in utterance
abs_nrg_diff	Difference between RMS energy of final and penultimate 200 ms regions
end_nrg_mean	Mean RMS energy in the final 200 ms region
norm_nrg_diff	Normalized difference between mean RMS energy of final and penultimate regions
rel_nrg_diff	Ratio of mean RMS energy of final and penultimate regions

Table 5
Accuracies (%) of DA classification experiments on the Switchboard-DAMSL corpus for different prosodic representations.

Prosodic representation	42 tags	7 tags
Chance (majority tag)	39.9	54.4
Acoustic correlates + decision tree	45.7	60.5
n-gram acoustic features + decision tree	52.1	66.3
n-gram acoustic features + maxent	54.4	69.4

absolute improvement of 6.4% in classification accuracy. The maxent model with the *n*-gram features offers further improvement compared to the decision tree classifier. This may be attributed to the integrated feature selection and modeling offered by the maxent framework.

The results also clearly demonstrate the suitability of the proposed *n*-gram representation for exploiting prosody in DA tagging. Closer analysis of the predictions made by the maxent intonational model (for the simplified SWBD-DAMSL tag set) indicate that majority of the correct predictions are for statements and acknowledgements, with a per dialog act accuracy of 76% and 56%, respectively. The precision and recall for the other categories are very low (less than 1%). In other words, even though the maxent intonation model performs much better than chance, the majority of correct predictions are limited to the two most frequent tags in the DA vocabulary. To evaluate the complementarity of our intonation model with respect to lexical information, in the next section, we perform DA tagging on both clean and recognized transcripts in conjunction with the *n*-gram prosodic contour representation.

7. Dialog act tagging using recognized transcripts

In most speech processing applications, dialog act tagging is either performed simultaneously with front-end automatic speech recognition (ASR) or as a post processing step. The lexical information at the output of ASR is typically noisy due to recognition errors. Thus, modeling the intonational characteristics of discourse segments that are independent of the hypothesized words can offer robustness in DA classification. To evaluate our framework on automatic speech recognition (ASR) output, the 29,869 test utterances were decoded with an ASR setup. The acoustic model for first-pass decoding was a speaker independent model trained on 220 hours of telephone speech from the Fisher English corpus. The language model (LM) was interpolated from the SWBD-DAMSL training set (182K words) and Fisher English corpus (1.5M words). The final hypothesis was obtained after speaker adaptive training using constrained maximum likelihood linear regression on the first-pass lattice. The word error rate (WER) for the test utterances was 34.4%.[3] While this is a relatively high WER, the experiment is intended to provide insights into DA tagging on noisy text.

Table 6 presents DA tagging results using lexical information from reference transcripts (true words) and recognition hypotheses. The accuracy using recognized words is 55.1% compared to 69.7% using the true transcript. The use of prosodic information in conjunction with the words obtained from the recognition output provides a relative improvement of 5.35%. The maxent models described so far use cues from the current utterance only. In the next section, we demonstrate how dialog context can be exploited in our framework.

8. Dialog act tagging using history

The dialog act tags that characterize discourse segments in a dialog are typically dependent on preceding context. For e.g., Questions are usually followed by Statements or Acknowledgments, and, Agreements often follow Statement-opinions. This aspect of dialog acts is usually captured by modeling the prior distribution of dialog act tags as a *k*th order Markov process, *k* being the number of preceding dialog act labels. Such an *n*-gram discourse model of DA tags coupled with locally decomposable likelihoods can be viewed as a *k*th order hidden markov model (HMM). An HMM-based representation of DA tagging, with the states

[3] The decoding was performed on all of 29K utterances for comparison across experiments. The standard deviation of WER was 14.0%.

Table 6
Dialog act tagging accuracies (in %) using lexical + syntactic + prosodic cues for true and recognized transcripts with the maximum entropy model. Only the current utterance was used to derive the *n*-gram features.

Cues used (current utt)	Features used	SWBD-DAMSL	
		42 tags	7 tags
True transcripts	Lexical	69.7	81.9
	Lexical + syntactic + Syntax-based prosody	70.4	82.5
	Lexical + syntactic + Syntax-based prosody + acoustics	70.4(3)	82.5(4)
Recognition output	Lexical	52.3	65.7
	Lexical + syntactic + Syntax-based prosody	53.1	66.8
	Lexical + syntactic + Syntax-based prosody + acoustics	55.1	69.9

corresponding to DAs and observations corresponding to utterances, coupled with a discourse LM, facilitates efficient dynamic programming to compute the most probable DA sequence using the Viterbi algorithm (Stolcke et al., 2000; Taylor et al., 2000; Ji and Bilmes, 2005). Mathematically, the HMM-based DA tagging can be expressed as,

$$D^* = \underset{D}{\operatorname{argmax}}\ P(D|U) = \underset{D}{\operatorname{argmax}}\ P(U|D) \times P(D). \tag{9}$$

The main drawback of such an approach is that one has to wait for the completion of entire conversation before decoding. Thus, optimal decoding can be performed only during offline processing. One way to overcome this problem is by using a greedy decoding approach that uses a discourse LM over the predictions of DA tags at each utterance. However, such an approach is clearly suboptimal and can be further exacerbated when applied to noisy ASR output. The results of such a greedy decoding scheme is presented in Tables 7 and 8.

In contrast to the above methods, we argue for a DA tagging model that uses context history in the form of only *n*-gram lexical and prosodic features from the previous utterances. Our objective is to approximate discourse context information indirectly using acoustic and lexical cues. Such a scheme facilitates online DA tagging and consequently, the decoding can be performed incrementally during automatic speech recognition. Even though the proposed scheme may still be suboptimal, it offers robustness in the decoding process, unlike greedy decoding schemes that can potentially propagate errors. We compare the proposed use of “static” contextual features with the scenario where one has accurate knowledge of previous DA tag. Such a comparison illustrates the gap between the best case scenario (optimal decoding with a bigram discourse LM using the Viterbi algorithm, will be less than or equal to this performance; the greedy approach maybe be worse) and the performance that can be achieved by using only the lexical and prosodic cues from previous utterances. The results are presented in Table 7.

The best case scenario, assuming accurate knowledge of words and the previous dialog act tag (bigram discourse context), results in a DA classification accuracy of 74.4% (see Table 7). A greedy decoding approach with the HMM-based framework and bigram discourse language model yields an DA tagging accuracy of 54.4%, which is much lower than the case when oracle information about previous dialog act tag is accurately known. On the other hand, using only the lexical and prosodic information from 1 previous utterance, yields 71.2% accuracy. The use of only static features from previous utterances is computationally inexpensive and the framework is more robust compared to using greedy DA predictions for each utterance. Adding context from 3 previous utterances[4] results in a classification accuracy of 72%. Similar trends can be observed for DA classification using the ASR output. It is interesting to observe that there is an accuracy drop of only 3% to 4% when using context in terms of lexical and prosodic content from previous utterances, compared to accurate (oracle) knowledge of previous DA. Such a scheme is clearly beneficial in speech applications that require online decoding of dialog act tags.

[4] Context beyond 3 previous utterances did not result in any significant improvement.

Table 7
Dialog act tagging accuracies (in %) using preceding context. current utterance refers to lexical + syntactic + prosodic cues of the current transcribed utterance. prev utterance refers to the lexical + syntactic + prosodic cues from the previous utterance.

Model	Cues used	Maptask	SWBD-DAMSL	
		12 moves	42 tags	7 tags
Greedy decoding	Current utterance + bigram discourse LM	60.1	54.4	76.4
	Current utterance + trigram discourse LM	58.2	54.9	76.8
Maxent	Current utterance	66.6	70.4	82.5
	Current utterance + 1 prev DA tag (oracle)	74.3	74.4	82.9
	Current utterance + 2 prev DA tags (oracle)	75.1	75.8	83.0
	Current utterance + 3 prev DA tags (oracle)	75.2	76.0	83.1
	Current utterance + 1 prev utterance	70.1	71.2	82.7
	Current utterance + 2 prev utterances	70.0	71.8	82.6
	Current utterance + 3 prev utterances	69.9	72.0	82.6

Table 8
Dialog act tagging accuracies (in %) using preceding context. Recognized utterance refers to lexical + syntactic + prosodic cues of the current ASR hypothesized utterance. prev utterance refers to the lexical + syntactic + prosodic cues from the preceding ASR hypotheses.

Model	Cues used	SWBD-DAMSL	
		42 tags	7 tags
Greedy Decoding	Recognized utterance + trigram discourse LM	47.63	57.27
Maxent	Current utterance	70.4	82.5
	Recognized utterance	55.1	69.9
	Recognized utterance + 3 prev DA tags (oracle)	59.7	73.9
	Recognized utterance + 3 prev utterances	56.2	70.8

9. Dialog act tagging using right context

Conventional dialog act tagging schemes (Jurafsky et al., 1997; Stolcke et al., 2000; Ji and Bilmes, 2005) typically use a dialog act grammar to predict the most probable next dialog act based on the previous ones. Exploiting discourse context in such a manner offers a convenient way of modeling the prior distribution of dialog acts in a generative model for dialog act tagging. Often, an n-gram model is chosen as a computationally convenient type of discourse grammar, as it allows for efficient decoding in the HMM framework. While the HMM-based approach to DA tagging is certainly intuitive and desirable in many left-to-right decoding systems, in this section, we are interested in evaluating the usefulness of right context in DA tagging. Since, our maximum entropy decomposes the sequence labeling problem into local classification problems, we can exploit right context of a current utterance during the tagging. In this case, Eq. (2) becomes,

$$\begin{aligned} D^* &= \underset{D}{\operatorname{argmax}}\; P(D|U) \approx \underset{D}{\operatorname{argmax}} \prod_{i=1}^{n} P(d_i|\mathbf{\Phi}(u_i,\ldots,u_{i+l})) \\ &= \underset{D}{\operatorname{argmax}} \prod_{i=1}^{n} P(d_i|\mathbf{\Phi}(W_i,\ldots,W_{i+l},S_i,\ldots,S_{i+l},A_i,\ldots,A_{i+l})), \qquad (10) \end{aligned}$$

where W_i is the word sequence, S_i is the syntactic feature sequence and A_i, the acoustic–prosodic observation belonging to utterances u_i. The variable l denotes right context.

Table 9 shows the results of using right context (words, part-of-speech tags, supertags, syntax-based prosody and acoustics-based prosody of future utterances) in the maximum entropy framework. Just as explained in Section 8, we use only the lexical, syntactic and prosodic information instead of using the actual dialog act tags. The results indicate that trends in improvement when right context is added to the current utterance is similar to that of adding left context for the Switchboard-DAMSL corpus. However, the addition of right

Table 9
Dialog act tagging accuracies (in %) using preceding context. current utterance refers to lexical + syntactic + prosodic cues of the current transcribed utterance. Next utterance refers to the lexical + syntactic + prosodic cues from the succeeding utterance and recognized utterance refers to utterance hypothesized by ASR.

Cues used	Maptask	SWBD-DAMSL	
	12 moves	42 tags	7 tags
Current utterance	66.6	70.4	82.5
Current utterance + 1 next utterance	67.4	71.4	82.8
Current utterance + 2 next utterances	67.3	71.4	82.7
Current utterance + 3 next utterances	67.0	71.3	82.6
Recognized utterance + 3 next utterances	–	56.1	70.7

context (1 next utterance) results in a degradation of about 2.6% in DA tagging accuracy in comparison with the use of left context (1 previous utterance) for the Maptask corpus. Hence, the experimental results indicate that right context is not as beneficial in DA tagging in comparison with left context.

10. Discussion

The maximum entropy framework for DA tagging presented in this work is not restricted to the data sets used in this paper. The framework is generalizable and can be used for multiple tasks that may require the joint use of lexical, syntactic, prosodic and additional cues for identifying dialog acts. Previous work on automatic DA tagging has mainly used lexico-syntactic information in the form of orthographic words and parts-of-speech. In this work, we exploited richer syntactic information such as supertags and prosody predicted from lexical and syntactic cues. These features offer a relative improvement of about 1.0% to 3.0% over using lexical information alone.

The proposed *n*-gram representation of the prosodic contour is trained using a regularized maximum entropy classifier. Thus, the proposed scheme avoids overfitting. In previous work, we have also demonstrated the suitability of such a representation for categorical prosody detection (Rangarajan Sridhar et al., 2007b) and achieved state-of-the-art results. The prosodic representation coupled with the maxent model achieves an accuracy of 54.4% on the SWBD-DAMSL corpus. Previous work on SWBD-DAMSL corpus with intonational cues (Stolcke et al., 2000) achieved an accuracy of 38.9% (chance being 35%). While a direct comparison with our work is not possible due to different training and test splits of the data, our test set consists of about 29K utterances, much larger than the 4K test set used in Stolcke et al. (2000).

To evaluate the complementarity of the lexico-syntactic and prosodic evidence, we performed a correlation analysis on the DA predictions using the two streams of information. We computed Yule's Q statistic (Kuncheva and Whitaker, 2003) for the two classifiers with different features. The value of Q can vary between -1 and 1, with Q taking a value of 0 for statistically independent classifiers. Classifiers that tend to recognize the same samples correctly will thus have positive values of Q. The value of Q for classifiers using lexico-syntactic (true transcripts) and prosodic evidence is 0.85, indicating that the outputs of the two classes are highly correlated. This also explains the relatively small improvement (0.7%) when the prosodic features are added to the classifier using only lexical and syntactic cues. On the other hand, the Q value between recognized transcripts and prosodic cues is 0.64, which in turn can be attributed to the higher improvement (2.8%) when prosodic features are added to the recognition output.

The DA tagging experiments reported on ASR output were performed on the entire test set for consistency across experiments. Our primary motivation was to evaluate the contribution of our intonation model when used with noisy text. We were not concerned with tuning the recognizer to obtain the best performance. However, it is easy to see that the DA tagging accuracy is directly related to the WER of the recognition system. For example, the DA tagging accuracy on a subset of SWBD-DAMSL utterances with 22.0% WER was 64.6%, in comparison with 52.3% accuracy on the entire test set with 34.4% WER.

The proposed use of dialog context from lexical, syntactic and prosodic cues of previous utterances performs well in comparison with previous work (Stolcke et al., 2000) that used the entire conversation for offline optimal decoding. On the SWBD-DAMSL corpus, Stolcke et al. (2000) achieved DA tagging accuracy of

71.0% with a bigram discourse model on true transcripts, while our framework achieves 72.0% accuracy. The best accuracy of 70.1% reported on the Maptask corpus also compares favorably to previous work (Carletta et al., 1997) reported on this corpus. The results indicate that exploiting discourse history information through actual lexical, syntactic and prosodic evidence is as good as representing them through a dialog act discourse model. Further, such a discourse context is limited to about 3 previous utterances. Adding further context does not offer additional knowledge in predicting the dialog act tag of the current utterance.

11. Conclusion and future work

We presented a maximum entropy discriminative model that jointly exploits lexical, syntactic and prosodic cues for automatic dialog act tagging. First, we presented a novel representation scheme for exploiting the intonational properties associated with certain dialog act categories. The *n*-gram feature representation of the prosodic contour, coupled with the maximum entropy learning scheme is effective for the task of distinguishing dialog acts based on intonation alone. The proposed feature representation outperforms conventional techniques such as extracting representative statistics such as mean, slope, variance, etc., from the acoustic correlates of prosody. It also supports the suprasegmental theory of prosody that advocates a sequential or continuous model of acoustic correlates over longer durations. Specifically, the *n*-gram feature representation resulted in an absolute improvement of 6.4% over using the acoustic correlates used in most previous work (Shriberg et al., 1998; Stolcke et al., 2000).

We also demonstrated the use of preceding context in terms of lexical, syntactic and prosodic cues from previous utterances for facilitating online DA tagging. Our maximum entropy framework approximates the previous dialog act state in terms of observed evidence and hence is not limited to offline DA classification that uses the entire conversation during the decoding process. Such a scheme also offers more robustness compared to greedy decoding procedures, which use a discourse model over DA tag predictions at each state. The proposed maxent model achieves DA tagging accuracy of 72% on the SWBD-DAMSL corpus, comparable to the 71% accuracy reported in Stolcke et al. (2000) using offline optimal decoding with a discourse model. Thus, the proposed framework can be used in a variety of speech applications that require online decoding of DA tags.

The methods and algorithms presented in this work were supervised. We plan to investigate unsupervised classification of dialog acts with the help of intonation as part of our future work. Another limitation of the current work is that we assume the knowledge of utterance boundaries for DA tagging. The problem of automatic sentence boundary detection has been well addressed in the literature and we intend to evaluate our framework on boundaries hypothesized by such a detector. Finally, the HMM-based framework and maximum entropy model (with left context) for DA tagging can be applied directly to ASR lattices and thus can enrich the lattices. Such an enriched lattice could be potentially used in applications such as speech-to-speech translation (Rangarajan Sridhar et al., 2008a; Rangarajan Sridhar et al., 2008b). We plan to perform DA tagging on ASR lattices as part of future work.

Acknowledgements

The authors would like to thank Giuseppe Di Fabbrizio of AT&T Labs Research. We would also like to thank Mississippi State University team for providing the resegmented Switchboard corpus. The work was supported in part by DARPA, ONR, NIH and Army.

References

Ang, J., Liu, Y., Shriberg, E., 2005. Automatic dialog act segmentation and classification in multiparty meetings. In: Proceedings of ICASSP.

Austin, J.L., 1962. How to do Things with Words. Clarendon Press, Oxford.

Bangalore, S., Joshi, A.K., 1999. Supertagging: an approach to almost parsing. Computational Linguistics 25 (2).

Bangalore, S., Di Fabbrizio, G., Stent, A., 2006. Learning the structure of task-driven human–human dialogs. In: Proceedings of ACL, Sydney, Australia, pp. 201–208.

Berger, A., Pietra, S.D., Pietra, V.D., 1996. A maximum entropy approach to natural language processing. Computational Linguistics 22 (1), 39–71.

Black, A.W., Campbell, N., 1995. Predicting the intonation of discourse segments from examples in dialogue speech. In: Proceedings of the ESCA Workshop on Spoken Dialogue Systems.

Bolinger, D.L., 1978. Intonation across languages. In: Greenberg, J.P., Ferguson, C.A., Moravcsik, E.A. (Eds.), Universals of Human Language, Phonology, vol. 2. Stanford University Press, Stanford.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A., 1997. The reliability of a dialogue structure coding scheme. Computational Linguistics 23, 13–31.

Chen, J., Vijay-Shanker, K. , 2000. Automated extraction of tags from the penn treebank. In: Proceedings of the 6th International Workshop on Parsing Technologies, Trento, Italy.

Cruttenden, A., 1989. Intonation. Cambridge University Press.

Dudik, M., Phillips, S., Schapire, R.E., 2004. Performance guarantees for regularized maximum entropy density estimation. In: Proceedings of COLT. Springer Verlag, Banff, Canada.

Fernandez, R., Picard, R.W., 2002. Dialog act classification from prosodic features using support vector machines. In: Proceedings of Speech Prosody, pp. 291–294.

Gravano, A., Benus, B., Chávez, J., Hirschberg, Wilcox, L., 2007. On the role of context and prosody in the interpretation of okay. In: Proceedings of ACL, Prague, Czech Republic.

Haffner, P., 2006. Scaling large margin classifiers for spoken language understanding. Speech Communication 48 (iv), 239–261.

Hamaker, J., Deshmukh, N., Ganapathiraju, A., Picone, J., 1998. Resegmentation and transcription of the SWITCHBOARD corpus. In: Proceedings of Speech Transcription Workshop.

Hirschberg, J., Litman, D., 1993. Empirical studies on the disambiguation of cue phrases. Computational Linguistics 19 (3), 501–530.

Hirschberg, J., Nakatani, C., 1998. Acoustic indicators of topic segmentation. In: Proceedings of the International Conference on Spoken Language Proc., pp. 976–979.

Ji, G., Bilmes, J., 2005. Dialog act tagging using graphical models. In: Proceedings of ICASSP.

Joshi, A., Schabes, Y., 1996. Tree-adjoining grammars. In: Salomaa, A., Rozenberg, G. (Eds.), Handbook of Formal Lanaguages and Automata. Springer-Verlag, Berlin.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C., 1997. Automatic detection of discourse structure for speech recognition and understanding. In: Proceedings of ASRU, Santa Barbara, CA, pp. 88–95.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, S., Taylor, P., Van Ess-Dykema, C., 1998. Switchboard discourse language modeling project report, Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, Technical Report Research Note 30.

Jurafsky, D., Shriberg, E., Fox, B., Curl, T., 1998. Lexical, prosodic, and syntactic cues for dialog acts. In: Proceedings of the ACL/COLING Workshop on Discourse Relations and Discourse Markers, Montreal, Canada, pp. 114–120.

Kuncheva, L.I., Whitaker, C.J., 2003. Measures of diversity in classifier ensembles. Machine Learning 51, 181–207.

Lavie, A., Levin, L., Qu, Y., Waibel, A., Gates, D., Gavalada, M., Mayfield, L., Taboada, M., 1996. Dialogue processing in a conversational speech translation system. In: Proceedings of ICSLP, pp. 554–557.

Liu, Y., Shriberg, E., Stolcke, A., 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. In: Proceedings of the Eurospeech, Geneva, pp. 957–960.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, H., Ostendorf, M., Harper, M., 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. IEEE Transactions on Audio, Speech and Language Processing 14 (5), 1526–1540.

Mrozinski, J., Whittaker, E.W.D., Chatain, P., Furui, S., 2006. Automatic sentence segmentation of speech for automatic summarization. In: Proceedings of ICASSP, vol. 1, pp. 14–19.

Murray, G., Renals, S., Moore, J., Carletta, J., 2006. Incorporating speaker and discourse features into speech summarization. In: Proceedings of HLT-NAACL, New York City, USA.

O'Connor, J.D., Arnold, G.F., 1973. Intonation of Colloquial English, second ed. Longman.

Pierrehumbert, J., 1980. The Phonology and Phonetics of English Intonation. Ph.D. Thesis, MIT.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S. Dec. 2006. Acoustic-syntactic maximum entropy model for automatic prosody labeling. In: Proceedings of IEEE/ACL Spoken Language Technology, Aruba.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2007. Exploiting prosodic features for dialog act tagging in a discriminative modeling framework. In: Proceedings of InterSpeech, Antwerp.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2007. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. In: Proceedings of NAACL-HLT.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2008. Enriching spoken language translation with dialog acts. In: Proceedings of ACL.

Rangarajan Sridhar, V.K., Bangalore, S., Narayanan, S., 2008. Factored translation models for enriching spoken language translation with prosody. In: Proceedings of Interspeech, Brisbane, Australia.

Reithinger, N., Engel, R., Kipp, M., Klesen, M., 1996. Predicting dialogue acts for a speech-to-speech translation system. In: Proceedings of ICSLP, vol. 2, pp. 654–657.

Ries, K., 1999. HMM and neural network based speech act detection. In: Proceedings of ICASSP, vol. 1, pp. 497–500.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., Van Ess-Dykema, C., 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? Language and Speech 41 (3–4), 439–487.

Shriberg, E., Stolcke, A., Hakkani-Tur, D., Tur, G., 2000. Prosody-based automatic segmentation of speech into sentences and topics. In: Speech Communication, No. 32 in Special Issue on Accessing Information in Spoken Audio, pp. 127–154.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., Meteer, M., 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics 26 (3), 339–373.

Taylor, P., King, S., Isard, S., Wright, H., 2000. Intonation and dialogue context as constraints for speech recognition. Language and Speech 41 (34), 493–512.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco.

Wu, C.H., Yan, G.L., Lin, C.L., 2002. Speech act modeling in a spoken dialog system using a fuzzy fragment-class Markov model. Speech Communication 38 (1–2), 183–199.

Xia, F., Palmer, M., Joshi, A., 2000. A uniform method of grammar extraction and its applications. In: Proceedings of Empirical Methods in Natural Language Processing.

XTAG, 2001. A lexicalized tree-adjoining grammar for English, Tech. Rep., University of Pennsylvania. <http://www.cis.upenn.edu/xtag/gramrelease.html>.

Yoshimura, T., Hayamizu, S., Ohmura, H., Tanaka, K., 1996. Pitch pattern clustering of user utterances in human–machine dialogue. In: Proceedings of ICSLP, vol. 2, pp. 837–840.

Zimmermann, M., Liu, Y., Shriberg, E., Stolcke, A., 2005. A* based joint segmentation and classification of dialog acts in multiparty meetings. In: Proceedings of the IEEE Speech Recognition and Understanding Workshop, San Juan, Puerto Rico, pp. 215–219.