Exploiting prosodic features for dialog act tagging in a discriminative modeling framework

Vivek Rangarajan[1], *Srinivas Bangalore*[2], *Shrikanth Narayanan*[1]

[1]Speech Analysis and Interpretation Laboratory
Viterbi School of Electrical Engineering, University of Southern California
[2]AT&T Research Labs
180 Park Avenue, Florham Park, NJ 07932, U.S.A.
vrangara@usc.edu, shri@sipi.usc.edu, srini@research.att.com

Abstract

Cue-based automatic dialog act tagging uses lexical, syntactic and prosodic knowledge in the identification of dialog acts. In this paper, we propose a discriminative framework for automatic dialog act tagging using maximum entropy modeling. We propose two schemes for integrating prosody in our modeling framework: (i) Syntax-based categorical prosody prediction from an automatic prosody labeler, (ii) A novel method to model continuous acoustic-prosodic observation sequence as a discrete sequence through the means of quantization. The proposed prosodic feature integration results in a relative improvement of 11.8% over using lexical and syntactic features alone on the Switchboard-DAMSL corpus. The performance of using the lexical, syntactic and prosodic features results in an dialog act tagging accuracy of 84.1%, close to the human agreement of 84%.

1. Introduction

Speech acts or dialog acts [1] are characterizations of actions performed by a speaker during the course of a conversation or a dialog. This characterization provides a representation of conversational function and is especially useful in systems that require an automatic interpretation of dialog act to facilitate a meaningful response or reaction. With the growing demand for integrated approaches to speech recognition, understanding, translation and synthesis, dialog act modeling has come to provide an important link in facilitating human-computer interactions.

Automatic interpretation of dialog acts has been addressed through two main approaches. The AI-style plan inferential interpretation of dialog acts that is designed through plan-inference heuristics [2] and the cue-based interpretation that uses knowledge sources such as, lexical [3, 4, 5], syntactic [6], prosodic [7, 8, 9] and discourse-structure [10]. Even though the plan-inference method can theoretically account for all variations in discourse, it is time-consuming in terms of manual design and computational overhead. On the contrary, data-driven cue-based approaches are computationally friendly and offer a reasonably robust framework to model and detect dialog acts automatically.

Automatic data-driven dialog act tagging is typically statistical in nature and uses various machine learning algorithms such as n-gram models [3, 4], hidden markov models [11], maximum entropy models [6, 12], neural networks [13], etc. These statistical models either use a flat chunk and label paradigm [9, 11, 12] or a hierarchical grammar-based framework [6] to model the dependencies and relations among dialog turns. They exploit multiple knowledge sources in the form of lexical (word identity, keywords), syntactic (parts-of-speech, syntactic structure), prosodic (pitch contour, pitch accents, boundary tones) or discourse structure (dialog history) cues as features in the identification of dialog acts. Prosody in particular has been a very useful feature (as it is domain-independent) that has received a fair amount of attention in cue-based dialog act tagging [7, 8, 9, 14]. Prosodic features such as parameterizations of the pitch contour, duration of segments, energy, as well as, categorical representation of pitch accents and boundary tones have been successfully used to improve dialog act tagging.

Dialog act tagging has been successfully integrated in speech recognition [4, 10], speech understanding [8], text-to-speech synthesis [9, 14] and speech translation [15] systems. Several corpora with domain-specific annotation schemes have been created to facilitate automatic learning of dialog acts [11, 16]. These corpora are hand-labeled for each utterance with a domain-specific dialog act tag set.

In this paper, we propose a discriminative framework for automatic dialog act tagging using maximum entropy modeling. We demonstrate the robustness of our approach in the use of lexical, syntactic and prosodic cues by testing on Maptask [16] and Switchboard-DAMSL [11] corpus. We propose two schemes for integrating prosody in our modeling framework: (i) Syntax-based categorical prosody prediction from an automatic prosody labeler [17], (ii) A novel method to model continuous acoustic-prosodic observation sequence as a discrete sequence through the means of quantization.

2. Related Work

Prosodic features have been used in dialog act tagging in three major ways: (i) Raw/normalized pitch contour, duration and energy, or transformations thereof [4, 5, 8], (ii) Discrete categorical representations of prosody through pitch accents and boundary tones [7, 14, 15] and, (iii) Parametric representation of pitch contour [9, 14]. Stol-

cke et al. [4] used prosodic decision trees to model the raw/normalized prosodic features. They used duration, pauses, pitch and speaking rate features as their prosodic feature vector. On the Switchboard-DAMSL dataset [11], they report a dialog act detection accuracy of 50.2% based on only prosody for a subset of five dialog acts (chance was 35%). Using the original word transcripts in an n-gram modeling framework they obtain 72% accuracy in dialog act detection. The Verbmobil project [15] used symbolic representation of prosodic events as additional features in dialog act tagging within a speech-to-speech translation system. Taylor et al. [9] have demonstrated the use of parametric representations of the pitch contour in dialog act classification. On a subset of the Maptask corpus (DCIEM Maptask corpus), they achieve an accuracy of 69% using the parametric representation of intonation. Prosodic features have been shown to improve dialog act tagging accuracy marginally for automatically recognized transcripts [4], as they offer more discrimination compared to possibly incorrect lexical information from the ASR. However, the incorporation of prosodic features in dialog act tagging has not resulted in significant improvements over lexical and syntactic features.

3. Data

We work with two corpora that have been extensively used for dialog act tagging, Maptask [16] and Switchboard-DAMSL [11]. Maptask [16] is a cooperative task involving two participants. The two speakers, *instruction giver* and *instruction follower* engage in a dialogue 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. The raw move information was augmented with the speaker information and non-verbal content (e.g., laughs, background noise) was removed. The Maptask tagging scheme has 12 unique dialog acts; augmented with speaker information this results in 24 tags. The corpus consists of 128 dialogs and 26181 utterances. We used a ten-fold cross validation for testing.

The Switchboard-DAMSL corpus [11] consists of 1155 dialogs and 218,898 utterances from the Switchboard corpus of telephone conversations, tagged with discourse labels from a shallow discourse tagset. The original tagset of 375 unique tags was clustered to obtain 42 dialog tags. A set of 173 dialogs, selected at random was used for testing. The Switchboard-DAMSL scheme splits long utterances into *slash units*. Thus, a speaker's turn can be divided in one or more slash units and a slash unit can extend over multiple turns. This is illustrated below:

sv B.64 *utt3: C but, F uh -*
b A.65 *utt1: Uh-huh. /*
+ B.66 *utt1: - people want all of that /*
sv B.66 *utt2: C and not all of those are necessities. /*
b A.67 *utt1: Right . /*

The labeling in Switchboard-DAMSL was performed on the basis of the whole slash unit. For e.g., this makes the disfluency turn in B.64 a Statement opinion (sv) rather a non-verbal token. Considering that we use a discriminative classifier, this could introduce noisy data since the context associated with the current labeling decision can appear later in the dialog. Hence, we used two classifiers: (i) non-merged - simply propagates the same label to each continuation, across slash units, (ii) merged - combines the units in one single utterance.

4. Dialog act tagging

We use a chunk based model for dialog act tagging as described in [6]. We model the prediction problem as a classification task in the following manner: given a sequence of utterances u_i in a dialog $U = u_1, u_2, \cdots, u_n$ and a dialog act vocabulary ($d_i \ \epsilon \ \mathcal{D}$), we need to predict the best dialog act sequence $D^* = d_1, d_2, \cdots, d_n$. In our experiments, we use a classifier to assign to each utterance a dialog act conditioned on vector of local contextual (lexical, syntactic, prosodic) features ($\mathbf{\Phi}$). We approximate the conditional probability to be within a bounded n-gram context. Thus,

$$D^* = \arg\max_D P(D|U) \approx \arg\max_D \prod_i^n p(d_i|\mathbf{\Phi}) \quad (1)$$

To estimate the conditional distribution $P(d_i|\mathbf{\Phi})$ we use the general technique of choosing the maximum entropy (maxent) distribution that estimates the average of each feature over the training data [18]. This can be written in terms of Gibbs distribution parameterized with weights λ, where V is the size of the dialog act tag set. Hence,

$$P(d_i|\mathbf{\Phi}) = \frac{e^{\lambda_{d_i}.\mathbf{\Phi}}}{\sum_{l=1}^{V} e^{\lambda_{d_i}.\mathbf{\Phi}}} \quad (2)$$

We use the machine learning toolkit LLAMA [19] 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. Each of the V classes in the tag set $\mathcal{D}$ is encoded as a bit vector such that, in the vector for class i, the i^{th} bit is one and all other bits are zero. Finally, V one-versus-other binary classifiers are used as follows.

$$P(y|\mathbf{\Phi}) = 1 - P(\bar{y}|\mathbf{\Phi}) = \frac{e^{\lambda_y.\mathbf{\Phi}}}{e^{\lambda_y.\mathbf{\Phi}} + e^{\lambda_{\bar{y}}.\mathbf{\Phi}}} \quad (3)$$

where $\lambda_{\bar{y}}$ is the parameter vector for the anti-label $\bar{y}$. To compute $P(l_i|\mathbf{\Phi})$, we use the class independence assumption and require that $y_i = 1$ and for all $j \neq i, y_j = 0$.

$$P(d_i|\mathbf{\Phi}) = P(y_i|\mathbf{\Phi}) \prod_{j \neq i}^{V} P(y_j|\mathbf{\Phi}) \quad (4)$$

4.1. Lexical and Syntactic features

The lexical cues we use are the speaker information, word trigrams from current and previous utterances; parts-of-speech and supertagged utterances constitute the syntactic cues. Supertags [20] encapsulate predicate-argument information in a local structure. They are composed with each other using substitution and adjunction operations of Tree-Adjoining Grammars (TAGs) to derive a dependency analysis of an utterance and its predicate-argument structure. We use only static features that are derived

Cues used	Maptask	Switchboard-DAMSL	
	moves	**non-merged**	**merged**
Chance (majority tag)	15.6	40.5	38.6
Lexical (current utt)	65.7	74.4	77.0
Lexical+Syntactic (current utt)	66.1	75.2	78.2
Lexical+Syntactic+Syntax-based prosody (current utt)	66.6	75.9	78.9
Lexical (3 previous utt)	69.6	79.7	81.7
Lexical+Syntactic (3 previous utt)	69.7	80.9	83.5
Lexical+Syntactic+Syntax-based prosody (3 previous utt)	69.9	82.0	83.9

Table 1: Dialog act tagging accuracies (in %) on Maptask and Switchboard-DAMSL corpora for lexical and syntactic cues

from the local context of the text being tagged. This obviates the need to search for the globally optimal sequence as in the case of using dynamic features. This is especially suitable for dialog act tagging during dialog management, as the prediction is done incrementally rather than waiting for the entire dialog before decoding.

In addition to the lexical and syntactic cues, we also use categorical prosody predicted from our previously developed automatic prosody labeler [17] to tag the utterances with prosodic labels. The prosody labeler uses lexical and syntactic information to predict binary pitch accent (**accent, none**) and boundary tone (**btone, none**) labels for each word. The prosody labeler was trained on the entire 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 and Supertags would provide a reasonable feature representation for prosody detection. Moreover, we expect the syntax-based prosody labeler to offer additional discriminatory evidence above the lexical and syntactic features, as the mapping between prosody and syntax is non-linear. The results are presented in Table 1.

4.2. Acoustic-prosodic features

Given that most dialog act classification tasks are typically performed with a front-end speech interface, in this section, we propose a novel scheme to model the acoustic-prosodic features obtained from the acoustic signal in dialog act tagging. We perform these experiments only on the Switchboard-DAMSL dataset as the Maptask corpus is not accompanied by utterance level segmentation. We obtain utterance level segmentation for the Switchboard-DAMSL annotations from the Mississippi State resegmentation of the Switchboard corpus [21]. We compute the pitch (f0), RMS energy (e) of the utterance over 10 msec frame intervals. Both the energy and the pitch were normalized with speaker specific means and variances. The length of the utterance was also used as a feature.

We model the continuous acoustic-prosodic observation sequence as a discretized sequence through the means of quantization (see Figure 1). We perform this on the normalized pitch and energy extracted from the last 100 frames[1] (1s) of each utterance. The quantized acoustic stream is then used as a feature vector and the conditional probabilities are approximated by an n-gram model. For this case, Eq.(1) becomes,

$$D^* \approx \arg\max_D \prod_i^n p(d_i|\mathbf{\Phi}) = \arg\max_D \prod_i^n p(d_i|a_i) \quad (5)$$

where $a_i = \{a_i^1, \cdots, a_i^k\}$ is the acoustic-prosodic feature sequence for utterance u_i and the variable k is the number of frames used in the analysis.

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

Quantization (precision 1):
-3.2 0.2 0.3 0.2 0.1 0.1 0.1

Figure 1: Quantization of the prosodic features (both temporal and feature scales)

The quantization while being lossy, reduces the vocabulary of the acoustic-prosodic features, and hence offers better estimates of the conditional probabilities. The quantized acoustic-prosodic cues are then modeled using the maximum entropy model described in Section 4. Such an n-gram model of quantized continuous features is similar to representing the acoustic-prosodic features with a piecewise linear fit as done in the tilt intonational model [9]. Essentially, we leave the choice of appropriate representations of the pitch and energy features to the maximum entropy discriminative classifier as opposed to extracting features such as f0 mean, range, slope of regression line, etc., used in [4]. The results of using the acoustic-prosodic features is presented in Table 2.

5. Discussion

The experiments reported in this paper have been performed on transcribed speech. However, the acoustic-prosodic feature modeling framework proposed in this work can be used on the acoustic signal corresponding to any single utterance without knowledge of lexical identity. Using the acoustic-prosodic feature by itself results in a an accuracy of 74.7%, still significantly better than chance. It is interesting to note that use of acoustic-prosodic, lexical, syntactic, syntax-based categorical prosodic cues progressively improves the dialog act tagging performance in that order. It is also important to note that all the experiments were performed on the

[1]We empirically found that 100 frames (1s) was sufficient to capture the patterns in f0 and energy. Shriberg et al. [8] have reported that 200ms window in the end and penultimate regions, respectively, captures the patterns reasonably well.

Cues used	Switchboard-DAMSL	
	non-merged	**merged**
Acoustics only	74.7	76.2
Lexical+Acoustics	83.7	85.6
Lexical+Syntactic+Acoustics	83.9	85.9
Lexical+Syntactic+Syntax-based prosody+Acoustics	84.1	86.2

Table 2: Dialog act tagging accuracies (in %) on Switchboard-DAMSL corpora for acoustic-prosodic cues (only current utt was used). All results are for quantization precision of 2.

complete tagset (42 dialog acts for Switchboard-DAMSL and 12 for Maptask).

The syntax-based prosodic cues offer a marginal improvement on the Maptask corpus, slightly better than previously published results [6]. While the syntax-based categorical prosody from our automatic prosody labeler results only in a relative improvement of 0.75-1.0%, the acoustic-prosodic features modeled through our framework offers as much as 11.8% relative improvement over using lexical and syntactic features alone. The proposed acoustic-prosodic maximum entropy model on the quantized feature values is a simple but very effective technique. The quantization precision can be adjusted based on the amount of available training data. Our results also demonstrate that it is better to leave the choice of the most discriminative acoustic-prosodic feature representation to the maximum entropy classifier rather than using representations of the prosodic contour based on heuristics (f0 slope, maximum value, range, etc.).

Even though the *merged* classifier breaks the structure of regular dialog, it performs consistently better than the *non-merged* classifier. While performing dialog act tagging on merged utterances is not plausible in a real conversation, it can be used for off-line tagging of archived meetings or lectures. The 84.1% accuracy reported for the *non-merged* case with our framework is close to the inter-labeler agreement of 84% on the Switchboard-DAMSL dataset.

The use of lexical, syntactic and prosodic features results in an accuracy of 84.1% which is still higher than using the lexical, syntactic and syntax-based prosody cues from 3 previous utterances. This makes our framework ideal for real-time dialog act detection and the discriminative framework further strengthens the proposed scheme as it obviates the need for searching for a globally optimal sequence, thus, avoiding latency.

6. Conclusions

We presented a discriminative framework for dialog act detection using maximum entropy modeling and demonstrated the integration of prosodic cues in addition to lexical and syntactic cues. The proposed prosodic feature integration results in a relative improvement of 11.8% over using lexical and syntactic features alone. The performance of using the lexical, syntactic and prosodic features results in 84.1% accuracy which is higher than using the lexical and syntactic features from 3 previous utterances (83.9%), making the proposed scheme suitable for real-time automatic dialog act tagging in dialog managers. The reported results are some of the best results on prosody integration in dialog act detection. We plan to use our best performing models in dialog management and spoken language understanding as part of our future work.

7. References

[1] H. Bunt, "Context and dialogue control," *Think*, pp. 19–31, 1994.

[2] J. Allen, G. Ferguson, B. Miller, and E. Ringer, "Spoken dialogue and interactive planning," in *Proceedings of ARPA Speech and Natural Language Workshop*, Austin, Texas, 1995, pp. 202–207.

[3] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts," in *Proc. ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada, Aug. 1998, pp. 114–120.

[4] A. Stolcke et al., "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, Sept. 2000.

[5] A. Venkataraman, A. Stolcke, and E. Shriberg, "Automatic dialog act labeling with minimal supervision," in *Proc. 9th Australian International Conference on Speech Science and Technology*, Melbourne, Dec. 2002.

[6] S. Bangalore, G. Di Fabbrizio, and A. Stent, "Learning the structure of task-driven human-human dialogs," in *Proceedings of ACL*, Sydney, Australia, July 2006, pp. 201–208.

[7] M. Mast et al., "Dialog act classification with the help of prosody," in *Proceedings of ICSLP*, 1996, pp. 1732–1735.

[8] E. Shriberg et al., "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 439–487, 1998.

[9] P. Taylor, S. King, S. Isard, and H. Wright, "Intonation and dialogue context as constraints for speech recognition," *Language and Speech*, vol. 41, no. 34, pp. 493–512, 2000.

[10] D. Jurafsky et al., "Automatic detection of discourse structure for speech recognition and understanding," in *Proceedings of ASRU*, Santa Barbara, CA, Dec. 1997, pp. 88–95.

[11] D. Jurafsky et al., "Switchboard discourse language modeling project report," Center for Speech and Language Processing, Johns Hopkins University, Baltimore, MD, Technical Report Research Note 30, 1998.

[12] M. Poesio and A. Mikheev, "The predictive power of game structure in dialogue act recognition: experimental results using maximum entropy estimation," in *Proceedings of ICSLP*, 1998.

[13] K. Ries, "HMM and neural network based speech act detection," in *Proc. of ICASSP*, vol. 1, March 1999, pp. 497–500.

[14] A. W. Black and N. Campbell, "Predicting the intonation of discourse segments from examples in dialogue speech," in *Proceedings of the ESCA Workshop on Spoken Dialogue Systems*, May 1995.

[15] N. Reithinger, R. Engel, M. Kipp, and M. Klesen, "Predicting dialogue acts for a speech-to-speech translation system," in *Proc. of ICSLP*, vol. 2, no. 3-6, Oct 1996, pp. 654–657.

[16] J. Carletta et al., "The reliability of a dialogue structure coding scheme," *Computational Linguistics*, vol. 23, pp. 13–31, 1997.

[17] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Acoustic-syntactic maximum entropy model for automatic prosody labeling," in *Proc. of IEEE/ACL Spoken Language Technology*, Aruba, Dec. 2006.

[18] A. Berger, S. D. Pietra, and V. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[19] P. Haffner, "Scaling large margin classifiers for spoken language understanding," *Speech Communication*, vol. 48, no. iv, pp. 239–261, 2006.

[20] S. Bangalore and A. Joshi, "Supertagging: An approach to almost parsing," *Computational Linguistics*, vol. 25, no. 2, June 1999.

[21] J. Hamaker et al., "Resegmentation and transcription of the SWITCHBOARD corpus," in *Proceedings of Speech Transcription Workshop*, 1998.