MODELING THE INTONATION OF DISCOURSE SEGMENTS FOR IMPROVED ONLINE DIALOG ACT TAGGING

Vivek Kumar Rangarajan Sridhar, Shrikanth Narayanan

Speech Analysis and Interpretation Laboratory
University of Southern California
Viterbi School of Engineering
vrangara@usc.edu, shri@sipi.usc.edu

Srinivas Bangalore

AT&T Labs - Research
180 Park Avenue
Florham Park, NJ 07932, U.S.A.
srini@research.att.com

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 acoustic correlates of prosody. We also propose a discriminative framework that exploits preceding context in the form of lexical and prosodic cues from previous discourse segments. Such a scheme facilitates online DA tagging and offers robustness in the decoding process, unlike greedy decoding schemes that can potentially propagate errors. Using only lexical and prosodic cues from 3 previous utterances, we achieve a DA tagging accuracy of 72% compared to the best case scenario with accurate knowledge of previous DA tag, which results in 74% accuracy.

Index Terms— dialog act tagging, prosody, maximum entropy model, discriminative 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 understanding and producing natural dialog. Dialog acts [1] are labels that represent communicative acts in a conversation or dialog. Such a representation can be useful in systems that require automatic interpretation of discourse to facilitate a meaningful response or reaction.

Automatic cue-based identification of dialog acts exploits multiple knowledge sources in the form of lexical, syntactic, prosodic and discourse structure cues. These cues have been modeled using stochastic models such as n-gram language models, hidden Markov models, neural networks, fuzzy systems and maximum entropy models. Conventional dialog act tagging systems rely on the words and syntax of utterances. However, in most applications that require front-end speech recognition, the lexical information obtained after decoding is typically noisy due to recognition errors. Moreover, some utterances are inherently ambiguous based on lexical information alone. For example, "okay" can be used in the context of a statement, question or acknowledgment [2].

While lexical information is a strong cue to DA identity, the prosodic information contained in the speech signal can provide another rich source of complementary information. In languages such as English and Spanish, discourse functions are characterized by distinct intonation patterns [3]. These intonation patterns can either be final f0 contour movements or characteristic global shapes of the pitch contour. For example, *yes-no* questions in English 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. [4, 5]. However, these acoustic correlates of prosody provide only a coarse summary of the macroscopic prosodic contour and hence may not exploit the prosodic profile completely. In this work, we exploit the prosodic contour by extracting n-gram features from the acoustic-prosodic sequence. The n-gram feature representation is shown to perform better in comparison with the approach using acoustic correlates of prosody.

We also present a discriminatively trained maximum entropy modeling framework that is suitable for online classification of DAs. Traditional DA systems typically combine the lexical and prosodic features in a HMM framework with a Markovian discourse grammar [4, 6]. 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 very sensitive to noisy input and may cause error propagation. In contrast, our approach uses contextual features captured in the form of just lexical and prosodic cues from previous utterances. Such a scheme is computationally inexpensive and facilitates robust online decoding that can be performed alongside with automatic speech recognition. We evaluate the proposed framework in light of the aforementioned objectives, by testing on the Switchboard DAMSL [6] corpus.

2. DATA

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 tagset. The original tagset of 375 unique tags was clustered to obtain 42 dialog tags as in [6]. A set of 173 dialogs, selected at random was used for testing. The test set consisted of 29869 discourse segments. The experiments were performed on the 42 tag vocabulary as well as a simplified tagset 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 [5]. 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. Figure 1 shows the distribution of the simplified tagset.

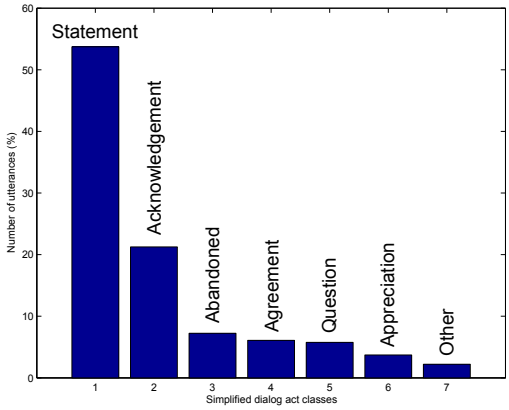

Fig. 1. The distribution of utterances for 7 tags in the Switchboard-DAMSL corpus.

3. 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 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 ($d_i \; \epsilon \; \mathcal{D}, |\mathcal{D}| = K$), we need to predict the best dialog act sequence $D^* = d_1, d_2, \cdots, d_N$. We approximate the string level global classification problem, using conditional independence assumptions to a product of local classification problems as shown in Eq.(1). The classifier is then used to assign to each word a dialog act label conditioned on a vector of local contextual features comprising lexical, syntactic and acoustic information.

$$D^* = \arg\max_D P(D|U) \approx \arg\max_D \prod_{i=1}^{N} p(d_i|\Phi(u_i)) \quad (1)$$

$$D^* \approx \arg\max_D \prod_{i=1}^{N} p(d_i|\Phi(W, S, A, i)) \quad (2)$$

where W is the word sequence, S is the syntactic sequence and A is the acoustic-prosodic observation for utterance u_i.

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

$$p(d|\mathbf{\Phi}) = \frac{e^{\lambda_d . \mathbf{\Phi}}}{\sum_{l=1}^{K} e^{\lambda_l . \mathbf{\Phi}}} \quad (3)$$

We use the machine learning toolkit LLAMA [7] to estimate the conditional distribution using maxent. LLAMA encodes multiclass maxent as binary maxent to increase the training speed and to scale to large data sets. An earlier formulation of this section was presented by the authors in [8], where the framework was tested only on true transcripts.

4. DA CLASSIFICATION USING PROSODY

Exploiting utterance level intonation characteristics in DA tagging presumes the capability to automatically segment the input dialog into discourse segments. However, we do not attempt to address the problem of utterance segmentation in this paper. The utterance level segmentations for the SWBD-DAMSL annotations were obtained from the Mississippi State resegmentation of the Switchboard corpus [9]. The obtained segmentations were checked for inconsistencies and cleaned up further. The pitch (f0) and the RMS energy (e) of the utterance were extracted 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). The length of the utterance was also used as a feature.

In this section, we propose a 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 acoustic correlates representation used in previous work [5]. Our objective is to compare the different prosodic representation schemes and investigate their strengths.

4.1. Sequence model of prosody with maxent framework

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 [5]. The n-gram features derived from the pitch and energy contour are modeled using the maxent framework described in Section 3. For this case, Eq.(2) becomes,

$$D^* \approx \arg\max_D \prod_{i=1}^{N} p(d_i|\mathbf{\Phi}(A,i)) = \arg\max_D \prod_{i=1}^{N} p(d_i|a_i) \quad (4)$$

where $a_i = \{a_i^1, \cdots, a_i^{k_{u_i}}\}$ 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.

We fixed the analysis window to the last 100 frames[1] (k_{u_i}) of the discourse segment corresponding to 1 second. 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 $\pm$20 frames. There was no significant degradation in performance for this window. However, the performance dropped for incorrect segmentation beyond $\pm$20 frames. Thus, the proposed model can offer some robustness to errors in utterance segmentation.

4.2. Acoustic correlates of prosody

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 [5]. Raw or normalized 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. 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). A decision tree classifier (J48 in WEKA toolkit [10]) was trained on the prosodic features for DA classification.

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

Table 1. Accuracies (%) of DA classification experiments for different prosodic representations.

We also fit a decision tree to the n-gram features (presented in Section 4.1) in order to compare the n-gram feature representation with that using acoustic correlates. The results are presented in Table 1. Results indicate that the n-gram feature representation performs better than using acoustic correlates, and offers an 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.

5. DA TAGGING USING RECOGNIZED TRANSCRIPTS

In most speech applications, dialog act tagging is either performed in lockstep 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. To evaluate our framework on automatic speech recognition (ASR) output, the 29869 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. Results in Table 2

Cues used (current utt)	**42 tags**	**7 tags**
True transcripts	69.7	81.9
Recognition output	52.3	65.7
Recognition output+acoustics	55.1	69.9

Table 2. Dialog act tagging accuracies (in %) using true and recognized transcripts with the maximum entropy model.

show the complementarity of the information in the prosodic stream relative to the lexical information. The sequence based acoustic-prosodic representation with the maximum entropy framework offers 2.8% improvement in accuracy over using the recognized transcripts. The performance using the recognition output is a function of the WER of the ASR system. With accurate knowledge of words (true transcripts), the DA classification accuracy is 69.7%.

6. DA TAGGING USING UTTERANCE HISTORY

The dialog act tags that characterize discourse segments in a dialog are typically dependent on preceding context. This aspect of dialog acts is usually captured by modeling the prior distribution of the tags as a k^{th} order Markov process. A HMM based representation of DA tagging, coupled with such a discourse LM, facilitates efficient dynamic programming to compute the most probable DA sequence using the Viterbi algorithm. 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

[1]This was determined empirically by optimization on a held-out set.

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

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

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.

In contrast to the above methods, we argue for a DA tagging model that uses context history in the form of 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 3.

Cues used	**42 tags**	**7 tags**
True transcripts + 1 prev DA tag	74.4	83.1
True transcripts + 3 prev utterances	72.0	82.4
Recognition output + 1 prev DA tag	59.7	73.9
Recognition output + 3 prev utterances	56.2	70.8

Table 3. Dialog act tagging accuracies (in %) using preceding context. Both lexical and prosodic information of utterances were considered.

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 3). On the other hand, using only the lexical and prosodic information from 1 previous utterance, yields 71.2%. 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 3-4% when using context in terms of lexical and prosodic content from previous utterances, compared to accurate (oracle) knowledge of previous DA.

7. CONCLUSION AND FUTURE WORK

We presented a maximum entropy intonation model for DA tagging that uses n-gram features of the normalized and quantized prosodic contour. We showed that the proposed n-gram feature representation is better for exploiting the prosodic characteristics of discourse segments in comparison with acoustic correlates of prosody.

We also showed that our discriminative model can be used for online tagging of DA tags in speech applications. Instead of using predicted DA information, our framework uses context captured in terms of lexical and prosodic cues from preceding utterances. The use of recognition output reduces the DA classification accuracy as expected, due to the relatively high WER for spontaneous speech recognition such as the Switchboard dialogs considered in our experiments. The maximum entropy intonation model still provides an improvement over using the hypothesized word sequence alone. 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. We also plan to use our models in a speech-to-speech translation framework by tagging the source language discourse segments with DA tags for facilitating enriched speech-to-speech translation.

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

8. REFERENCES

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

[2] A. Gravano, B. Benus, J. Chávez, Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of okay," in *Proceedings of ACL*, Prague, Czech Republic, 2007.

[3] D. L. Bolinger, "Intonation across languages," in *Universals of human language*, ser. Phonology, J. P. Greenberg, C. A. Ferguson, and E. A. Moravcsik, Eds. Stanford: Stanford University Press, 1978, vol. 2.

[4] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, Sept. 2000.

[5] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 439–487, 1998.

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

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

[8] V. K. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting prosodic features for dialog act tagging in a discriminative modeling framework," in *Proc. of InterSpeech*, Antwerp, 2007.

[9] J. Hamaker, N. Deshmukh, A. Ganapathiraju, and J. Picone, "Resegmentation and transcription of the SWITCHBOARD corpus," in *Proceedings of Speech Transcription Workshop*, 1998.

[10] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd, Ed. Morgan Kaufmann, San Francisco, 2005.