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 McClintock 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.