Dialogue Act Detection from Human-Human Spoken Conversations

Nithin Ramachandran
Department of Computer Science
and Information Technology,
KMCT College of Engineering,
University of Calicut, Kerala, India

ABSTRACT

Accurate detection of dialogue acts is essential for understanding human conversations and to recognize emotions. This requires 1) the segmentation of human-human dialogs into turns, 2) the intra-turn segmentation into DA boundaries and 3) the classification of each segment according to a DA tag. Most dialogue act classification models approaches the problem of identifying the different DA segments within an utterance in separate fashion: first, DA boundary segmentation within an utterance was addressed with generative or discriminative approaches then, DA labels were assigned to such boundaries based on multi-classification. This paper, presents an effective approach to improve the accuracy of dialogue act recognition from speech signal by combining acoustic and linguistic features. This paper adopts the use of a silence removal algorithm based on Mahalanobis Distance for the segmentation of human-human dialogs into turns and proposes the keyword spotting feature to reduce the ambiguity of opinion vs. non-opinion statements and agreements vs. acknowledgements, occurs while classifying the dialogue acts.

General Terms

Speech and Language Processing, Spoken Language Understanding

Keywords

Dialogue Acts, Silence Removal Algorithms, Conditional Random Fields.

1. INTRODUCTION

Automatic recognition of user emotional states is a very challenging task that has attracted the attention of the research community for several decades. The goal is to design methods to make computers interact more naturally with human beings. This is a very complex task due to a variety of reasons. One is the absence of a generally agreed definition of emotion and of qualitatively different types of emotion. Another is that we still have an incomplete understanding of how humans process emotions, as even people have difficulty in distinguishing between them. Thus, in many cases a given emotion is perceived differently by different people. Studies in emotion recognition made by the research community have been applied to enhance the quality or efficiency of several services provided by computers. The Dialog Act (DA) feature set as a typical pragmatic feature is helpful to classify emotions in user utterances because DAs can represent the current dialog state of a human-human interaction and impose the utterance level context in a dialogue. There are multiple ways to define dialog act set which ranges from generic to specific.

In Spoken Language Understanding, the identification of Dialog Acts (DAs) within an utterance, i.e. its illocutionary acts of communication, is a complementary process to concept extraction. Indeed, as the same concept may occur in a question, an answer or a clarification request, both levels of analysis are necessary for the complete understanding of conversations. Identifying DAs within an utterance is not a trivial task, as utterances may contain more than one DA; hence, prior to DA “classification”, utterances must be segmented according to DA boundaries.

2. RELATED WORKS

Dialogue act classification, which categorizes the intention behind the user’s move (e.g., asking a question, providing declarative information). Because of the importance of dialogue act classification within dialogue systems, it has been an active area of research for some time. Early work on automatic dialogue act classification modeled discourse structure with hidden Markov model [1] or SVMs [2] experimenting with lexical and prosodic features, and applying the dialogue act model as a constraint to aid in automatic speech recognition. A recently proposed alternative approach uses a discriminative approach, namely Conditional Random Fields (CRFs), [3] to simultaneously segment an utterance into its DA boundaries and label such segments according to a DA tag. Of the approaches noted above, this paper uses CRF for dialogue act segmentation. However, it takes a step beyond the previous work by including the use of a silence removal algorithm [4] based on Mahalanobis Distance which uses the statistical properties of the background noise and physiological aspects of speech production the segmentation of human-human dialogs into turns and the keyword spotting features to reduce the ambiguity of opinion vs. non-opinion statements and agreements vs. acknowledgements, occurs when classifying as dialogue acts [3].

3. FEATURE EXTRACTION

The basic algorithm for dialogue act segmentation and classification is very similar to sequence classification problem. Most of the dialogue act detection system uses the feature-based statistical classification with combination of several features using appropriate machine learning methods. In this approach, all of the feature set may not be equally useful and important, therefore creating the need for the reasonable feature extraction methods. The feature sets for textual modality have commonly included the acoustic and linguistic features. In addition to these feature sets, we investigated the use of the keyword spotting features to improve the performance of dialogue act detection in the following sections.

3.1 Acoustic Features

In the proposed approach, firstly two basic acoustic features: pitch, energy are estimated. The previous research showed that dialogue act reaction is strongly related to pitch and energy of the speech. The following 10 features were automatically extracted over the entire user turn using Praat, a program for speech analysis and synthesis [7]: overall energy minimum, maximum, median, and standard deviation, to approximate loudness information; overall fundamental frequency (f0) minimum, maximum, median, standard deviation, and mean absolute slope, to approximate pitch contour; and ratio of voiced frames to total frames, to approximate speaking rate.

3.2 Linguistic Features

As a traditional linguistic feature, n-grams up to length of n = 2 is extracted. In addition, the proposed method uses some lexical and syntactic features from the result of part-of-speech (POS) tagging such as last POS tag and last morpheme .

A primary word types ‘dialogue act keywords’ is manually defined for each dialogue act and these keywords were used to extract confusion pairs of dialogue acts accurately from the input sentence. Analysis of the experiments using CRF Model [3] has shown that out of the top five confusion pairs, 70% involve opinion vs. non-opinion statements, while the remaining ones affect agreements and acknowledgements. Some keywords used for opinion are In my view, To my mind, I feel, If you ask me, To be honest, As far as I’m concerned, I think, It seems, I would argue that, I do not believe, I do not agree etc. Some keywords for agreement are Of course, that’s right, hmmm, etc. The feature set used are summarized in the Table 1.

Table 1: Set of Features used for dialogue act detection.

Feature Set	**Description**
Linguistic Features	Keyword, Pos tag, N-grams
Acoustic Features	Pitch, Energy

4. DIALOGUE ACT DETECTION TASK SETUP

Dialog act detection is the task of identifying the function or goal of a given utterance: thus, it provides complementary information to the identification of domain concepts in the utterance, and a domain independent dialog act scheme can be applied. For modeling a speech corpus, casual English conversations from BBC learning English is used. The proposed method adopts 18 DA tagsets from DAMSL [10] and ADAMACH taxonomy, [5] where utterances have been manually transcribed and annotated according to the DAs. The tag set used are shown in Table 2.

Detection of dialogue acts is done in three steps. 1) The segmentation of human-human dialogs into turns, 2) The intra-turn segmentation into DA boundaries and 3) The classification of each segment according to a DA tag. Acoustic and Linguistic features are used for this purpose. The dialogue act detection process is summarized in Figure 1.

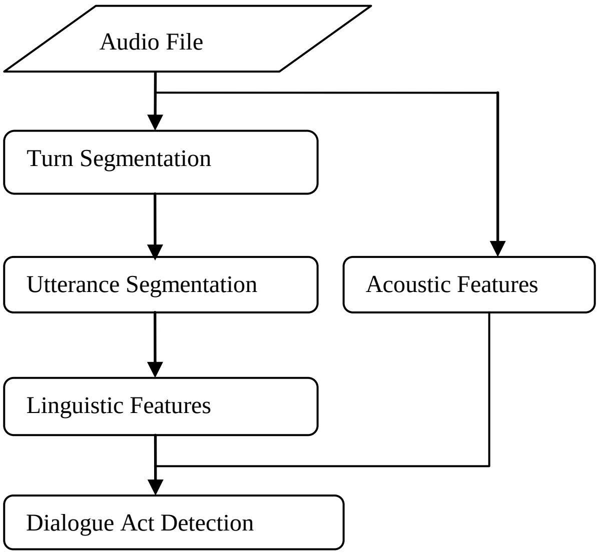


Fig 1: Flow chart of dialogue act detection task

This paper, apply silence removal and endpoint detection algorithms [4] for speech and speaker recognition, and to segment the human-human dialogs into turns. For detecting silence/unvoiced part from the speech sample, the algorithm uses uni-dimensional Mahalanobis Distance [11] function which itself is a Linear Pattern Classifier [11], [12]. The algorithm has two parts. First part assigns label to the samples by using a statistical properties of background noise while the second part smoothens the labeling by the physiological aspects from the speech production process.

For segmentation of DA the proposed method adopt a discriminative approach namely Conditional Random Fields (CRFs), [6] to simultaneously segment an utterance into its DA boundaries and label such segments according to a DA tag. In the learning model, a dialog is represented as an ordered list of turns *ti*, each bearing an utterance *ui*. The latter is annotated with an ordered list of dialog acts (DAs) *dai*0,..,*daiN* , where DA values (*lij*) are taken from a DA taxonomy. Each DA *daij* is transcribed with a word sequence *si*. For instance, turn *t*0, bearing utterance *u*0: “Hi, my phone isn’t working”, may be represented as the sequence of *da*00, with label *l*00 = *GREET* and surface *s*00 = “Hi”, and *da*01, with value *l*01 = *STATEMENT* and word sequence *s*01 =“my phone isn’t working”. Given such a representation, DA segmentation and classification is formalized as a sequence classification problem, i.e. the problem of predicting a single DA label *lij* that applies to a word sequence *sij* in a turn *ti*. To estimate *P*(*lij* |*sij*) given the training dialogs, a combination of features extracted from the utterance is used: these mainly consist of lexical features such as word and Part-of-Speech (POS) n-grams.

Table 2: Dialogue Act Tagset and Labels

Tag Set	**Label**
STATEMENT	[ST]
ACKNOWLEDGEMENT	[ACK]
OPINION	[OP]
GREET	[GR]
AGREEMENT	[AGR]

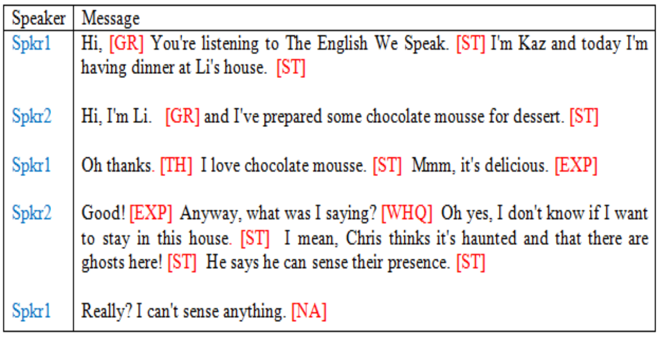
APPRECIATION	[APP]
YES/NOQUESTION	[YN]
YES ANWERS	[YA]
WH-QUESTIONS	[WHQ]
NO ANSWERS	[NA]
APOLOGY	[APO]
THANKING	[TH]
OPENING	[OP]
CLOSING	[CLOSE]
REQUEST	[RQ]
UNINTERRAPTABLE	[UNI]
BACKGROUND	[BCK]
EXPRESSION	[EXP]

As a learning algorithm, the first-order linear-chain CRFs, a category of probabilistic learners frequently used for labeling and segmenting structured data [6] is used. CRFs are undirected graphical models used to specify the conditional probability of assigning output labels given a set of input observations. A conditional probability distribution is defined over label sequences given a particular observation sequence of (e.g. DA surfaces), rather than a joint distribution over both label and observation sequences, CRFs simultaneously segment and assign labels to the tokens of an unsegmented and unlabelled input.

5. EXPERIMENTS AND RESULTS

We experiment with the casual chats. Since there was no existing available data, we developed the data for evaluation ourselves. We collected online casual chats from BBC learning English.Conversations from our corpus annotated with dialog acts are shown in Table 3.

Table 3: An example of a beginning of a dialogue in our corpus showing utterance boundaries and dialogue-act tags in transcript.



Speaker	Message
Spkr1	Hi, [GR] You're listening to The English We Speak. [ST] I'm Kaz and today I'm having dinner at Li's house. [ST]
Spkr2	Hi, I'm Li. [GR] and I've prepared some chocolate mousse for dessert. [ST]
Spkr1	Oh thanks. [TH] I love chocolate mousse. [ST] Mmm, it's delicious. [EXP]
Spkr2	Good! [EXP] Anyway, what was I saying? [WHQ] Oh yes, I don't know if I want to stay in this house. [ST] I mean, Chris thinks it's haunted and that there are ghosts here! [ST] He says he can sense their presence. [ST]
Spkr1	Really? I can't sense anything. [NA]

We applied the Mahalanobis Distance based method to the segmentation of human-human dialogs into turns by detecting silence/unvoiced part from the speech sample. We used CRFs for simultaneous segmentation of an utterance into its DA boundaries and label such segments according to a DA tag. The combination of keyword spotting features, linguistic features and the acoustic features are used to detect dialogue act accurately.

The accuracy is measured with test set using turn level, utterance level and DA accuracy (*AccT, AccU, AccD*); these are respectively defined as the number of correctly segmented and labelled turns out of the total number of turns, the number of correct utterance-level predictions out of the total number of utterance and the percentage of correctly labelled dialogue acts. The experiments were conducted using an audio file in the wave format. The results were analyzed by comparing with the manually detected dialogue acts. It could successfully reduce the ambiguity of the confused dialogue act pairs. The detection accuracy of dialogue acts given different feature sets are shown in Table 4.

Table 4: The detection accuracy of dialogue acts and feature sets used

Feature Sets Used	AccT	AccU	AccD
Acoustic + POS + N grams	71.2%	72.4%	73.4%
Acoustic + POS+N grams + Keywords	76.3%	77.8%	78.2%

The use of keywords made an increase of about five percentages in the accuracy of dialogue act detection. The results are shown graphically in figure 2.

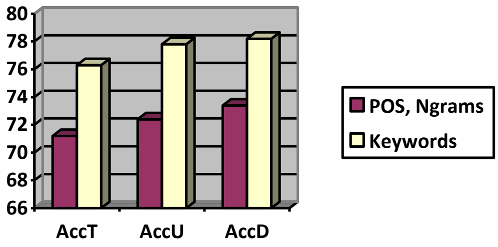


Fig 2: The dialogue act detection accuracy and the feature set used

6. CONCLUSIONS

This paper, presents an approach to detect dialogue acts in utterances. A hybrid method combining the acoustic, linguistic and the keyword spotting features is proposed to the automatic detection of the dialogue acts of user for the affective user interfaces. The work focused on the confusion pairs of dialogue acts like opinion vs. non-opinion statements and agreements vs. acknowledgements, occurs while classifying as dialogue acts. The use of key word spotting feature were proved to reduce this ambiguity. In addition, the method the silence removal algorithm

based on Mahalanobis Distance which improved the accuracy of the segmentation of human-human dialogs into turns. The experiments were conducted using a modeled corpus. In future this work can be extended to repeat the experiments using standard corpus like SWITCHBOARD. It can also be extend to predict the user emotions from speech using dialogue acts with a better accuracy.

7. REFERENCES

[1] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper. 2004. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*

[2] S. Quarteroni and G. Riccardi, “Dialog Act Classification in Human Human and Human Machine Conversations,” in *Proc. INTERSPEECH*, 2010.

[3] Silvia Quarteroni, Alexei V. Ivanov, Giuseppe Riccardi, “Simultaneous Dialog Act Segmentation And Classification from Human-Human Spoken Conversations,” IEEE 2011.

[4] G. Saha, Sandipan Chakroborty, Suman Senapati, “A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications”, Proc. Eleventh Conference on Speech Processing, IIT Khragpur 2005.

[5] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi,“Annotating spoken dialogs: from speech segments to dialog acts and frame semantics,” in *Proc. SRSL*, 2009.

[6] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labelling sequence data,” in *Proc. ICML*, 2001.

[7] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp.341–345, 2001. [Online]. Available: http://www.praat.org

[8] Atal, B.; Rabiner, L., “A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition” Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 24 , Issue: 3 , Jun 1976, Pages: 201 - 212.

[9] D. G. Childers, M. Hand, J. M. Larar, “ Silent and Voiced/Unvoied/Mixed excitation(FourWay),Classification of Speech”, IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.

[10] A. Stolcke, K Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C.Van Ess-dykema, andM.Meteer, “Dialogue Act modeling and automatic tagging And recognition of conversational speech”, Computational Linguistics vol.26, 2000.

[11] Richard. O. Duda, Peter E. Hart, David G. Strok, “Pattern Classification”, A Wiley Inter science publication, John Wiley & Sons, Inc, Second Edition, 2001.

[12] Sarma, V.; Venugopal, D., “Studies on pattern recognition approach to voiced-unvoiced-silence classification”, Acoustics, Speech, and Signal Processing, IEEE International conference on ICASSP '78. ,Volume 3:April 1978, pages 1-4.