# EXPLOITING DIALOGUE ACT TAGGING AND PROSODIC INFORMATION FOR ACTION ITEM IDENTIFICATION

*Fan Yang, Gokhan Tur, Elizabeth Shriberg*

SRI International
Speech Technology and Research Lab
333 Ravenswood Ave, Menlo Park CA 94025
{fly,gokhan,ees}@speech.sri.com

**ABSTRACT**

An important task for multiparty meeting understanding is extracting action items. Action items are a set of tasks that are agreed on by the participants for execution after the meeting, with specific due dates and owners. Dialogue acts, the pragmatic function of an utterance, such as question or backchannel, have been reported to be useful for various dialogue understanding tasks. On the other hand, prosodic information, such as pitch, volume, and speech rate, has been reported to be useful for segmenting a dialogue into utterances or detecting questions. In this paper we investigate the use of dialogue act tagging to improve the identification of action item descriptions and prosodic information to improve action item agreements. Our results indicate that dialogue act tagging improves the identification of action item descriptions by 5% over lexical information, and prosodic information helps discriminating backchannels from agreements with 25% absolute improvement over a baseline.

***Index Terms***— action item, dialogue act, prosody

## 1. INTRODUCTION

There is growing interest in studying multiparty meetings, in which participants share information, discuss issues, and make decisions. One important outcome of a meeting is a set of *action items* (AI), informally defined as a list of tasks that are agreed on by the participants for execution after the meeting, with specific due dates and owners. Figure 1 shows an example where conversants create an action item for C to clean up the web page by mid-July.

Researchers have been exploring approaches to automatically identify action items from recorded meetings. Morgan et al. experimented with automatically determining whether an utterance is related to an action item [1]. They annotated 54 meetings from the ICSI corpus [2] and then trained a maximum entropy model using features from words, context, syntax, timing, prosody, semantics, and dialogue acts.

Purver et al. argued that it is too rough clustering all utterances that are related to action items into a class [3]. An action item usually comprises four components: (1) description: what the task is, (2) owner: who is responsible for the task, (3) time frame: when the task should be finished, and (4) agreement: whether the task proposal is agreed upon by participants. Utterances that are related to different components of an action item have very different features. For example, an utterance related to task description is usually a long statement, such as "you also need to look at your web page" as u1 in Figure 1, while an utterance related to agreement can be very short, such as "okay" as u4. Clustering all these utterances with distinct features into a class for machine learning might make the training of a classifier more difficult, which could in turn lead to low accuracy. Moreover, determining whether an utterance is related to an action item is not sufficient. Richer information is needed to locate the extent of an action item, and to extract its associated properties, such as time frame and owner.

| | | |
|---|---|---|
| u1 | A: | you also need to look at your web page |
| u2 | A: | and clean it up by mid July |
| u3 | B: | p d a free |
| u4 | C: | okay |

**Fig. 1**. An example of an action item

Purver et al. proposed a three-step approach to identify action items [3, 4]. The first step was to determine whether an utterance is related to any components of an action item. For each component, an independent subclassifier was trained to recognize the related utterances. The second step was to locate the extent of action items. A super classifier was trained using the outcome of the first step, hypothesized utterances that are related to each component of action item, over a window of utterances. Finally, the region that contains an action item was parsed to give a summary of the action item.

In this paper we take on Purver et al.'s three-step approach and aim to create more useful features to improve its performance. More specifically, we propose using rich dialogue act tagging and prosodic information for the subclassifiers in the first step to recognize action-item-related utterances. A dialogue act (DA) is an approximate representation of the illocutionary force of an utterance, such as question or backchannel [5]. Dialogue acts are designed to be task independent by definition. The main goal of using dialogue acts is to provide a basis for further discourse analysis and understanding. Dialogue acts have been reported to be useful for plan recognition [6], dialogue management [7], and discourse structure analysis [8]. Prosodic information, such as pitch, volume, and speech rate, has been reported to be useful for segmenting dialogue into utterances [9], annotating dialogue acts [5], identifying discourse markers [10], classifying discourse functions of affirmative words (such as *okay*, *alright*, and *right*) [11], and analyzing discourse structure [12]. Although Purver et al. reported that dialogue acts and prosodic information yielded no improvement for labeling action items [4], this is probably due to the fact that they were using very shallow structures. For example, the dialogue act taxonomy that they used did not even distinguish proposal from acceptance (both were annotated as *statement*). We hypothesize that a richer set of dialogue act tags and prosodic information may help to identify action items: dialogue acts to improve the identification of action item descriptions and prosodic information to improve action item agreements.

## 2. DIALOGUE ACT TAGGING FOR AI DESCRIPTIONS

The first set of features we try for improving the action item extraction is dialogue act tags, more specifically *action motivator* (AM) tags. Below, we first briefly describe the dialogue acts and then explain how we exploited them to get better action item descriptions.

### 2.1. Dialogue Act Tagging

We follow the ICSI Meeting Recorder Dialogue Act (MRDA) tagging standard [13]. It is a hierarchical tagging scheme where an utterance is given a dialogue act label containing three components: a general tag, some specific tags, and a disruption tag. The general tag is a mandatory component of every label, and simply classifies each utterance as a statement, one of different types of question, or one of floor-related units. Only one general tag is present in each dialogue act label. Specific tags provide further description of the utterance, for example, distinguishing whether an utterance gives a proposal or accepts a proposal. Specific tags are appended to the general tag when necessary and are not used alone. The disruption tag indicates when a speaker trails off, is interrupted, or is indecipherable.

In the annotation scheme, the dialogue act tags, including general tags, specific tags, and disruption tags, are categorized into thirteen groups according to syntactic, semantic, pragmatic, and functional similarities of the utterances that they mark. These thirteen groups are statements, questions, floor mechanisms, backchannels and acknowledgments, responses, action motivators, checks, restated information, supportive functions, politeness mechanisms, further descriptions, disruption forms, and nonlabeled.

The group of action motivators is of interest to us in this research. Action motivators are specific dialogue act tags pertaining to immediate or future actions. The group contains three tags: command, suggestion, and commitment. Commands and suggestions are annotated to utterances in which the speaker wants the hearer(s) to perform some actions, and commitments are utterances in which the speaker offers to do something. Figure 2 gives some examples of these tags. Intuitively, because action motivators lead to future actions, they are probably also action item descriptions, such as E1, E3, and E5. Of course, not all action motivators constitute an action item description, such as E2, E4, and E6.

### 2.2. Identification of Action Item Descriptions

In this study we used the tool AdaBoost [14] to take advantage of its strength in text processing. We first examined the use of lexical features, i.e. word unigrams and bigrams, for identifying action item descriptions as a baseline. We then augmented lexical features with the dialogue acts of action motivators. An ideal solution would be to use all specific dialogue acts but there are in total 56 dialogue act tags, and it is rather challenging to develop a model for automatically labeling a corpus with such a detailed tagging system. We trained a binary classifier discriminating whether or not an utterance is an action motivator. This classifier is trained using only lexical features. The output of this AM classifier is used as an additional feature in two ways, either as a binary decision or a continuous value of the confidence estimated by the classifier.

## 3. PROSODIC INFORMATION FOR AI AGREEMENTS

In this study we hypothesize the use of prosodic features for improving action item agreements. We first describe the prosodic features we extracted and then explain how we used them.

**Examples of commands**
*E1*: so maybe just c c hari and say that you've just been asked to handle the large vocabulary part here
*E2*: give me the microphone
**Examples of suggestions**
*E3*: i really would like to suggest looking um a little bit at the kinds of errors
*E4*: should we take turns?
**Examples of commitments**
*E5*: i'll send it out to the list telling people to look at it
*E6*: I'll wait

**Fig. 2**. Examples of action motivators

### 3.1. Prosodic Features

Prosodic features are extracted similar to [9], including pause, energy, pitch, and speech rate. Energy features include the mean, maximum, minimum, and range of intensity in an utterance, after straight-line approximations of energy contours. Pitch features include the mean, maximum, minimum, range, first and last F0 values of an utterance, after stylization to remove halving and doubling errors. Speech-rate features include durations of the word, the longest syllable, the longest vowel, the last syllable, and the last vowel. When appropriate, normalized ratios and Z-scores on the speaker's channel over the whole meeting are also included in the feature set.

### 3.2. Identification of Action Item Agreements

Identifying action item agreements is even challenging, because the information from the words is not sufficient. For example, a "yeah" can be a backchannel that signals the speaker to go on, an acknowledgment indicating understanding, an agreement to a fact statement about the world (e.g., "George W. Bush is the president of U.S."), or an agreement to an action item. We thus propose to use prosodic information to exclude backchannels, and then we can focus on the non-backchannel utterances for action item agreement, probably by using contextual information.

We trained a decision tree classifier using prosodic features to discriminate backchannels. We chose decision tree learning because its output is interpretable and our experiments show that its performance for this task is comparable to other discriminative classifiers.

## 4. EXPERIMENT AND RESULTS

To evaluate these approaches we used the ICSI MRDA Corpus [13]. This corpus is a collection of naturally occurring multiparty conversations. Most of them are regular group meetings at ICSI in which participants discussed research topics. The corpus contains 75 meetings, for a total of about 72 hours.

We used Purver et al.'s action item annotations [4]; 18 meetings are annotated for which of the four components of action item an utterance is related to. Of the total 28,251 utterances, 323 are related to action item description, 226 related to owner, 110 related to time frame, and 351 related to action item agreement. Note that an utterance can be related to multiple components of an action item. For example, the utterance "clean it up by mid July" in Figure 1 is related to both action item description and time frame.

**Table 1**. Top five dialogue acts related to AI descriptions

| Recall | | Precision | | F | |
|---|---|---|---|---|---|
| s | 74.3% | cc | 33.9% | cc | 18.8% |
| cs | 18.6% | co | 12.3% | cs | 13.1% |
| cc | 13.0% | bs | 12.0% | co | 10.4% |
| rt | 11.8% | cs | 10.1% | t | 6.3% |
| co | 9.0% | t | 10.0% | e | 4.2% |

Short-term dialogue act denotations: 's' for statement, 'co' for command, 'cs' for suggestion, 'cc' for commitment, 'bs' for summary, 't' for meeting agenda, 'rt' indicates rising tone at the end of an utterance, 'e' for elaboration.

### 4.1. Action Item Description Experiments

We describe the experiments using dialogue acts to improve action item identification. We first examined the correlation between dialogue acts and action item descriptions. We then ran machine learning experiments to automatically identify action item descriptions.

#### *4.1.1. Correlation between DAs and AI Descriptions*

For each dialogue act tag, we counted the number of utterances annotated as action item description. We then divided this number by the total number of utterances annotated as action item description (323) to calculate *recall*, and divided by the total number of utterances annotated with the dialogue act tag to calculate *precision*. F is the harmonic mean of recall and precision. Table 1 lists the top five dialogue acts that are related to action item description in terms of recall, precision, and F measure.

Not surprisingly, commitment (cc), suggestion (cs), and command (co), which together form the group of action motivators, are the top three dialogue acts in terms of F. All three of them are in the top five lists for both recall and precision. In fact, action motivators as a group correlate with action item description at 40.6% for recall, 13.7% for precision, and 20.5% for F. These results suggest that distinguishing the group of action motivators from other dialogue acts might be useful for identifying action item descriptions.

#### *4.1.2. Results*

We ran some exploratory machine learning experiments to investigate how to make use of dialogue acts to improve the identification of action item descriptions. We divided the data into five subsets and ran five-fold cross validation. In each iteration, three subsets were used as training data, one subset was used as development data to optimize parameters, and one was used as testing data.

Table 2 compares the results of our experiments. For reference, we also show the performance from Purver et al. [4] (line 1). Purver et al. were able to achieve 20% in recall, 12% in precision, and 15% in F by using features including lexical, contextual, syntactic, and timing information. These low numbers suggest that identifying action item descriptions is a very difficult task.

In our experiments, the baseline performance is attained by using only lexical features, i.e. word unigrams and bigrams (line 2). We then experimented using only dialogue acts as the feature set (line 3), and the combination of both words and dialogue acts (line 4). Using words alone has pretty low recall (7.4%), while using dialogue acts is able to identify almost half of the action item descriptions but with lower precision. The combination of both is thus able to improve recall for 5.6% while maintaining the precision. These results suggest that rich dialogue acts, which include general tags, specific tags, and disruption tags, are useful features for the identification of action item descriptions.

**Table 2**. Results on dialogue acts for AI descriptions

| | Recall | Precision | F |
|---|---|---|---|
| Purver et al. | 20% | 12% | 15% |
| Words | 7.4% | 15.7% | 10.1% |
| DA | 42.7% | 13.9% | 21.0% |
| Words+DA | 13.0% | 15.8% | 14.3% |
| Words+AM | 21.4% | 14.4% | 17.2% |
| Words+HypoAM | 10.8% | 9.0% | 9.8% |
| Words+ConfAM | 14.6% | 16.3% | 15.4% |

Inspired by our finding that action motivators are the dialogue acts most correlated to action item description, we experimented with words and action motivators as features for identifying action item descriptions. Line 5 in Table 2 shows the results. The use of action motivators results in a 7.1% improvement in terms of F measure over lexicon only.

We then trained a model to automatically label action motivators. We used the 57 ICSI meetings that were not annotated with action items as training data, and trained with the tool AdaBoost using words (unigram and bigram) as features. We then applied the trained model to the 18 meetings that were annotated with action items. For each utterance, AdaBoost generated two outputs: a binary prediction of whether it is an action motivator (HypoAM), and a confidence value between 0 and 1 predicting how likely it is to be an action motivator (ConfAM). For HypoAM, we got 53.4% in recall, 26.0% in precision, and 35% in F (compared to 6.5% baseline by assuming that all utterances are action motivators).

The HypoAM and ConfAM were then each combined with words as features to identify action item descriptions. Lines 6 and 7 in Table 2 show the results. The combination of HypoAM and words actually decreases performance, probably due to the forced binary decision. However, the combination of ConfAM and words improves both recall and precision over words only, and results in 5% absolute improvement in F. The performance in terms of F is comparable to Purver et al.'s model using features combining words, timing, syntax, semantics, and context.

### 4.2. Action Item Agreement Experiments

As stated in Section 3, the main idea is to discriminate backchannels from other utterances. By doing this, we are not limited to the action item labeled data. Instead, we used the whole corpus, which gives us more data to balance out noise affecting prosody, such as (speaker) individual differences.

Because 96% of backchannels are one-word utterances, we focused on utterances that contain only one word. This removes another noise factor that affects prosody, namely, the length of utterance (in number of words). However, the data somewhat skew: non-backchannels are about 1.6 times more frequent than backchannels. To better understand the prosodic characteristic of backchannels, we down-sampled by randomly selecting from the non-backchannels so that both classes had the same amount of data. We then ran three-fold cross validation to evaluate the performance of decision tree training. This procedure of down-sampling and cross-validation was repeated 10 times. Results are shown in Table 3A. The average accuracy is 73%, with a standard deviation of 0.2%. Recall for backchannels is 79%, precision is 71%, and F measure is 75%. Compared with the

**Table 3**. Results on prosody for backchannels

| | | Accuracy | Recall | Precision | F |
|---|---|---|---|---|---|
| A | baseline | 50% | 50% | 50% | 50% |
| | prosody | 73% | 79% | 71% | 75% |
| B | baseline | 53% | 40% | 34% | 36% |
| | prosody | 73% | 61% | 60% | 61% |

baseline performance of 50% (after down-sampling), these results suggest that most backchannels can be prosodically identified. In the post-analysis of the learned decision trees, we found that backchannels tend to have smaller pauses at the turn transition, a smaller pitch range, and a faster speaking rate.

We then used the 57 meetings without action item annotations as training data, and tested it on the 18 meetings that are annotated with action items. Results are shown in Table 3B. The prosodic features achieve an accuracy of 73%. Recall for backchannels is 61%, precision is 60%, and F measure is 61%. To evaluate performance of the prosodic features, we assumed that we did not have such features available and calculated the Monte Carlo baseline by randomly selecting backchannels according to the prior distribution in the training data, which is 53% in accuracy, 40% in recall, 34% in precision, and 36% in F measure for backchannels. Compared with the baseline, the prosodic features increase accuracy by 20%, and improve F measure for backchannels by 25%. These experiment results suggest that it is possible to use prosody information to exclude backchannels for the identification of action item agreements.

The next step is to identify action item agreements from the non-backchannel utterances, or by using the confidence from the decision tree. We leave this as future work.

## 5. CONCLUSION AND FUTURE WORK

Identifying action items from recorded meetings is a very difficult task. Performance reported in the literature is quite limited. In this paper we have shown that the use of rich dialogue acts and prosodic information can help to improve the identification of action item descriptions and agreements. Moreover, these features, such as confidence score of action motivators and prosody, can be automatically extracted without expensive human labeling.

We are also investigating more features. The fourth dialogue act tag correlated to action item description as shown in Table 1 is meeting agenda (t). Meeting agenda generally constrains the topic of the meeting, and also leads to decision making during the meeting. We speculate that meeting agenda, if available, might also provide useful information for identifying action items. The fifth dialogue act tag correlated to action item description is task elaboration (e). An action item description might cross more than one utterance; there are also utterances that provide more detailed description of the action item. We are trying to identify these utterances by measuring utterance similarity to surrounding action motivators.

We are also experimenting with a more refined approach for identifying action items. Instead of identifying all four components of action items and then using them to determine the extent of an action item, we can first try to identify action item descriptions and time frames. We use this information to determine the extent of a potential action item by applying a relatively large window. We then try to locate the action item owner and agreement inside this window. This constrains the search for action item agreements to be around descriptions, and hopefully will yield higher precision. Finally, we can apply a smaller window to fine-tune the extent of action items, and discard task proposals that are denied.

## 6. ACKNOWLEDGMENTS

We thank Matthew Purver, Patrick Ehlen, and Kristin Precoda for helpful discussions. This work was supported by DARPA CALO (NBCHD-030010) funding at SRI. The opinions and conclusions are those of the authors and not necessarily endorsed by the sponsors.

## 7. REFERENCES

[1] W. Morgan, P. Chang, S. Gupta, and J. Brenier, "Automatically detecting action items in audio meeting recordings," in *Proceedings of 7th SIGDIAL*, 2007.

[2] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proceedings ICASSP*, 2003.

[3] M. Purver, P. Ehlen, and J. Niekrasz, "Detecting action items in multi-party meetings: Annotation and initial experiments," in *Proceedings of MLMI*, Washington, DC, 2006.

[4] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi, "Detecting and summarizing action items in multi-party dialogue," in *Proceedings of the 9th SIGdial*, Sept. 2007.

[5] 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. 339–371, 2000.

[6] D. Litman and J. Allen, "A plan recognition model for subdialogues in conversations," *Cognitive Science*, vol. 11, pp. 163–200, 1987.

[7] C. Sidner, "An artificial discourse language for collaborative negotiation," in *Proceedings of the 12th National Conference on Artificial Intelligence*, 1994.

[8] D. Traum and E. Hinkelman, "Conversation acts in task-oriented spoken dialogue," *Computational Intelligence*, vol. 8, no. 3, pp. 575–599, 1992, Special Issue: Computational Approaches to Non-Literal Language.

[9] E. Shriberg, A. Stolcke, D. Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000, Special Issue on Accessing Information in Spoken Audio.

[10] J. Hirschberg and D. Litman, "Now let's talk about now: identifying cue phrases intonationally," in *Proceedings of the 25th ACL*, 1987, pp. 163–171.

[11] A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue," in *Proceedings of INTERSPEECH*, 2007, pp. 1613–1616.

[12] J. Hirschberg and C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *Proceedings of 34th ACL*, 1996, pp. 286–293.

[13] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act corpus," in *Proceedings of the 5th SIGDIAL*, 2004.

[14] R. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.