Detecting Action Items in Meetings

Gabriel Murray[1] and Steve Renals[2]

[1] University of British Columbia, Vancouver, Canada
gabrielm@cs.ubc.ca
[2] University of Edinburgh, Edinburgh, Scotland
s.renals@ed.ac.uk

Abstract. We present a method for detecting action items in spontaneous meeting speech. Using a supervised approach incorporating prosodic, lexical and structural features, we can classify such items with a high degree of accuracy. We also examine how well various feature subclasses can perform this task on their own.

1 Introduction

Meetings tend to occur in series with regular intervals. While some meetings will be one-off occasions, many others occur weekly or bi-weekly with more or less the same group of participants. As a consequence, the discussion within a given meeting might reference the discussion from a previous meeting, or describe what will happen between the current and upcoming meetings. It is this latter phenomenon of stated *action items* that we are interested in detecting in the current research. Providing a meeting participant with such action items from a previous meeting would be very useful for reminding the individual of what needs to be accomplished before the upcoming meeting.

In this paper we describe a supervised method for detecting these action items, presenting results on a corpus of spontaneous meeting speech. We analyze how well prosodic, lexical, structural and speaker-related features aid this particular task.

2 Experimental Setup

In this section we describe the meeting corpus used, the relevant action item annotations, and the classifier used for these experiments.

2.1 Corpora

For these experiments, we use the AMI meetings corpus [1]. The corpus consists of about 100 hours of recorded and annotated meetings, divided into *scenario* and *non-scenario* meetings. In the scenario portion, groups of four participants role-play in a series of four meetings. Here we use only the scenario meetings from the AMI corpus, numbering 138 in total, with 20 meetings used for our test set. The participants consist of both native and non-native English speakers.

The corpus contains both hand-authored and automatic speech recognition (ASR) transcripts. The ASR system employs the standard framework of context-dependent HMM/GMM acoustic modelling and trigram language models, and features a word error rate (WER) of 38.9%.

2.2 Annotation

For each meeting in the corpus, multiple human annotators are asked to write abstractive summaries of the meeting discussion. The abstract summary consists of a general abstract section in addition to abstract subsections describing *decisions*, *actions* and *problems* from the meeting. The annotators then go through the meeting transcript and link meeting dialogue acts (DAs) to sentences within the abstract, creating a many-to-many mapping of sentences and DAs. We can then determine which DAs represent action items by whether or not they are linked to sentences in the *actions* portion of the transcript. The instruction given to the annotators for writing the *actions* subsection was to "name the next steps that each member of the group will take until the next meeting." There is an average of just under three action item DAs per meeting, but the number depends greatly on which meeting in the series it is – for example, the final meetings in each series contain few action items.

Two examples of action item DAs are given below, taken from meeting IS1003c:

- Speaker A: So you will have Baba and David Jordan you will have to work together on the prototype
- Speaker A: and you will have next time to show us modelling a clay remote control

In these experiments we employed a manual DA segmentation, although automatic approaches are available [3].

2.3 Classifier

The classifier used is the *liblinear* logistic regression classifier[3]. The *liblinear* toolkit incorporates simple feature subset selection based on calculating the f statistic for each feature and performing cross-validation with subsets of various sizes, comparing the resultant balanced accuracy scores. The f statistic for each feature is first calculated [2], and then feature subsets of size n are tried, where n equals 19, 17, 15, 13, 11, 9, 7, 5, and 3, with the n best features included at each step based on the f statistic. The feature subset size with the highest balanced accuracy during cross-validation is selected as the feature set for training. The logistic regression model is then trained on the training data using that subset.

3 Features Description

Table 1 lists and briefly describes the set of the features used. The prosodic features consist of energy , F0, pause, duration and a rate-of-speech measure. We calculate both the duration of the complete DA, as well as of the uninterrupted portion. The structural features include the DA's position in the meeting and position within the speaker's turn (which may contain multiple DAs). There are two measures of speaker dominance: the dominance of the speaker in terms of meeting DAs and in terms of total speaking time. There are two term-weighting metrics, *tf.idf* an d *su.idf*, the former favoring words that

[3] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

are frequent in the given document but rare across all documents, and the latter favoring words that are used with varying frequency by the different speakers [9]. We also include the number of filled pauses in the dialogue act, and the number of abstractive cuewords. These abstractive cuewords are automatically derived from the training data. We examine terms that occur often in the abstracts of meetings but less often in the *extracts* of meetings. We score each word according to the ratio of these two frequencies,

$$TF(t,j)/TF(t,k)$$

where $TF(t,j)$ is the frequency of term t in the set of abstracts j from the training set meetings and $TF(t,k)$ is the frequency of term t in the set of extracts k from the training set meetings. These scores are used to rank the words from most abstractive to least abstractive, and we keep the top 50 words as our list of meta cuewords. The top 5 abstractive cuewords are "team", "group", "specialist", "member", and "manager." For both the manual and ASR feature databases, each DA then has a feature indicating how many of these high-level terms it contains.

Feature ID	Description
Prosodic Features	
ENMN	mean energy
F0MN	mean F0
ENMX	max energy
F0MX	max F0
F0SD	F0 stdev.
PPAU	precedent pause
SPAU	subsequent pause
ROS	rate of speech
Structural Features	
MPOS	meeting position
TPOS	turn position
Speaker Features	
DOMD	speaker dominance (DAs)
DOMT	speaker dominance (seconds)
Length Features	
DDUR	DA duration
UINT	uninterrupted length
WCNT	number of words
Lexical Features	
SUI	su.idf sum
TFI	tf.idf sum
ACUE	abstractive cuewords
FPAU	filled pauses

Table 1. Features Key

4 Results

Figure 1 depicts the f statistics for the features used. The most interesting result is that the abstractive cuewords feature is by far the best single feature according to this measure. The position of the DA in the meeting is also a very useful feature for this task.

Using manual transcripts, the optimal feature set as determined by feature subset selection is comprised of only three features: abstractive cuewords, DA position in meeting, and DA duration. However, with ASR there is a total of nine features selected: abstractive cuewords, DA position in meeting, uninterrupted length, word count, duration, *tf.idf* score, *su.idf* score, and both measures of speaker dominance.

The action item DAs tend to have higher mean and max energy, and higher max F0 and F0 standard deviation than in the negative class. They tend to occur very late in the meeting and also later in a given speaker's turn. They have a much longer duration, higher word count, longer precedent pause, and shorter subsequent pause. They tend to be spoken by the meeting participants who are more dominant in the meeting overall. The rate-of-speech is higher, as are both term-weight scores. The number of abstractive cuewords is dramatically higher, and there tend to be more filled pauses.

Figure 2 shows the ROC curves for both manual and ASR transcripts. The area under the ROC curve (AUROC) is very high in each case: 0.91 for manual transcripts and 0.93 for ASR transcripts, with 0.50 equal to chance performance. This shows that action items can be detected with a high degree of accuracy, and that the classification is robust to ASR errors. This resilience to ASR errors is similar to the finding in automatic speech summarization that summarization results do not greatly deteriorate on speech recognition output [13, 10].

4.1 Feature Subsets

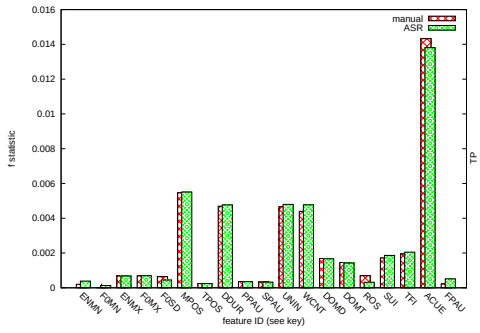


Fig. 1. f statistics for AMI database features

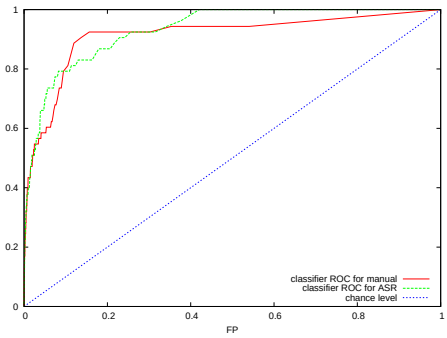
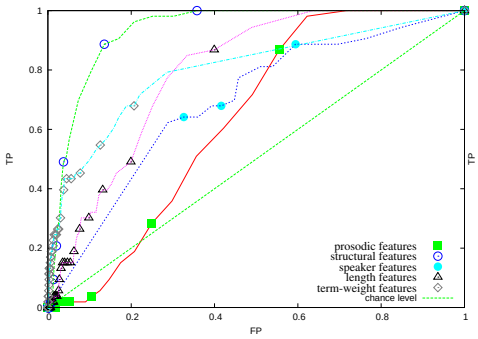


Fig. 2. Classifier ROC Curves, Manual/ASR

Though the f statistics provide us with interesting information about the usefulness of individual features, we would also like to analyze how particular feature *classes* aid the detection of action items. We therefore separate the features into five classes: prosodic, structural, speaker, length and lexical features. Note that we do not consider DA duration and uninterrupted duration to be prosodic features, but rather length features along with DA word count. We then build logistic regression classifiers for each feature class and run the classifiers over the test data. Figure 3 shows the ROC curves and the AUROCs for the feature classes using manual transcripts. The structural class

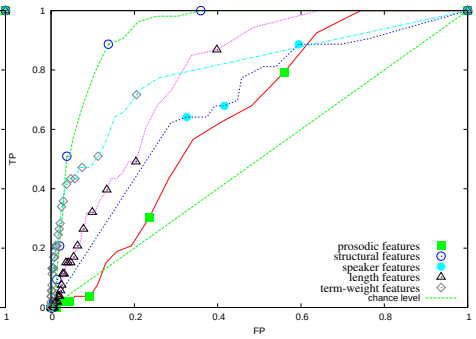
performs the best, with an AUROC of 0.93. This is somewhat surprising, as the structural class contains only two features: DA position in the meeting and DA position in the turn. The length and lexical classes are comparable to each other, with AUROCs of 0.80, while prosodic and speaker features are less useful on their own.

The story is much the same with ASR transcripts. Structural features again are the best performing feature class, and all of the feature classes are robust to ASR errors. Figure 4 shows the ROC curves and AUROCs for ASR transcripts.



Fea. Subset	AUROC
Prosodic	0.64
Structural	0.93
Speaker	0.69
Length	0.80
Lexical	0.80

Fig. 3. AUROC Values, Manual Transcripts



Fea. Subset	AUROC
Prosodic	0.65
Structural	0.93
Speaker	0.69
Length	0.79
Lexical	0.80

Fig. 4. AUROC Values, ASR Transcripts

5 Discussion

It is encouraging to find that action items can be detected with a high degree of accuracy with the given features. Even a small set of lexical and structural features can yield very good performance. It is interesting to note that while abstractive cuewords are the best single feature according to the f statistic, the best feature class is the structural class. Using only information about DA position in the meeting and in the speaker's turn is still enough to detect the action items. Prosodic features are less useful for this task than for speech summarization work [8, 7]. While none of the prosodic features are selected for either manual or ASR transcripts, we do however show that they perform well above chance level when used on their own.

Related work has been carried out by Purver et al. [12, 11] as part of the CALO project, using ICSI meeting data [6]. In that research, the authors used a variety of lexical, structural and prosodic features to detect not just action items in general, but subclasses of action items such as explicit mentions of the action item timeframe and the action item "owner." Like automatic decision detection [5], this work can also be considered a type of focused extractive summarization [10, 4]. By extracting DAs based on

more meaningful criteria than simply informativeness/uninformativeness distinctions, we can create structured or hierarchical summaries.

6 Conclusion

We have shown that action items can be detected with high accuracy using structural and lexical cues. We have also described how these action items are realized in terms of structural, lexical, prosodic, and speaker features. Breaking the features into several classes, we have assessed the performance of each class on its own.

Acknowledgements This work is supported by the European IST Programme Project AMIDA (FP6-0033812). Thanks to the AMI-ASR team for providing the ASR.

References

1. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In *Proc. of MLMI 2005, Edinburgh, UK*, pages 28–39, 2005.
2. Y.-W. Chen and C.-J. Lin. Combining SVMs with various feature selection strategies. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature extraction, foundations and applications*. Springer, 2006.
3. A. Dielmann and S. Renals. DBN based joint dialogue act recognition of multiparty meetings. In *Proc. of ICASSP 2007, Honolulu, USA*, pages 133–136, 2007.
4. M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP 2006, Sydney, Australia*, pages 364–372, 2006.
5. P.-Y. Hsueh, J. Kilgour, J. Carletta, J. Moore, and S. Renals. Automatic decision detection in meeting speech. In *Proc. of MLMI 2007, Brno, Czech Republic*, 2007.
6. 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 *Proc. of IEEE ICASSP 2003, Hong Kong, China*, pages 364–367, 2003.
7. S. Maskey and J. Hirschberg. Comparing lexial, acoustic/prosodic, discourse and structural features for speech summarization. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 621–624, 2005.
8. G. Murray. *Using Speech-Specific Features for Automatic Speech Summarization*. PhD thesis, University of Edinburgh, 2007.
9. G. Murray and S. Renals. Term-weighting for summarization of multi-party spoken dialogues. In *Proc. of MLMI 2007, Brno, Czech Republic*, pages 155–166, 2007.
10. G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. In *Proc. of Interspeech 2005, Lisbon, Portugal*, pages 593–596, 2005.
11. M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. Detecting and summarizing action items in multi-party dialogue. In *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium*, 2007.
12. M. Purver, P. Ehlen, and J. Niekrasz. Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proc. of MLMI 2006, Bethesda, MD, USA*, pages 200–211. 2006.
13. R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarization of spoken audio through information extraction. In *Proc. of the ESCA Workshop on Accessing Information in Spoken Audio, Cambridge UK*, pages 111–116, 1999.