Automatically Detecting Action Items in Audio Meeting Recordings

William Morgan **Pi-Chuan Chang** **Surabhi Gupta** **Jason M. Brenier**

Department of Computer Science
Stanford University
353 Serra Mall
Stanford, CA 94305-9205
ruby@cs.stanford.edu
pcchang@cs.stanford.edu
surabhi@cs.stanford.edu

Department of Linguistics
Center for Spoken Language Research
Institute of Cognitive Science
University of Colorado at Boulder
594 UCB
Boulder, Colorado 80309-0594
jbrenier@colorado.edu

Abstract

Identification of action items in meeting recordings can provide immediate access to salient information in a medium notoriously difficult to search and summarize. To this end, we use a maximum entropy model to automatically detect action item-related utterances from multi-party audio meeting recordings. We compare the effect of lexical, temporal, syntactic, semantic, and prosodic features on system performance. We show that on a corpus of action item annotations on the ICSI meeting recordings, characterized by high imbalance and low inter-annotator agreement, the system performs at an F measure of 31.92%. While this is low compared to better-studied tasks on more mature corpora, the relative usefulness of the features towards this task is indicative of their usefulness on more consistent annotations, as well as to related tasks.

1 Introduction

Meetings are a ubiquitous feature of workplace environments, and recordings of meetings provide obvious benefit in that they can be replayed or searched through at a later date. As recording technology becomes more easily available and storage space becomes less costly, the feasibility of producing and storing these recordings increases. This is particularly true for audio recordings, which are cheaper to produce and store than full audio-video recordings.

However, audio recordings are notoriously difficult to search or to summarize. This is doubly true of multi-party recordings, which, in addition to the difficulties presented by single-party recordings, typically contain backchannels, elaborations, and side topics, all of which further confound search and summarization processes. Making efficient use of large meeting corpora thus requires intelligent summary and review techniques.

One possible user goal given a corpus of meeting recordings is to discover the *action items* decided within the meetings. Action items are decisions made within the meeting that require post-meeting attention or labor. Rapid identification of action items can provide immediate access to salient portions of the meetings. A review of action items can also function as (part of) a summary of the meeting content.

To this end, we explore the task of applying maximum entropy classifiers to the task of automatically detecting action item utterances in audio recordings of multi-party meetings. Although available corpora for action items are not ideal, it is hoped that the feature analysis presented here will be of use to later work on other corpora.

2 Related work

Multi-party meetings have attracted a significant amount of recent research attention. The creation of the ICSI corpus (Janin et al., 2003), comprised of 72 hours of meeting recordings with an average of 6 speakers per meeting, with associated transcripts, has spurred further annotations for various types of information, including dialog acts (Shriberg et al., 2004), topic hierarchies and action items (Gruenstein et al., 2005), and "hot spots" (Wrede and Shriberg, 2003).

The classification of individual utterances based on their role in the dialog, i.e. as opposed to their semantic payload, has a long history, especially in the context of *dialog act* (DA) classification.

Research on DA classification initially focused on two-party conversational speech (Mast et al., 1996; Stolcke et al., 1998; Shriberg et al., 1998) and, more recently, has extended to multi-party audio recordings like the ICSI corpus (Shriberg et al., 2004). Machine learning techniques such as graphical models (Ji and Bilmes, 2005), maximum entropy models (Ang et al., 2005), and hidden Markov models (Zimmermann et al., 2005) have been used to classify utterances from multi-party conversations.

It is only more recently that work focused specifically on action items themselves has been developed. SVMs have been successfully applied to the task of extracting action items from email messages (Bennett and Carbonell, 2005; Corston-Oliver et al., 2004). Bennett and Carbonell, in particular, distinguish the task of action item detection in email from the more well-studied task of text classification, noting the finer granularity of the action item task and the difference of semantics vs. intent. (Although recent work has begun to blur this latter division, e.g. Cohen et al. (2004).)

In the audio domain, annotations for action item utterances on several recorded meeting corpora, including the ICSI corpus, have recently become available (Gruenstein et al., 2005), enabling work on this topic.

3 Data

We use action item annotations produced by Gruenstein et al. (2005). This corpus provides topic hierarchy and action item annotations for the ICSI meeting corpus as well as other corpora of meetings; due to the ready availability of other types of annotations for the ICSI corpus, we focus solely on the annotations for these meetings. Figure 1 gives an example of the annotations.

The corpus covers 54 ICSI meetings annotated by two human annotators, and several other meetings annotated by one annotator. Of the 54 meetings with dual annotations, 6 contain no action items. For this study we consider only those meetings which contain action items and which are annotated by both annotators.

As the annotations were produced by a small number of untrained annotators, an immediate question is the degree of consistency and reliability. Inter-annotator agreement is typically measured by the kappa statistic (Carletta, 1996), defined as:

$$\kappa = \frac{P(O) - P(E)}{1 - P(E)}$$

where $P(O)$ is the probability of the observed agreement, and $P(E)$ the probability of the "expected agreement" (i.e., under the assumption the two sets of annotations are independent). The kappa statistic ranges from -1 to 1, indicating perfect disagreement and perfect agreement, respectively.

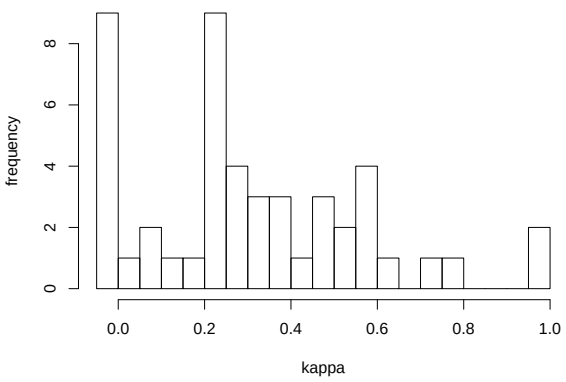

Figure 2: Distribution of κ (inter-annotator agreement) across the 54 ICSI meetings tagged by two annotators. Of the two meetings with $\kappa = 1.0$, one has only two action items and the other only four.

Overall inter-annotator agreement as measured by κ on the action item corpus is poor, as noted in Purver et al. (2006), with an overall κ of 0.364 and values for individual meetings ranging from 1.0 to less than zero. Figure 2 shows the distribution of κ across all 54 annotated ICSI meetings.

To reduce the effect of poor inter-annotator agreement, we focus on the top 15 meetings as ranked by κ; the minimum κ in this set is 0.435. Although this reduces the total amount of data available, our intention is that this subset of the most consistent annotations will form a higher-quality corpus.

While the corpus classifies related action item utterances into action item "groups," in this study we wish to treat the annotations as simply binary attributes. Visual analysis of annotations for several meetings outside the set of chosen 15 suggests that the union of the two sets of annotations yields the most consistent resulting annotation; thus, for this study, we consider an utterance to be an action item if at least one of the annotators marked it as such.

The 15-meeting subset contains 24,250 utter-

A1	A2	
X	X	So that will be sort of the assignment for next week, is to—
X	X	to—for slides and whatever net you picked and what it can do and—and how far you've gotten. Pppt!
X	-	Well, I'd like to also,
X	X	though, uh, ha- have a first cut at what the
X	X	belief-net looks like.
-	X	Even if it's really crude.
-	-	OK? So, you know,
-	-	here a- here are—
-	X	So we're supposed to @@ about features and whatnot, and—

Figure 1: Example transcript and action item annotations (marked "X") from annotators A1 and A2. "@@" signifies an unintelligible word. This transcript is from an ICSI meeting recording and has $\kappa = 0.373$, ranking it 16^{th} out of 54 meetings in annotator agreement.

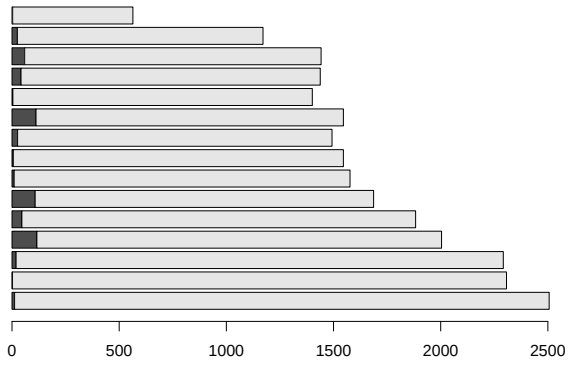

Figure 3: Number of total and action item utterances across the 15 selected meetings. There are 24,250 utterances total, 590 of which (2.4%) are action item utterances.

ances total; under the union strategy above, 590 of these are action item utterances. Figure 3 shows the number of action item utterances and the number of total utterances in the 15 selected meetings.

One noteworthy feature of the ICSI corpus underlying the action item annotations is the "digit reading task," in which the participants of meetings take turns reading aloud strings of digits. This task was designed to provide a constrained-vocabulary training set of speech recognition developers interested in multi-party speech. In this study we did not remove these sections; the net effect is that some portions of the data consist of these fairly atypical utterances.

4 Experimental methodology

We formulate the action item detection task as one of binary classification of utterances. We apply a maximum entropy (maxent) model (Berger et al., 1996) to this task.

Maxent models seek to maximize the conditional probability of a class c given the observations X using the exponential form

$$P(c|X) = \frac{1}{Z(X)} \exp\left[\sum_i \lambda_{i,c}\, f_{i,c}(X)\right]$$

where $f_{i,c}(X)$ is the ith feature of the data X in class c, $\lambda_{i,c}$ is the corresponding weight, and $Z(X)$ is a normalization term. Maxent models choose the weights $\lambda_{i,c}$ so as to maximize the entropy of the induced distribution while remaining consistent with the data and labels; the intuition is that such a distribution makes the fewest assumptions about the underlying data.

Our maxent model is regularized by a quadratic prior and uses quasi-Newton parameter optimization. Due to the limited amount of training data (see Section 3) and to avoid overfitting, we employ 10-fold cross validation in each experiment.

To evaluate system performance, we calculate the F measure (F) of precision (P) and recall (R), defined as:

$$\begin{aligned} P &= \frac{|A \cap C|}{|A|} \\ R &= \frac{|A \cap C|}{|C|} \\ F &= \frac{2PR}{P+R} \end{aligned}$$

where A is the set of utterances marked as action items by the system, and C is the set of (all) correct action item utterances.

The use of precision and recall is motivated by the fact that the large imbalance between positive and negative examples in the corpus (Section 3) means that simpler metrics like accuracy are insufficient—a system that simply classifies every utterance as negative will achieve an accuracy of 97.5%, which clearly is not a good reflection of desired behavior. Recall and F measure for such a system, however, will be zero.

Likewise, a system that flips a coin weighted in proportion to the number of positive examples in the entire corpus will have an accuracy of 95.25%, but will only achieve $P = R = F = 2.4\%$.

5 Features

As noted in Section 3, we treat the task of producing action item annotations as a binary classification task. To this end, we consider the following sets of features. (Note that all real-valued features were range-normalized so as to lie in $[0, 1]$ and that no binning was employed.)

5.1 Immediate lexical features

We extract word unigram and bigram features from the transcript for each utterance. We normalize for case and for certain contractions; for example, "I'll" is transformed into "I will".

Note that these are oracle features, as the transcripts are human-produced and not the product of automatic speech recognizer (ASR) system output.

5.2 Contextual lexical features

We extract word unigram and bigram features from the transcript for the previous and next utterances across all speakers in the meeting.

5.3 Syntactic features

Under the hypothesis that action item utterances will exhibit particular syntactic patterns, we use a conditional Markov model part-of-speech (POS) tagger (Toutanova and Manning, 2000) trained on the Switchboard corpus (Godfrey et al., 1992) to tag utterance words for part of speech. We use the following binary POS features:

- Presence of UH tag, denoting the presence of an "interjection" (including filled pauses, unfilled pauses, and discourse markers).
- Presence of MD tag, denoting presence of a modal verb.
- Number of NN* tags, denoting the number of nouns.
- Number of VB* tags, denoting the number of verbs.
- Presence of VBD tag, denoting the presence of a past-tense verb.

5.4 Prosodic features

Under the hypothesis that action item utterances will exhibit particular prosodic behavior—for example, that they are emphasized, or are pitched a certain way—we performed pitch extraction using an auto-correlation method within the sound analysis package Praat (Boersma and Weenink, 2005). From the meeting audio files we extract the following prosodic features, on a per-utterance basis: (pitch measures are in Hz; intensity in energy; normalization in all cases is z-normalization)

- Pitch and intensity range, minimum, and maximum.
- Pitch and intensity mean.
- Pitch and intensity median (0.5 quantile).
- Pitch and intensity standard deviation.
- Pitch slope, processed to eliminate halving/doubling.
- Number of voiced frames.
- Duration-normalized pitch and intensity ranges and voiced frame count.
- Speaker-normalized pitch and intensity means.

5.5 Temporal features

Under the hypothesis that the length of an utterance or its location within the meeting as a whole will determine its likelihood of being an action item—for example, shorter statements near the end of the meeting might be more likely to be action items—we extract the duration of each utterance and the time from its occurrence until the end of the meeting. (Note that the use of this feature precludes operating in an online setting, where the end of the meeting may not be known in advance.)

5.6 General semantic features

Under the hypothesis that action item utterances will frequently involve temporal expressions—e.g. "Let's have the paper written by *next Tuesday*"—we use Identifinder (Bikel et al., 1997) to mark temporal expressions ("TIMEX" tags) in utterance transcripts, and create a binary feature denoting

the existence of a temporal expression in each utterance.

Note that as Identifinder was trained on broadcast news corpora, applying it to the very different domain of multi-party meeting transcripts may not result in optimal behavior.

5.7 Dialog-specific semantic features

Under the hypothesis that action item utterances may be closely correlated with specific dialog act tags, we use the dialog act annotations from the ICSI Meeting Recorder Dialog Act Corpus. (Shriberg et al., 2004) As these DA annotations do not correspond one-to-one with utterances in the ICSI corpus, we align them in the most liberal way possible, i.e., if at least one word in an utterance is annotated for a particular DA, we mark the entirety of that utterance as exhibiting that DA.

We consider both fine-grained and coarse-grained dialog acts.[1] The former yields 56 features, indicating occurrence of DA tags such as "appreciation," "rhetorical question," and "task management"; the latter consists of only 7 classes—"disruption," "backchannel," "filler," "statement," "question," "unlabeled," and "unknown."

6 Results

The final performance for the maxent model across different feature sets is given in Table 1. F measures scores range from 13.81 to 31.92. Figure 4 shows the interpolated precision-recall curves for several of these feature sets; these graphs display the level of precision that can be achieved if one is willing to sacrifice some recall, and vice versa.

Although ideally, all combinations of features should be evaluated separately, the large number of features in this precludes this strategy. The combination of features explored here was chosen so as to start from simpler features and successively add more complex ones. We start with transcript features that are immediate and context-independent ("unigram", "bigram", "TIMEX"); then add transcript features that require context ("temporal", "context"), then non-transcript (i.e. audio signal) features ("prosodic"), and finally add features that require both the transcript and the audio signal ("DA").

[1] We use the `map_01` grouping defined in the MRDA corpus to collapse the tags.

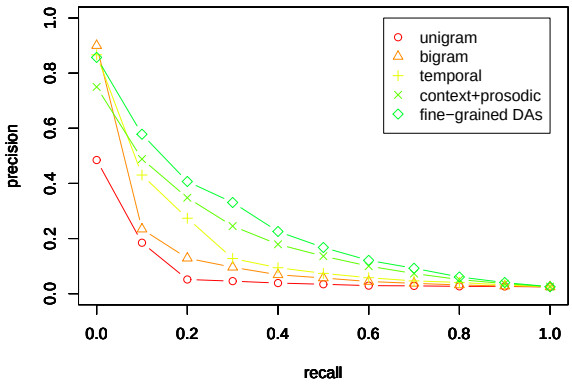

Figure 4: Interpolated precision-recall curve for several (cumulative) feature sets. This graph suggests the level of precision that can be achieved if one is willing to sacrifice some recall, and vice versa.

In total, nine combinations of features were considered. In every case except that of syntactic and coarse-grained dialog act features, the additional features improved system performance and these features were used in succeeding experiments. Syntactic and coarse-grained DA features resulted in a drop in performance and were discarded from succeeding systems.

7 Analysis

The unigram and bigram features provide significant discriminative power. Tables 2 and 3 give the top features, as determined by weight, for the models trained only on these features. It is clear from Table 3 that the detailed end-of-utterance punctuation in the human-generated transcripts provide valuable discriminative power.

The performance gain from adding TIMEX tagging features is small and likely not statistically significant. Post-hoc analysis of the TIMEX tagging (Section 5.6) suggests that Identifinder tagging accuracy is quite plausible in general, but exhibits an unfortunate tendency to mark the digit-reading (see Section 3) portion of the meetings as temporal expressions. It is plausible that removing these utterances from the meetings would allow this feature a higher accuracy.

Based on the low feature weight assigned, utterance length appears to provide no significant value to the model. However, the time until the meeting is over ranks as the highest-weighted feature in the unigram+bigram+TIMEX+temporal feature set. This feature is thus responsible for the 39.25%

features	number	F	% imp.
unigram	6844	13.81	
unigram+bigram	61281	16.72	21.07
unigram+bigram+TIMEX	61284	16.84	0.72
unigram+bigram+TIMEX+temporal	61286	23.45	39.25
unigram+bigram+TIMEX+temporal+syntactic	*61291*	*21.94*	*-6.44*
unigram+bigram+TIMEX+temporal+context	183833	25.62	9.25
unigram+bigram+TIMEX+temporal+context+prosodic	183871	27.44	7.10
unigram+bigram+TIMEX+temporal+context+prosodic+coarse DAs	*183878*	*26.47*	*-3.53*
unigram+bigram+TIMEX+temporal+context+prosodic+fine DAs	183927	31.92	16.33

Table 1: Performance of the maxent classifier as measured by F measure, the relative improvement from the preceding feature set, and the number of features, across all feature sets tried. Italicized lines denote the addition of features which do not improve performance; these are omitted from succeeding systems.

feature	+/-	λ
"pull"	+	2.2100
"email"	+	1.7883
"needs"	+	1.7212
"added"	+	1.6613
"mm-hmm"	-	1.5937
"present"	+	1.5740
"nine"	-	1.5019
"!"	-	1.5001
"five"	-	1.4944
"together"	+	1.4882

Table 2: Features, evidence type (positive denotes action item), and weight for the top ten features in the unigram-only model. "Nine" and "five" are common words in the digit-reading task (see Section 3).

feature	+/-	λ
"- $"	-	1.4308
"i will"	+	1.4128
", $"	-	1.3115
"uh $"	-	1.2752
"w- $"	-	1.2419
". $"	-	1.2247
"email"	+	1.2062
"six $"	-	1.1874
"* in"	-	1.1833
"so $"	-	1.1819

Table 3: Features, evidence type and weight for the top ten features in the unigram+bigram model. The symbol * denotes the beginning of an utterance and $ the end. All of the top ten features are bigrams except for the unigrams "email".

feature	+/-	λ
mean intensity (norm.)	-	1.4288
mean pitch (norm.)	-	1.0661
intensity range	+	1.0510
"i will"	+	0.8657
"email"	+	0.8113
reformulate/summarize (DA)	+	0.7946
"just go" (next)	+	0.7190
"i will" (prev.)	+	0.7074
"the paper"	+	0.6788
understanding check (DA)	+	0.6547

Table 4: Features, evidence type and weight for the top ten features on the best-performing model. Bigrams labeled "prev." and "next" correspond to the lexemes from previous and next utterances, respectively. Prosodic features labeled as "norm." have been normalized on a per-speaker basis.

boost in F measure in row 3 of Table 1.

The addition of part-of-speech tags actually decreases system performance. It is unclear why this is the case. It may be that the unigram and bigram features already adequately capture any distinctions these features make, or simply that these features are generally not useful for distinguishing action items.

Contextual features, on the other hand, improve system performance significantly. A post-hoc analysis of the action item annotations makes clear why: action items are often split across multiple utterances (e.g. as in Figure 1), only a portion of which contain lexical cues sufficient to distinguish them as such. Contextual features thus allow utterances immediately surrounding these "obvious" action items to be tagged as well.

Prosodic features yield a 7.10% increase in F measure, and analysis shows that speaker-normalized intensity and pitch, and the range in intensity of an utterance, are valuable discriminative features. The subsequent addition of coarse-grained dialog act tags does not further improve system performance. It is likely this is due to reasons similar to those for POS tags—either the categories are insufficient to distinguish action item utterances, or whatever usefulness they provide is subsumed by other features.

Table 4 shows the feature weights for the top-ranked features on the best-scoring system. The addition of the fine-grained DA tags results in a significant increase in performance.The F measure of this best feature set is 31.92%.

8 Conclusions

We have shown that several classes of features are useful for the task of action item annotation from multi-party meeting corpora. Simple lexical features, their contextual versions, the time until the end of the meeting, prosodic features, and fine-grained dialog acts each contribute significant increases in system performance.

While the raw system performance numbers of Table 1 are low relative to other, better-studied tasks on other, more mature corpora, we believe the relative usefulness of the features towards this task is indicative of their usefulness on more consistent annotations, as well as to related tasks.

The Gruenstein et al. (2005) corpus provides a valuable and necessary resource for research in this area, but several factors raise the question of annotation quality. The low κ scores in Section 3 are indicative of annotation problems. Post-hoc error analysis yields many examples of utterances which are somewhat difficult to imagine as possible, never mind desirable, to tag. The fact that the extremely useful oracular information present in the fine-grained DA annotation does *not* raise performance to the high levels that one might expect further suggests that the annotations are not ideal—or, at the least, that they are inconsistent with the DA annotations.[2]

This analysis is consistent with the findings of Purver et al. (2006), who achieve an F measure of less than 25% when applying SVMs to the classification task to the same corpus, and motivate the development of a new corpus of action item annotations.

9 Future work

In Section 6 we showed that contextual lexical features are useful for the task of action item detection, at least in the fairly limited manner employed in our implementation, which simply looks at immediate previous and immediate next utterances. It seems likely that applying a sequence model such as an HMM or conditional random field (CRFs) will act as a generalization of this feature and may further improve performance.

Addition of features such as speaker change and "hot spots" (Wrede and Shriberg, 2003) may also aid classification. Conversely, it is possible that feature selection techniques may improve performance by helping to eliminate poor-quality features. In this work we have followed an "everything but the kitchen sink" approach, in part because we were curious about which features would prove useful. The effect of adding POS and coarse-grained DA features illustrates that this is not necessarily the ideal strategy in terms of ultimate system performance.

In general, the features evaluated in this work are an indiscriminate mix of human- and automatically-generated features; of the human-generated features, some are plausible to generate automatically, at some loss of quality (e.g. transcripts) while others are unlikely to be automatically generated in the foreseeable future (e.g. fine-grained dialog acts). Future work may focus on the effects that automatic generation of the former has on overall system performance (although this may require higher-quality annotations to be useful.) For example, the detailed end-of-utterance punctuation present in the human transcripts provides valuable discriminative power (Table 3), but current ASR systems are not likely to be able to provide this level of detail. Switching to ASR output will have a negative effect on performance.

One final issue is that of utterance segmentation. The scheme used in the ICSI meeting corpus does not necessarily correspond to the ideal segmentation for other tasks. The action item annotations were performed on these segmentations, and in this study we did not attempt resegmentation, but in the future it may prove valuable to collapse,

[2]Which is not to say they are devoid of significant value—training and testing our best system on the corpus with the 590 positive classifications randomly shuffled across all utterances yields an F measure of only 4.82.

for example, successive un-interrupted utterances from the same speaker into a single utterance.

In conclusion, while overall system performance does not approach levels typical of better-studied classification tasks such as named-entity recognition, we believe that this is a largely a product of the current action item annotation quality. We believe that the feature analysis presented here is useful, for this task and for other related tasks, and that, provided with a set of more consistent action item annotations, the current system can be used as is to achieve better performance.

Acknowledgments

The authors wish to thank Dan Jurafsky, Chris Manning, Stanley Peters, Matthew Purver, and several anonymous reviewers for valuable advice and comments.

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the ICASSP*.

Paul N. Bennett and Jaime Carbonell. 2005. Detecting action-items in e-mail. In *Proceedings of SIGIR*.

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the Conference on Applied NLP*.

Paul Boersma and David Weenink. 2005. Praat: doing phonetics by computer v4.4.12 (computer program).

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of EMNLP*.

Simon Corston-Oliver, Eric Ringger, Michael Gamon, and Richard Campbell. 2004. Task-focused summarization of email. In *Text Summarization Branches Out: Proceedings of the ACL Workshop*.

J. Godfrey, E. Holliman, and J.McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICAASP*.

Alexander Gruenstein, John Niekrasz, and Matthew Purver. 2005. Meeting structure annotation: Data and tools. In *Proceedings of the 6th SIGDIAL Workshop on Discourse and Dialogue*.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the ICASSP*.

Gang Ji and Jeff Bilmes. 2005. Dialog act tagging using graphical models. In *Proceedings of the ICASSP*.

Marion Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, E.G. Schukat-Talamazzini, and V. Warnke. 1996. Dialog act classification with the help of prosody. In *Proceedings of the ICSLP*.

Matthew Purver, Patrick Ehlen, and John Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In *Proceedings of the 3rd Joint Workshop on MLMI*.

Elizabeth Shriberg, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van EssDykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3–4):439–487.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGDIAL Workshop on Discourse and Dialogue*.

Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van EssDykema. 1998. Dialog act modeling for conversational speech. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP*.

Britta Wrede and Elizabeth Shriberg. 2003. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of the European Conference on Speech Communication and Technology*.

Matthias Zimmermann, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2005. Toward joint segmentation and classification of dialog acts in multiparty meetings. In *Proceedings of the 2nd Joint Workshop on MLMI*.