# Detecting and Summarizing Action Items in Multi-Party Dialogue*

**Matthew Purver**[1], **John Dowding**[2], **John Niekrasz**[1], **Patrick Ehlen**[1], **Sharareh Noorbaloochi**[1] and **Stanley Peters**[1]

[1]Center for the Study of Language and Information
Stanford University, Stanford, CA, 94305 USA
{mpurver, niekrasz, ehlen, sharare, peters}@stanford.edu

[2]University of California at Santa Cruz
Santa Cruz, California, USA
jdowding@ucsc.edu

## Abstract

This paper addresses the problem of identifying *action items* discussed in open-domain conversational speech, and does so in two stages: firstly, detecting the subdialogues in which action items are proposed, discussed and committed to; and secondly, extracting the phrases that accurately capture or summarize the tasks they involve. While the detection problem is hard, we show that we can improve accuracy by taking account of dialogue structure. We then describe a semantic parser that identifies potential summarizing phrases, and show that for some task properties these can be more informative than plain utterance transcriptions.

## 1 Introduction

Multi-party conversation, usually in the form of meetings, is the primary way to share information and make decisions in organized work environments. There is growing interest in the development of automatic methods to extract and analyze the information content of meetings in various ways, including automatic transcription, targeted browsing, and topic detection and segmentation – see (Stolcke et al., 2005; Tucker and Whittaker, 2005; Galley et al., 2003), amongst others.

In this paper we are interested in identifying *action items* – public commitments to perform a given task – both in terms of *detecting the subdialogues* in which those action items are discussed (along with the roles certain utterances perform in that discussion), and of *producing useful descriptive summaries* of the tasks they involve. While these summaries are the obvious end product in the first instance (perhaps presented as an automatically-prepared to-do list), subdialogue detection is also a useful output *per se*, as it allows users to browse the meeting recording or transcript in a targeted way.

Section 3 discusses the detection of subdialogues – short passages of conversation in which the action items are typically discussed, summarized, agreed and committed to – using a hierarchical classifier which exploits local dialogue structure. Multiple independent sub-classifiers are used to detect utterances which play particular roles in the dialogue (e.g. agreement or commitment), and an overall super-classifier then detects the critical passages based on patterns of these roles. We show that this method performs better than a flat, utterance-based approach; as far as we are aware, these are the first results for this task on realistic data.

Section 4 then investigates the production of summaries. For this, we use an open-domain semantic parser to extract phrases from within the utterances which describe one of two important properties: the *task* itself and the *timeframe* over which it is to be performed. We describe how such a parser can be built from generally available lexical resources and tailored to the particular problem of parsing speech recognition output, and show how a regression model can be used to rank the candidate parser outputs. For the timeframes, this produces

*This work was supported by the CALO project (DARPA grant NBCH-D-03-0010). We also thank Gokhan Tür, Andreas Stolcke and Liz Shriberg for provision of ASR output and dialogue act tags for the ICSI corpus.

more informative results than the alternative of presenting the entire 1-best utterance transcriptions.

## 2 Background

**Subdialogue Detection** User studies show that participants regard action items as one of a meeting's most important outputs (Lisowska et al., 2004; Banerjee et al., 2005). However, spoken action item detection seems to be a relatively new task. There is related work with email text: (Corston-Oliver et al., 2004; Bennett and Carbonell, 2005) both showed success classifying sentences or entire messages as action item- or task-related. Performance was reasonable, with f-scores around 0.6 for sentences and 0.8 for whole messages; the features used included lexical, syntactic and semantic features (n-grams, PoS-tags, named entities) as well as more email-specific features (e.g. header information).

However, applying the same methods to dialogue data is problematic. Morgan et al. (2006) applied a similar method to a portion of the ICSI Meeting Corpus (Janin et al., 2003) annotated for action items by Gruenstein et al. (2005). While they found that similar lexical, syntactic and contextual features were useful (together with other dialogue-specific features, including dialogue act type and prosodic information), performance was poor, with f-scores limited to approximately 0.3, even given manual transcripts and dialogue act tags. One major reason for this is the fragmented nature of conversational decision-making: in contrast to email text, the descriptions of tasks and their properties tend not to come in single sentences, but may be distributed over many utterances. These utterances may take many different forms and play very distinct roles in the dialogue (suggestions, commitments, (dis)agreements, etc.) and thus form a rather heterogeneous set on which it is hard to achieve good overall classification performance. For the same reasons, human annotators also have trouble deciding which utterances are relevant: Gruenstein et al. (2005)'s inter-annotator agreement was as low as $\kappa = 0.36$.

In (Purver et al., 2006), we proposed an approach to this problem using individual classifiers to detect a set of distinct action item-related utterance classes: *task description*, *timeframe*, *ownership* and *agreement*. The more homogeneous nature of these classes seemed to produce better classification accuracy, and action item discussions could be hypothesized using a simple heuristic to detect clusters of multiple classes. However, this was only evaluated on a small corpus of simulated meetings (5 c.10-minute meetings, simulated by actors given a detailed scenario), and only on gold-standard manual transcriptions. The first half of this paper applies that proposal to a larger, less domain-specific, naturally-occurring dataset, and also extends it to include the learning of a super-classifier from data.

Note that while previous work in the detection and modelling of *decisions* (Verbree et al., 2006; Hsueh and Moore, 2007) is related, the tasks are not the same. Firstly, our job is to identify public commitments to tasks, rather than general decisions about strategy, or decisions not to do anything (see e.g. Hsueh and Moore (2007)'s example Fig. 1). Secondly, our data is essentially open-domain, making e.g. simple lexical cues less useful than they are in a domain with repeated fixed topics. Note also that our results are not directly comparable with those of Hsueh and Moore (2007), who detect decision-making acts from a human-extracted summary rather than a raw meeting transcript, making positive examples much less sparse.

**Summarization & Phrase Extraction** Detecting subdialogues and utterances, though, is only part of the task – we need a succinct summary if we are to present a list of action items to a user. Ideally, this summary should contain at least the identity of the owner, a description of the task, and a specification of the timeframe. Ownership may occasionally be expressed by explicit use of a name, but is more often specified through the interaction itself – proposals of ownership usually either volunteer the speaker *"I guess I'll ..."* or request commitment from the addressee *"Could you maybe ..."*. Establishing identity therefore becomes a problem of speaker and addressee identification, which we leave aside for now, but see e.g. (Katzenmaier et al., 2004; Jovanovic et al., 2006; Gupta et al., 2007).

Timeframe and task, however, are expressed explicitly; but detecting the relevant utterances only gets us part of the way. Example (1) shows an utterance containing a task description:

(1) *What I have down for action items is we're sup-*

*posed to find out about our human subject*

Arguably the best phrase within this utterance to describe the task is *find out about our human subject*, as opposed to other larger or smaller phrases. Notably, although the utterance contains the phrase *action item* — likely a strong clue to the detection of this utterance as action item-related — this phrase itself is not particularly useful in a summary.

## 3 Subdialogue Detection

### 3.1 Approach

Following the proposal of (Purver et al., 2006), the insight we intend to exploit is that while the relevant utterances may be hard to identify on their own, the subdialogues which contain them do have characteristic structural patterns. Example (2) illustrates the idea: no single utterance contains a complete description of the task, and while some features (the phrases *by uh Tuesday* and *send it*, perhaps) might suggest action items, they may be equally likely to appear in unrelated utterances. However, the structure gives us more to go on: A proposes something involving B's agency, B considers it, and finally B agrees and commits to something.

(2) A: Well maybe by uh Tuesday you could
B: Uh-huh
A: revise the uh
C: proposal
B: Mmm Tuesday let's see
A: and send it around
B: OK sure sounds good

There are two ways in which this might help us with the detection task. Firstly, if these *action-item-specific dialogue acts* (AIDAs) form more homogeneous sets than the general class of "action-item-related utterance", we should be able to detect them more reliably. Secondly, if they are more-or-less independent, we can use the co-occurrence of multiple act types to increase our overall subdialogue detection accuracy.[1]

### 3.2 Data

Following (Purver et al., 2006), we take the relevant AIDA classes to be:

[1] In fact, there is a third: the different information associated with each act type helps in summarization – but see below.

| | | |
|---|---|---|
| D | *description* | discussion of the task to be performed |
| T | *timeframe* | discussion of the required timeframe |
| O | *owner* | assignment of responsibility (to self or other) |
| A | *agreement* | explicit agreement or commitment |

Table 1: Action item dialogue act (AIDA) classes.

We annotated 18 meetings from the ICSI Meeting Corpus (Janin et al., 2003), recordings of naturally-occurring research group meetings. The meetings are divided up by subject area; our set contains 12 from one area and 6 from 4 further areas. Three authors annotated between 9 and 13 meetings each, with all three overlapping on 3 meetings and two overlapping on a further 4. Inter-annotator agreement improved significantly on (Gruenstein et al., 2005), with pairwise $\kappa$ values for each individual AIDA class from 0.64 to 0.78. Positive examples are sparser, though, with only 1.4% of utterances being marked with any AIDA class. Note that while utterances can perform multiple AIDAs (see (2) above), there is a large degree of independence between the class distributions. Cosine distances between the distributions show high independence between A and all other classes, and reasonable independence for all other pairings except perhaps D-O (here, 0 represents total independence, 1 exact correlation):

| A-T | A-D | A-O | T-D | T-O | D-O |
|---|---|---|---|---|---|
| 0.06 | 0.03 | 0.07 | 0.23 | 0.29 | 0.55 |

Table 2: Between-class cosine distances.

### 3.3 Experiments

We trained 4 independent classifiers for the detection of each individual AIDA class; features were derived from various properties of the utterances in context (see below). We then trained a super-classifier, whose features were the hypothesized class labels and confidence scores from the sub-classifiers, over a 10-utterance window. In all cases, we performed 18-fold cross-validation, with each fold training on 17 meetings and testing on the remaining 1. All classifiers were linear-kernel support

vector machines, using *SVMlight* (Joachims, 1999).

We can evaluate performance at two levels: firstly, the accuracy of the individual AIDA sub-classifiers, and secondly, the resulting accuracy of the super-classifier in detecting subdialogue regions. The sub-classifiers can be evaluated on a per-utterance basis; it is less obvious how to evaluate the super-classifier as it detects windows rather than utterances, and we would like to give credit for windows which overlap with gold-standard subdialogues even if not matching them exactly. We therefore use two metrics; one divides the discourse into 30-second windows and evaluates on a per-window basis; one evaluates on a per-subdialogue basis, judging hypothesized regions which overlap by more than 50% with a gold-standard subdialogue as being correct.

As a baseline, we compare to a standard flat classification approach, as taken by (Morgan et al., 2006; Hsueh and Moore, 2007); we trained a single classifier on the same annotations, but for the simple binary decision of whether an utterance is action-item-related (a member of any AIDA class) or not.

### 3.4 Features

We extracted utterance features similar to those of (Morgan et al., 2006; Hsueh and Moore, 2007): n-grams, durational and locational features from the transcriptions; general dialogue act tags from the ICSI-MRDA annotations (Shriberg et al., 2004); TIMEX temporal expression tags using MITRE's rule-based TempEx tool; and prosodic features from the audio files using Praat. We also allowed "context" features, consisting of the same utterance features (suitably indexed) from the immediately preceding 5 utterances. Table 3 shows the complete set.

| | |
|---|---|
| Lexical | ngrams length 1-3 |
| Utterance | length in words & duration in seconds<br>percentage through meeting |
| Prosodic | pitch & intensity min/max/mean/deviation<br>pitch slope<br>number of voiced frames |
| TIMEX | Number of time expression tags |
| MRDA | MRDA dialogue act class |
| Context | features as above for utts $i-1 \ldots i-5$ |

Table 3: Features for subdialogue detection.

However, use of lexical and dialogue act features brings up the question of robustness: ASR word error rates are high in this domain, and general dialogue act tagging accuracy low (Ang et al., 2005). We therefore investigated the use of ASR output (obtained using SRI's Decipher (Stolcke et al., 2005)) for lexical features, both via 1-best transcriptions and word confusion networks (WCNs), which encode multiple scored hypotheses for each word (Tür et al., 2002).[2] We also examined performance both with and without MRDA dialogue act tag features.

### 3.5 Results

**Overall** Performance with unigram, utterance and context features is shown in Table 4. While per-utterance results are still low (f-scores all below 0.3), commensurate with Morgan et al. (2006)'s results with flat classification, we see that the use of the super-classifier to detect subdialogue regions does give us results which might be of practical use, with overlap f-scores near 0.5. Words were the most useful feature, with no improvement gained by increasing n-gram length above 1; prosodic features give no improvement. While MRDA and TIMEX features do give small improvements at the sub-classifier level, we see no overall subdialogue accuracy gain – we are currently investigating whether super-classifier improvements can help with this.[3]

| | *Sub-classifiers* | | | | *Super-classifier* | |
|---|---|---|---|---|---|---|
| | D | T | O | A | 30sec | Overlap |
| Recall | .19 | .15 | .21 | .18 | .51 | .59 |
| Precn. | .18 | .46 | .27 | .16 | .31 | .37 |
| F1 | .19 | .22 | .24 | .17 | .39 | .45 |

Table 4: Structured classifier; lexical + utterance features, 5-utterance context.

**Baseline comparison** Comparison with the flat baseline classifier (Table 5) shows that the structured approach gives a significant advantage; we hypothesize that this is because commitments in dialogue arise via the interaction itself as much as from individual utterances. Interestingly, although our approach consistently outperforms the baseline,

[2] While we do not know the exact ASR word error rate on our meeting set, Stolcke et al. (2005) report 24% WER on meetings from the same corpus.

[3] Note that although accuracies are much lower than those reported by Hsueh and Moore (2007), the tasks are not the same: in particular, they detect relevant dialogue acts from a manually extracted summary, rather than a whole meeting. See Section 2.

the delta decreases as more contextual information becomes available – Figure 1 shows how f-scores vary as a unigram feature set is expanded to include unigrams from preceding utterances. It may be that contextual features implicitly provide some of the structural information explicitly modelled in the structured approach. We plan to investigate this effect on larger datasets when available.

| | *30sec* | | | *Overlap* | | |
|---|---|---|---|---|---|---|
| | Re | Pr | F1 | Re | Pr | F1 |
| Structured | .51 | .31 | .39 | .59 | .37 | .45 |
| Flat | .65 | .23 | .34 | .64 | .24 | .35 |

Table 5: Classifier comparison; lexical + utterance features, 5-utterance context.

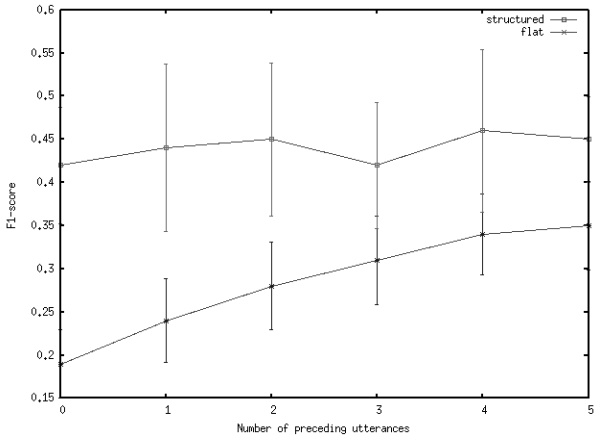


Figure 1: F-scores for structured vs. flat classifiers with 95% confidence bars; unigram features from increasing numbers of utterances in context.

**Robustness** Investigation of the effect of ASR output shows a drop in overlap f-score of 8-9% (absolute) or 17-20% (relative) – see Table 6. Use of WCNs improves over 1-best hypotheses by 1-2%. While this is a large drop, we are encouraged by the fact that this overall loss in accuracy is smaller than the loss at the sub-classifier level, where f-scores drop by around 35% on average, and up to 50% (relative). This suggests that the presence of multiple independent sub-classifiers is able (to some extent, at least) to make up for the drop in their individual performance. As more data becomes available and sub-classifier performance becomes more robust, we anticipate better overall results.

| | Structured | | | | | Flat | |
|---|---|---|---|---|---|---|---|
| | *Sub-classifiers* | | | | *Super* | | |
| | D | T | O | A | O'lap | Utt | O'lap |
| Manual | .19 | .22 | .24 | .17 | .45 | .24 | .35 |
| 1-best | .16 | .19 | .15 | .11 | .36 | .19 | .32 |
| WCNs | .15 | .14 | .18 | .07 | .37 | .19 | .33 |

Table 6: F1-scores against ASR type; lexical + utterance features, 5-utterance context.

Comparison to the baseline flat classifier shows that the structured approach is less robust (unsurprisingly, perhaps, given its more complex nature); the relative drop in the baseline overlap f-scores is lower. However, the resulting absolute performances are still higher for the structured approach, although the difference is no longer statistically significant over the number of meetings we have.

**Summary** We see that using our discourse-structural approach gives significantly improved performance over a comparable flat approach when using manual transcripts. While there is a drop in performance when using (highly errorful) ASR output, performance is still above the baseline.

## 4 Parsing and Summarization

We now turn to the second task: extracting useful phrases for summarization.

### 4.1 Approach

To extract timeframe and task descriptions, we exploit the fact that the critical phrases which contain them display certain characteristic syntactic and semantic features. Since the meeting topics and tasks are not known in advance, we expect that any approach which learns these features purely from a training set is unlikely to generalize well to unseen data. We therefore use a general rule-based parser with an open-domain, broad-coverage lexicon. The grammar, however, is small: as our data is highly ungrammatical, disfluent and errorful, we have developed a semantic parser that attempts to find basic predicate-argument structures of the major phrase types S, VP, NP, and PP, not necessarily trying to find larger structures (such as coordination and relative clauses) where reliability would be low.

**Lexical Resources** Our lexicon is built from publicly available lexical resources for English, including COMLEX, VerbNet, WordNet, and NOMLEX. Others have shared this basic approach (Shi and Mihalcea, 2005; Crouch and King, 2005; Swift, 2005).

COMLEX (Grishman et al., 1994) provides detailed morphological and syntactic information for the 40,000 most common words of English, as well as basic lexical information (e.g. adjective gradability, verb subcategorization, noun mass/count nature). VerbNet (Kipper et al., 2000) provides semantic information for 5,000 verbs, including frames and thematic roles, along with syntactic mappings and selectional restrictions for role fillers. WordNet (Miller, 1995) then provides us with another 15,539 nouns, and the semantic class information for all nouns. These semantic classes are hand-aligned to the selectional classes used in VerbNet, based on the upper ontology of EuroWordNet (Vossen, 1997). NOMLEX (Macleod et al., 1998) provides syntactic information for event nominalizations and a mapping from noun arguments to VerbNet syntactic positions; this allows us to give nominalizations a semantics compatible with verb events, and assert selectional restrictions. To add proper names, we used US Census data for people, KnowItAll (Downey et al., 2007) for companies, and WSJ data for person and organization names. Proper names account for about 1/3 of the entries in the lexicon.

These resources are combined and converted to the Prolog-based format used in the Gemini framework (Dowding et al., 1993), which includes a fast bottom-up robust parser in which syntactic and semantic information is applied interleaved. To facilitate extracting semantic features, we use Minimal Recursion Semantics (Copestake et al., 2005), a flat semantic representation; we have also modified Gemini to parse WCNs as well as flat transcriptions. Gemini computes parse probabilities on the context-free background of the grammar; in these experiments, probabilities were trained on WSJ data.

### 4.2 Experiments

Our parsing approach intentionally produces multiple short fragments rather than one full utterance parse. Combining this with the high number of paths through a WCN means that our primary problem is to extract a few useful phrases from amongst a very high number of alternatives. We approached this as a regression problem, and attempted to learn a model to rank phrases according to their likelihood of appearing in an action item description (again using *SVMlight*). We cross-validated over the same 18-meeting dataset, considering only those utterances manually annotated as containing timeframe and task descriptions (the T and D AIDA classes). To provide target phrases for evaluation, annotators marked those portions of the manual utterance transcriptions which should be extracted (note that these often do not match any WCN path exactly).

For each segment returned by the parser we extracted features of three general types: properties of the raw WCN paths, properties of the parsed phrases, and lexical features reflecting the identity of the words themselves – a list is given in Table 7. As lexical features are likely to be more domain-specific, and increase the size of the feature space dramatically, we prefer to avoid them if possible. Initial feature selection experiments indicate that the most useful features are acoustic probability, phrase type and verb semantic class, suggesting that syntactic and semantic information are indeed valuable.

| | |
|---|---|
| WCN | phrase length (words & WCN arcs)<br>start/end point (absolute & percentage)<br>acoustic probability<br>acoustic probability shortfall (delta below highest probability for this segment) |
| Parse | parse probability<br>phrase type (S/VP/NP/PP)<br>main verb VerbNet class<br>head noun WordNet synset<br>nominalization (yes, no)<br>number of thematic roles filled<br>noun class of *agent* thematic role (if any) |
| Lexical | main verb<br>head noun<br>all unigrams in the phrase |
| TIMEX | Number of time expression tags |

Table 7: Features for parse fragment ranking.

### 4.3 Results

Choosing an evaluation metric is not straightforward: standard parse evaluation methods (e.g. checking crossing brackets against a treebank) are not applicable to our task of choosing useful fragments. Instead, we evaluate success based on how much of the human-annotated task descriptions are covered by the top-ranked fragment chosen by the

regression model. For recall we take the total proportion of the desired description covered; for precision, the total proportion of the chosen phrase which overlaps with the desired description; we then produce a corresponding f-score. We compare to a baseline of using the entire 1-best utterance transcription, and the ideal ceiling of choosing the fragment with the best f-score (still less than 1, due to ASR errors and parse segmentation). For timeframe utterances, we also compare to a second baseline of using those fragments of the 1-best transcription tagged as TIMEX expressions.

Results are shown in Table 8 for *timeframe* phrases, and Table 9 for *task description* phrases. For timeframes, the best feature set gives an f-score of .51 and precision of .62, outperforming both baselines but still some way below the ideal ceiling. Semantic classes and phrase-head lexical features help performance, although including other unigrams did not; TIMEX tags help, although a TIMEX-only baseline does badly.

| | Recall | Precision | F1 |
|---|---|---|---|
| Baseline 1: TIMEX | .26 | .36 | .31 |
| Baseline 2: 1-best | .76 | .27 | .39 |
| No sem/lex features | .33 | .47 | .38 |
| + semantic classes | .36 | .53 | .43 |
| + head verb/noun | .39 | .59 | .47 |
| + TIMEX | .43 | .62 | .51 |
| Ceiling: best F1 | .64 | .80 | .71 |

Table 8: Fragment ranking results: *timeframe*.

However, results for description phrases are poor, with no feature set outperforming the baseline. This is partly as the baseline recall is already quite high; note that using the parser does increase precision. Lexical features actually harm performance, perhaps unsurprisingly given the wider range of vocabulary compared to timeframes. The problem is also more difficult, hence the ideal figures are lower too; but inspection of errors suggests that inaccurate sentence segmentation (based only on pause length in these data) causes many of the problems, with many utterances annotated as providing only single words to the ideal phrase. We expect that improved sentence segmentation will improve performance, and are currently investigating this.

| | Recall | Precision | F1 |
|---|---|---|---|
| Baseline: 1-best | .66 | .32 | .43 |
| No sem/lex features | .22 | .41 | .29 |
| + semantic classes | .35 | .41 | .38 |
| + head verb/noun | .31 | .41 | .35 |
| Ceiling: best F1 | .50 | .78 | .61 |

Table 9: Fragment ranking results: *description*.

## 5 Conclusions & Future Work

Both problems are hard, and overall performance is correspondingly lower than that achieved on less difficult tasks or less sparse data. However, they do appear tractable, even on errorful ASR output, with some encouraging initial performances obtained. Importantly, we have shown the benefits of using discourse structure in classification, and semantic features in summarization.

To improve detection performance, we are investigating more effective super-classifiers, incorporating existing task lists to provide reliable information about possible tasks to be discussed, and leveraging user interaction for learning – allowing users to confirm, delete or edit hypothesized action items, and using this as feedback to allow incremental learning (Purver et al., 2007).

For summarization, one of the major limitations of our approach is that we only consider phrases from within a single acoustically-segmented utterance, while many ideal descriptions combine information from more than one. We plan to investigate improved segmentation, and generation of summaries from multiple utterances.

## References


J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of ICASSP*.

S. Banerjee, C. Rosé, and A. Rudnicky. 2005. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human-Computer Interaction (CHI)*.

P. N. Bennett and J. Carbonell. 2005. Detecting action-items in e-mail. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

A. Copestake, D. Flickinger, C. Pollard, and I. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2–3):281–332.

S. Corston-Oliver, E. Ringger, M. Gamon, and R. Campbell. 2004. Task-focused summarization of email. In *Proceedings of the 2004 ACL Workshop Text Summarization Branches Out*.

R. Crouch and T. King. 2005. Unifying lexical resources. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

J. Dowding, J. M. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, and D. Moran. 1993. GEMINI: a natural language system for spoken-language understanding. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*.

D. Downey, M. Broadhead, and O. Etzioni. 2007. Locating complex named entities in web text. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

R. Grishman, C. Macleod, and A. Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*.

A. Gruenstein, J. Niekrasz, and M. Purver. 2005. Meeting structure annotation: data and tools. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*.

S. Gupta, M. Purver, and D. Jurafsky. 2007. Disambiguating between generic and referential "you" in dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*.

P.-Y. Hsueh and J. Moore. 2007. What decisions have you made?: Automatic decision detection in meeting conversations. In *Proceedings of NAACL/HLT*.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus. In *Proceedings of the 2003 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*. MIT Press.

N. Jovanovic, R. op den Akker, and A. Nijholt. 2006. Addressee identification in face-to-face meetings. In *Proceedings of the 11th Conference of the European Chapter of the ACL (EACL)*.

M. Katzenmaier, R. Stiefelhagen, and T. Schultz. 2004. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces*.

K. Kipper, H. T. Dang, and M. Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*.

A. Lisowska, A. Popescu-Belis, and S. Armstrong. 2004. User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

C. Macleod, R. Grishman, A. Meyers, L. Barrett, and R. Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proceedings of EURALEX 98*.

G. A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

W. Morgan, P.-C. Chang, S. Gupta, and J. M. Brenier. 2006. Automatically detecting action items in audio meeting recordings. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*.

M. Purver, P. Ehlen, and J. Niekrasz. 2006. Detecting action items in multi-party meetings: Annotation and initial experiments. In S. Renals, S. Bengio, and J. Fiscus, editors, *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Revised Selected Papers*, volume 4299 of *Lecture Notes in Computer Science*, pages 200–211. Springer.

M. Purver, J. Niekrasz, and P. Ehlen. 2007. Automatic annotation of dialogue structure from simple user interaction. In *Proceedings of the 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI'07)*.

L. Shi and R. Mihalcea. 2005. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*.

A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng. 2005. Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system. In *Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation*.

M. Swift. 2005. Towards automatic verb acquisition from VerbNet for spoken dialog processing. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

S. Tucker and S. Whittaker. 2005. Reviewing multimedia meeting records: Current approaches. In *Proceedings of the 2005 (ICMI) International Workshop on Multimodal Multiparty Meeting Processing*.

G. Tür, J. Wright, A. Gorin, G. Riccardi, and D. Hakkani-Tür. 2002. Improving spoken language understanding using word confusion networks. In *Proceedings of the 7th International Conference on Spoken Language Processing (INTERSPEECH - ICSLP)*.

A. Verbree, R. Rienks, and D. Heylen. 2006. First steps towards the automatic construction of argument-diagrams from real discussions. In *Proceedings of the 1st International Conference on Computational Models of Argument, September 11 2006, Frontiers in Artificial Intelligence and Applications*, volume 144.

P. Vossen. 1997. EuroWordNet: a multilingual database for information retrieval. In *Proceedings of the 1997 DELOS Workshop on Cross-language Information Retrieval*.