Shallow Discourse Structure for Action Item Detection

Matthew Purver, Patrick Ehlen, and John Niekrasz
Center for the Study of Language and Information
Stanford University
Stanford, CA 94305
{mpurver,ehlen,niekrasz}@stanford.edu

Abstract

We investigated automatic *action item detection* from transcripts of multi-party meetings. Unlike previous work (Gruenstein et al., 2005), we use a new hierarchical annotation scheme based on the roles utterances play in the action item assignment process, and propose an approach to automatic detection that promises improved classification accuracy while enabling the extraction of useful information for summarization and reporting.

1 Introduction

Action items are specific kinds of decisions common in multi-party meetings, characterized by the concrete assignment of tasks together with certain properties such as an associated timeframe and reponsible party. Our aims are firstly to automatically detect the regions of discourse which establish action items, so their surface form can be used for a targeted report or summary; and secondly, to identify the important properties of the action items (such as the associated tasks and deadlines) that would foster concise and informative semantically-based reporting (for example, adding task specifications to a user's to-do list). We believe both of these aims are facilitated by taking into account the roles different utterances play in the decision-making process – in short, a shallow notion of discourse structure.

2 Background

Related Work Corston-Oliver et al. (2004) attempted to identify action items in e-mails, using classifiers trained on annotations of individual sentences within each e-mail. Sentences were annotated with one of a set of "dialogue" act classes; one class `Task` included any sentence containing items that seemed appropriate to add to an ongoing to-do list. They report good inter-annotator agreement over their general tagging exercise ($\kappa > 0.8$), although individual figures for the `Task` class are not given. They then concentrated on `Task` sentences, establishing a set of predictive features (in which word n-grams emerged as "highly predictive") and achieved reasonable per-sentence classification performance (with f-scores around 0.6).

While there are related tags for dialogue act tagging schema – like DAMSL (Core and Allen, 1997), which includes tags such as `Action-Directive` and `Commit`, and the ICSI MRDA schema (Shriberg et al., 2004) which includes a `commit` tag – these classes are too general to allow identification of action items specifically. One comparable attempt in spoken discourse took a flat approach, annotating utterances as action-item-related or not (Gruenstein et al., 2005) over the ICSI and ISL meeting corpora (Janin et al., 2003; Burger et al., 2002). Their inter-annotator agreement was low ($\kappa = .36$). While this may have been partly due to their methods, it is notable that (Core and Allen, 1997) reported even lower agreement ($\kappa = .15$) on their `Commit` dialogue acts. Morgan et al. (forthcoming) then used these annotations to attempt auto-

Proceedings of the Analyzing Conversations in Text and Speech (ACTS) Workshop at HLT-NAACL 2006, pages 31–34,
New York City, New York, June 2006.

matic classification, but achieved poor performance (with f-scores around 0.3 at best).

Action Items Action items typically embody the transfer of group responsibility to an individual. This need not be the person who actually performs the action (they might delegate the task to a subordinate), but publicly commits to seeing that the action is carried out; we call this person the *owner* of the action item. Because this action is a social action that is coordinated by more than one person, its initiation is reinforced by *agreement* and uptake among the owner and other participants that the action should and will be done. And to distinguish this action from immediate actions that occur during the meeting and from more vague future actions that are still in the planning stage, an action item will be specified as expected to be carried out within a *timeframe* that begins at some point after the meeting and extends no further than the not-too-distant future. So an action item, as a type of social action, often comprises four components: a *task description*, a *timeframe*, an *owner*, and a round of *agreement* among the owner and others. The related discourse tends to reflect this, and we attempt to exploit this fact here.

3 Baseline Experiments

We applied Gruenstein et al. (2005)'s flat annotation schema to transcripts from a sequence of 5 short related meetings with 3 participants recorded as part of the CALO project. Each meeting was simulated in that its participants were given a scenario, but was not scripted. In order to avoid entirely data- or scenario-specific results (and also to provide an acceptable amount of training data), we then added a random selection of 6 ICSI and 1 ISL meetings from Gruenstein et al. (2005)'s annotations. Like (Corston-Oliver et al., 2004) we used support vector machines (Vapnik, 1995) via the classifier *SVM-light* (Joachims, 1999). Their full set of features are not available to us, but we experimented with combinations of words and n-grams and assessed classification performance via a 5-fold validation on each of the CALO meetings. In each case, we trained classifiers on the other 4 meetings in the CALO sequence, plus the fixed ICSI/ISL training selection. Performance (per utterance, on the binary classification problem) is shown in Table 1; overall f-score figures are poor even on these short meetings. These figures were obtained using words (unigrams, after text normalization and stemming) as features – we investigated other discriminative classifier methods, and the use of 2- and 3-grams as features, but no improvements were gained.

Mtg.	Utts	AI Utts.	Precision	Recall	F-Score
1	191	22	0.31	0.50	0.38
2	156	27	0.36	0.33	0.35
3	196	18	0.28	0.55	0.37
4	212	15	0.20	0.60	0.30
5	198	9	0.19	0.67	0.30

Table 1: Baseline Classification Performance

4 Hierarchical Annotations

Two problems are apparent: firstly, accuracy is lower than desired; secondly, identifying utterances related to action items does not allow us to actually identify those action items and extract their properties (deadline, owner etc.). But if the utterances related to these properties form distinct sub-classes which have their own distinct features, treating them separately and combining the results (along the lines of (Klein et al., 2002)) might allow better performance, while also identifying the utterances where each property's value is extracted. Thus, we produced an annotation schema which distinguishes among these four classes. The first three correspond to the discussion and assignment of the individual properties of the action item (`task description`, `timeframe` and `owner`); the final `agreement` class covers utterances which explicitly show that the action item is agreed upon.

Since the `task description` subclass extracts a description of the task, it must include any utterances that specify the action to be performed, including those that provide required antecedents for anaphoric references. The `owner` subclass includes any utterances that explicitly specify the responsible party (e.g. "I'll take care of that", or "John, we'll leave that to you"), but not those whose function might be taken to do so implicitly (such as agreements by the responsible party). The `timeframe` subclass includes any utterances that explicitly refer to when a task may start or when it is expected to be finished; note that this is often not specified with

a date or temporal expression, but rather e.g. "by the end of next week," or "before the trip to Aruba". Finally, the `agreement` subclass includes any utterances in which people agree that the action should and will be done; not only acknowledgements by the owner themselves, but also when other people express their agreement.

A single utterance may be assigned to more than one class: "**John, you** need to do that **by next Monday**" might count as `owner` and `timeframe`. Likewise, there may be more than one utterance of each class for a single action item: John's response "OK, **I'll** do that" would also be classed as `owner` (as well as `agreement`). While we do not require all of these subclasses to be present for a set of utterances to qualify as denoting an action item, we expect any action item to include most of them.

We applied this annotation schema to the same 12 meetings. Initial reliability between two annotators on the single ISL meeting (chosen as it presented a significantly more complex set of action items than others in this set) was encouraging. The best agreement was achieved on `timeframe` utterances ($\kappa = .86$), with `owner` utterances slightly less good (between $\kappa = .77$), and `agreement` and `description` utterances worse but still acceptable ($\kappa = .73$). Further annotation is in progress.

5 Experiments

We trained individual classifiers for each of the utterance sub-classes, and cross-validated as before. For `agreement` utterances, we used a naive n-gram classifier similar to that of (Webb et al., 2005) for dialogue act detection, scoring utterances via a set of most predictive n-grams of length 1–3 and making a classification decision by comparing the maximum score to a threshold (where the n-grams, their scores and the threshold are automatically extracted from the training data). For `owner`, `timeframe` and `task description` utterances, we used SVMs as before, using word unigrams as features (2- and 3-grams gave no improvement – probably due to the small amount of training data). Performance varied greatly by sub-class (see Table 2), with some (e.g. agreement) achieving higher accuracy than the baseline flat classifications, but others being worse. As there is now significantly less training data available to each sub-class than there was for all utterances grouped together in the baseline experiment, worse performance might be expected; yet some sub-classes perform better. The worst performing class is `owner`. Examination of the data shows that `owner` utterances are more likely than other classes to be assigned to more than one category; they may therefore have more feature overlap with other classes, leading to less accurate classification. Use of relevant sub-strings for training (rather than full utterances) may help; as may part-of-speech information – while proper names may be useful features, the name tokens themselves are sparse and may be better substituted with a generic tag.

Class	Precision	Recall	F-Score
description	0.23	0.41	0.29
owner	0.12	0.28	0.17
timeframe	0.19	0.38	0.26
agreement	0.48	0.44	0.40

Table 2: Sub-class Classification Performance

Even with poor performance for some of the subclassifiers, we should still be able to combine them to get a benefit as long as their true positives correlate better than their false positives (intuitively, if they make mistakes in different places). So far we have only conducted an initial naive experiment, in which we combine the individual classifier decisions in a weighted sum over a window (currently set to 5 utterances). If the sum over the window reaches a given threshold, we hypothesize an action item, and take the highest-confidence utterance given by each sub-classifier in that window to provide the corresponding property. As shown in Table 3, this gives reasonable performance on most meetings, although it does badly on meeting 5 (apparently because no explicit agreement takes place, while our manual weights emphasized agreement).[1] Most encouragingly, the correct examples provide some useful "best" sub-class utterances, from which the relevant properties could be extracted.

These results can probably be significantly improved: rather than sum over the binary classification outputs of each classifier, we can use their confidence scores or posterior probabilities, and learn

[1] Accuracy here is currently assessed only over correct detection of an action item in a window, not correct assignment of all sub-classes.

Mtg.	AIs	Correct	False+	False-	F-Score
1	3	2	1	1	0.67
2	4	1	0	3	0.40
3	5	2	1	3	0.50
4	4	4	0	0	1.00
5	3	0	1	3	0.00

Table 3: Combined Classification Performance

the combination weights to give a more robust approach. There is still a long way to go to evaluate this approach over more data, including the accuracy and utility of the resulting sub-class utterance hypotheses.

6 Discussion and Future Work

So accounting for the structure of action items appears essential to detecting them in spoken discourse. Otherwise, classification accuracy is limited. We believe that accuracy can be improved, and the detected utterances can be used to provide the properties of the action item itself. An interesting question is how and whether the structure we use here relates to discourse structure in more general use. If a relation exists, this would shed light on the decision-making process we are attempting to (begin to) model, and might allow us to use other (more plentiful) annotated data.

Our future efforts focus on annotating more meetings to obtain large training and testing sets. We also wish to examine performance when working from speech recognition hypotheses (as opposed to the human transcripts used here), and the best way to incorporate multiple hypotheses (either as n-best lists or word confusion networks). We are actively investigating alternative approaches to sub-classifier combination: better performance (and a more robust and trainable overall system) might be obtained by using a Bayesian network, or a maximum entropy classifier as used by (Klein et al., 2002). Finally, we are developing an interface to a new large-vocabulary version of the Gemini parser (Dowding et al., 1993) which will allow us to use semantic parse information as features in the individual sub-class classifiers, and also to extract entity and event representations from the classified utterances for automatic addition of entries to calendars and to-do lists.

References

S. Burger, V. MacLaren, and H. Yu. 2002. The ISL Meeting Corpus: The impact of meeting type on speech style. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2002)*.

M. Core and J. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In D. Traum, editor, *AAAI Fall Symposium on Communicative Action in Humans and Machines*.

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

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 *Proc. 31st Annual Meeting of the Association for Computational Linguistics*.

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*.

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 *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*.

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.

D. Klein, K. Toutanova, H. T. Ilhan, S. D. Kamvar, and C. D. Manning. 2002. Combining heterogeneous classifiers for word-sense disambiguation. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*.

W. Morgan, S. Gupta, and P.-C. Chang. forthcoming. Automatically detecting action items in audio meeting recordings. Ms., under review.

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

S. Siegel and J. N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

V. N. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.

N. Webb, M. Hepple, and Y. Wilks. 2005. Dialogue act classification using intra-utterance features. In *Proc. AAAI Workshop on Spoken Language Understanding*.