Aspectual Properties of Conversational Activities

Rebecca J. Passonneau and **Boxuan Guan** and **Cho Ho Yeung**
becky@ccls.columbia.edu and bg2469@columbia.edu and cy2277@columbia.edu
Columbia University, New York, NY, USA

Yuan Du
ydu@fb.com
Facebook, New York, NY, USA

Emma Conner
econner@oberlin.edu
Oberlin College, Oberlin, OH, USA

Abstract

Segmentation of spoken discourse into distinct conversational activities has been applied to broadcast news, meetings, monologs, and two-party dialogs. This paper considers the aspectual properties of discourse segments, meaning how they transpire in time. Classifiers were constructed to distinguish between segment boundaries and non-boundaries, where the sizes of utterance spans to represent data instances were varied, and the locations of segment boundaries relative to these instances. Classifier performance was better for representations that included the end of one discourse segment combined with the beginning of the next. In addition, classification accuracy was better for segments in which speakers accomplish goals with distinctive start and end points.

1 Introduction

People engage in dialogue to address a wide range of goals. It has long been observed that discourse can be structured into units that correspond to distinct goals and activities (Grosz and Sidner, 1986; Passonneau and Litman, 1997). This is conceptually distinct from structuring discourse into the topical units addressed in (Hearst, 1997). The ability to recognize where distinct activities occur in spoken discourse could support offline applications to spoken corpora such as search (Ward and Werner, 2013), summarization (Murray et al., 2005), and question answering. Further, a deeper understanding of the relation of conversational activities to observable features of utterance sequences could inform the design of interactive systems for online applications such as information gathering, service requests, tutoring, and companionship. Automatic identification of such units, however, has been difficult to achieve. This paper considers the *aspectual* properties of speakers' conversational activities, meaning how they transpire in time. We hypothesize that recognition of a transition to a new conversational activity depends on recognizing not only the start of a new activity but also the end of the preceding one, on the grounds that the relative contrast between endings and beginnings might matter as much or more than absolute characteristics consistent across all beginnings or all endings. We further hypothesize that transitions to certain kinds of conversational activity may be easier to detect than others.

Following Austin's view that speech constitutes action of different kinds (Austin, 1962), we assume that different kinds of communicative action have different ways of transpiring in time, just as other actions do. Conversational activities that address objective goals, for example, can have very well-demarcated beginnings and endings, as when two people choose a restaurant to go to for dinner. Conversational participants can, however, address goals that need not have a specific resolution, such as shared complaints about the lack of good Chinese restaurants. This distinction between different kinds of actions that speakers perform through their communicative behavior is analogous to the distinction in linguistic semantics pertaining to verbal aspect, between states, processes and transition events (or accomplishments and achievements) (Vendler, 1957) (Dowty, 1986). States (e.g., *being at a standstill*) have no perceptible change from moment to moment; processes (e.g., *walking*) have detectable differences in state from moment to moment with no clearly demarcated change of state during the process; transition events (e.g., *starting to walk; walking to the end of the block*) involve a transition from one state or process to another.

To investigate the aspectual properties of discourse segments, we constructed classifiers to de-

Proceedings of the SIGDIAL 2014 Conference, pages 228–237,
Philadelphia, U.S.A., 18-20 June 2014.

tect discourse segment boundaries based on features of utterances. We considered the aspectual properties of discourse segments in two ways. First, to investigate the relative contribution of features from segment endings versus beginnings, we experimented with different sizes of utterance sequences, and different locations of segment boundaries relative to these sequences. Second, we considered different categories of segments, based on the speculation that segment transitions that are easier to recognize would be associated with conversational activities that have a well-demarcated event structure, in constrast to activities that involve goals to maintain or sustain aspects of interaction.

The following section describes related work in this area, as well as the difficulties in achieving good performance. Most work on identification of discourse segments (or other forms of discourse structure in spoken interaction) depends on a prior phase of annotation (e.g., (Galley et al., 2003; Passonneau and Litman, 1997)). We studied a corpus of eighty-two transcribed and annotated telephone dialogues between library patrons and librarians that had been annotated with units analogous to speech acts, and subsequently annotated with discourse segments comprised of these units. The annotation yielded eight distinct kinds of discourse segment, where a segment results from a linear segmentation of a discourse into strictly sequential units. (While the segmentation is sequential, the units can have hierarchical relations.) We found that classifiers to detect segment boundaries performed best with boundaries represented by features of sequences of utterances that spanned the end of one segment and the beginning of the next. Error analysis indicated that performance was better for boundaries that initiate conversational activities with clear beginnings and endings.

2 Related Work

Segmentation of spoken language interaction into distinct discourse units has been applied to meetings as well as to two-party discourse using acoustic features, lexical features, and very heterogeneous features. In our previous work, we used a very heterogeneous set of features to segment monologues into units that had been identified by annotators as corresonding to distinct intentional units (Passonneau and Litman, 1997). Text tiling (Hearst, 1997) has been applied to segmentation of meetings into distinct agenda segments using both prior and following context (Banerjee and Rudnicky, 2006). Results had high precision and low recall. We also find that recall is more challenging than precision. Topic modeling methods have also been applied to the identification of topical segments in speech (Purver et al., 2006) (Eisenstein and Barzilay, 2008), with improvements over earlier work on the ICSI meeting corpus (Galley et al., 2003) (Malioutov and Barzilay, 2006).

An analog of text tiling that uses acoustic patterns rather than lexical items has been applied to the segmentation of speech into stories using segmental dynamic time warping (SDTW) (Park and Glass, 2008). The method is based on the intuition of aligning utterances by similar acoustic patterns, possibly representing common words and phrases. Results on TDT2 Mandarin Broadcast News corpus were moderately good for short episodes with F=0.71 beating the baseline for lexical text tiling of 0.66, but poor on long episodes.

An alternative method of relying solely on acoustic information has been applied to importance prediction at a very fine granularity (Ward and Richart-Ruiz, 2013). Four basic classes of prosodic features derived from PCA were used (Ward and Vega, 2012): volume, pitch height, pitch range and speaking rate cross various widths of time intervals. The data was labeled by annotators using an importance scale of 1 to 5, and linear regression was used to predict the label for instances consisting of frames. The method performed well with a correlation of 0.82 and mean average error of 0.75 (5-fold cross validation).

The identification of different kinds of units in discourse is somewhat related to the notion of genre identification, e.g. (Obin et al., 2010) (Ries et al., 2000). Results from this area have been applied to segmentation of conversation by a combination of topic and style (Ries, 2002).

3 Data and Annotations

The corpus consists of recordings, transcripts and annotations on the transcripts of a set of 82 calls recorded in 2005 between patrons of the Andrew Heiskell Braille and Talking Book Library of New York City.[1] An annotation for dialog acts with a

[1] The audio files and transcripts are available for download from the Columbia University Data Commons. The annotations and raw features will be released in the near future.

reduced set of dialog act types and adjacency pair relations (Dialogue Function Units, DFUs) was developed, originally for comparison of dialogues across modalities (Hu et al., 2009). A subsequent phase of annotation at the discourse level that makes use of the dialog act annotation was later applied. This later annotation, referred to as Task Success and Cost Annotation (TSCA), was aimed at identifying individual dialog tasks analogous to those carried out by spoken dialog systems, to facilitate comparison of human-human dialog with human-machine dialog. Interannotator reliability of both annotations was measured using Krippendorff's alpha (Krippendorff, 1980) at levels of 0.66 and above for individual dialogues (Passonneau et al., 2011). The corpus consists of 24,760 words, or 302 words per dialog.

Briefly, the second phase of annotation involved grouping DFUs into larger sequences in which the participants continued to pursue a single coordinated activity, and labeling the large discourse units for their discourse function. The human annotation instructions avoided reference to overt signals of dialog structure. Rather, annotators were asked to judge the semantic and pragmatic functions of utterances. The annotations have been described in previous work (Hu et al., 2009; Passonneau et al., 2011); the annotation guidelines are available online.[2]

The location of a transition between one conversational activity and the next is represented as occurring between adjacent utterances. There are 9,340 utterance in the corpus, or 114 per dialog. About 10.6 percent of the utterances (994) start a new discourse unit. Within each unit, the speakers establish a conversational goal explicitly or implicitly, and continue to address the goal until it is achieved, suspended, or abandoned. The discourse segments were of the following seven categories, with an additional *Other category* for none of the above (examples from the corpus are shown after each segment category description; words in brackets represent overlapping talk of the two speakers):

- Conventional: The participants engage in conventionalized behavior, e.g., greetings (at the beginning of the call) or goodbyes (at the end of the call).

  Librarian: andrew heiskell library
  Librarian: how are you
  Patron: good morning
  Librarian: good morning

- Book-Request: The participants address a patron's request for a book, which can be a specific book that first needs to be identified, or which can be a non-specific request for a book fitting some criterion (e.g., a mystery the patron has not read before).

  Patron: do you have any fannie flagg stories
  Librarian: flag
  Patron: yeah
  Patron: F L A <Pause>
  Patron: A G G I think it is

- Inform: One of the participants provides the other with general information that does not support a Book Request, e.g., the patron provides identifying information so the librarian can pull up the patron's record.

  Patron: well I'll call him again then
  Patron: and I'll get the name [today]
  Librarian [talk] to him and call me back
  Patron: <pause> i- i'll call him
  Patron: and then i'll call you okay
  Librarian: okay

- Librarian-Proposal: The participants address the librarian's suggestion of a specific book or a kind of book that might meet the patron's desires.

  Librarian: I have ellis but not bret
  Patron: ah wa wa what do you have by him
  Librarian: by cose
  Librarian: C O S E
  Librarian: I have the rage of a privileged class
  Patron: that's all right

- Request-Action: One of the participants asks the other to perform an action, e.g., the patron asks that certain authors be added to the patron's list of preferences

  Patron: also <pause> uh
  Patron: <pause> of the favorite author list
  Librarian: mmhm
  Patron: would you um
  Patron: remove t jefferson parker
  Librarian: okay

[2] See links at `http://www1.ccls.columbia.edu/~Loqui/resources.html` for transcription guidelines, and annotation manuals.

- Information-Request: One of the participants seeks information from the other, e.g., the patron wants to know if

  Patron: this is the talking books right
  Librarian: yes
  Librarian: this is the library for the blind

- Sidebar: The librarian temporarily takes a call from another Patron only long enough to place the new caller on hold

  Librarian: hold on one second
  Librarian: Andrew Heiskell Library
  Librarian: please hold

- Other

Of these seven kinds of discourse units, Book-Requests and Librarian-Proposals are the most clearly delimited by beginning and ending points. At the beginning of a Book-Request, the patron establishes that she wants a book, and the end is identified by the mutual achievement of the librarian and patron of either a successful resolution, meaning the identification of a particular book in the library's collection that the patron will accept, or a failure of the current attempt, which often leads to a new revised book request. Librarian-Proposals are very parallel to Book-Requests; the difference is that the librarian makes a suggestion of a specific book or kind of book which must be identified for the patron, and which the patron then accepts or rejects.

4 Experiments

The experiments to automatically identify the locations of the annotated discourse units apply machine learning to instances consisting of utterance sequences that represent the two classes, presence versus absence of a boundary. We hyothesize that the enormous challenges for identifying discourse structure in human-machine dialogue can be better addressed through complementary reliance on semantics and interaction structure (behavioral cues), and each can reinforce the other. The main focus of the experiments reported here is on data representation to address the questions, what features of the context support the ability to segment a dialogue into conversational activity units, and how much context is necessary?

A disadvantage of the dataset is its relatively small size, especially given the extreme skew with the positive class consisting of only 10% of the instances. On the other hand, the small size made detailed annotation feasible, and the corpus is well-suited to our research question in that it represents naturally occurring, spontaneous human-human telephone discourse. Therefore. the manner in which the dialogs evolve over time is entirely natural. Our major question of interest is how much of the time-course of the discourse is required for a machine learner to identify the start of a new discourse unit. To examine this question, we vary two dimensions of the representation of the instances for learning. The first is the number of utterances around the location of the start of a new discourse unit. The second is the set of features to represent each instance, which as we will see below, affects to some degree how many utterances to include before and after the start of a new discourse unit.

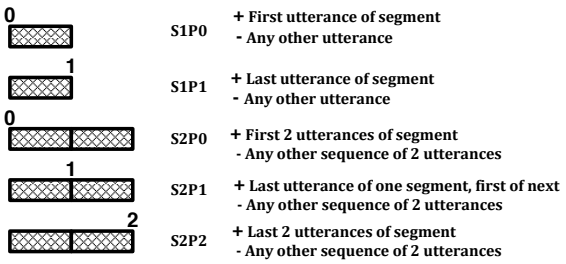

Figure 1: Schematic representation of instance spans and labels. Bars on the left show the number of utterances (size) and position of segment boundary (position) for five of the fourteen types of instances. Positive and negative labels are shown on the descriptions at the right.

Four machine learning methods were tested using the Weka toolkit (Hall et al., 2009): Naive Bayes, J48 Decision Trees, Logistic Regression and Multilayer Perceptron. Of these, J48 had the best and most consistent performance, which we speculate is due to a combination of the small size of the dataset, and non-linearity of the data. Because J48 is doing feature selection while building the tree, it can identify different threshholds for the same features, depending on the location in the tree. All results reported here are for J48.

4.1 Labels and Instance Spans

We refer to a sequence of utterances, and a potential location of the onset of a discourse unit relative to that sequence, as a *span*. We varied the num-

ber of utterances for each span from 1 to 4, and the location of the start of a new unit to be at the beginning of the first utterance, at the end of the last utterance, or between any pair of utterances in the span. For a single utterance, there will be two types of instances, as shown in Figure 1. Each instance type is represented as S<N>P<M> where N is the number of utterances in the span and M is how many utterances there are before the boundary. S1P0 denotes size 1 spans with the boundary at position 0; positively labeled instances represent the first utterance of a segment. S1P1 denotes size 1 spans with the boundary at position 1; positively labeled instances represent the last utterance of a segment. The experiments used all labelings for spans from size 1 to 4, yielding 14 types of instances. For multi-utterance spans that occur at the beginning or end of a discourse, dummy utterances are used to fill out the spans.

4.2 Features

We use three sets of features. A set we refer to as *discourse* features consists of a mixed set of acoustic features and lexicogrammatical features that have been associated with discourse structure, such as discourse cue words (Hirschberg and Litman, 1993). Table 1 lists the 35 discourse features. The second set is a bag-of-words (BOW) vector representation, and the third is the combination of the discourse and BOW features. We used alternative sets of features on the assumption that the performance of a machine learner across the different instance spans will vary, depending on the aspects of the utterance that the features capture. We see some expected differences in performance between the discourse features and BOW, with BOW benefitting more than the discourse features from longer spans. Unexpectedly, we see no gain in performance from the combination of both feature sets.

The discourse features consist of acoustic features, pause features, word and utterance length features, proper noun features and speaker change. The acoustic features and the (unfilled) pause location and duration features were extracted using Praat, a cross-platform tool for speech analysis. The features pertaining to filled pauses (e.g., *um, uh*) were extracted from the transcripts.

4.3 Conditions and Evaluation

The experimental conditions varied the feature set, the selection of training data versus testing data, and the fourteen kinds of instance spans and labels. Three feature sets consisted of the discourse features from Table 1 (discourse), bag-of-words (bow), and the combination of the two (combo). In all experiments, the data was randomly split into 75% for training, and 25% for testing, using two methods to select instances. In randomization by dialog, all utterances from a single dialog were kept together and 75% of the dialogs were selected for training. In randomization by utterance, 75% of all utterances were randomly selected for training, without regard to which dialog they came from. This was done to test the hypothesis that the bow representation would be more sensitive to changes of vocabulary across dialogs. The three feature sets, fourteen data representations and two randomization methods yield 84 experimental conditions.

While N-fold cross-validation is a popular method to estimate a classifier's prediction error, it is not a perfect substitute for isolating the training data from the test data (Ng, 1997). The cross-validation estimate of prediction error is relatively unbiased, but it can be highly variable (Efron and Tibshirani, 1997)(Rodriguez et al., 2010). To avoid the inherent risk of overfitting (Ng, 1997), one recommendation is to use cross-validation to compare models, and to reserve a test set to verify that a selected classifier has superior generalization (Rao and Fung, 2008). To assess whether performance measures of different models are genuinely different requires error bounds on the result, which is not done with cross-validation. We perform train-test splits of the data to minimize overfitting, and bootstrap confidence intervals for each classifier's accuracy (and other metrics) in order to measure the variance, and thereby assess whether the performance error bounds of two conditions are distinct.

5 Results

Given that for this data, the rate of segment boundary instances (positive labels) is about 10%, a baseline classifier that always predicts a non-segment will have about 90% overall accuracy. The baseline column in Table 2 shows the average accuracy that would be achieved by this simple baseline on the test data for a given run, along with the bootstrapped confidence interval for this baseline over the 50 runs. In the 84 experiments, the baseline ranged from 90% (+/- 1%) to 89% (+/-

	Interaction feature	
1	Speaker	whether there is a speaker switch between preceding utterance and current utterance
	Acoustic features	
2	Pitch_MIN	Minimum pitch of the utterance
3	Pitch_MAX	Maximum pitch of the utterance
4	Pitch_MEAN	Mean pitch of the utterance
5	Pitch_STDV	Standard deviation of the pitch of the utterance
6	Pitch_RANGE	Maximim pitch of the utterance less the minimum pitch
7	Pitch_CHANGE	Pitch_MEAN of the current utterance less the Pitch_MEAN of the preceding utterance
8	Intensity_MIN	Minimum intensity of the utterance
9	Intensity_MAX	Maximum intensity of the utterance
10	Intensity_MEAN	Mean intensity of the utterance
11	Intensity_STDV	Standard deviation of the intensity of the utterance
12	Intensity_RANGE	Intensity_MAX less Intensity_MIN
13	Intensity_CHANGE	Intensity_MEAN of the current utterance less Intensity_MEAN of preceding utterance
14	LR1	Utterance duration
15	LR1_Normalized	Utterance duration normalized by each speaker independently
	Lexical features	
16	LR2_1	Word count
17	LR2_2	Word count normalized by speaker
18	LR3_1	Words per second
19	LR3_2	Words per second by speaker
20	LR4	Average word length
21	LR5	Maximum word length
22	LR6_1	Average frequency of characters in the utterance
23	LR6_2	Number of low frequency characters
24	IR	Number of content words
25	PN_1	Number of named entities
26	PN_2	Whether the utterance contains a new named entity
	Pause features	
27	Pause_DURT	total duration of all pauses
28	Pause_RATIO	proportion utterance consisting of pauses
29	FP1	Presence of a filled pause at the beginning of an utterance
30	FP2	Presence of a filled pause at the end of an utterance
31	FP3	Presence of a filled pause in the middle of an utterance
32	P1	Presence of a pause tag at the beginning of an utterance
33	P2	Presence of a pause tag at the end of an utterance
34	P3	Presence of a pause tag in the middle of an utterance

Table 1: Discourse Features

1%). Crucially, however, the simple baseline will fail to identify any of the members of the positive class. Though it is difficult to beat the baseline on overall accuracy, the question addressed here is what level of accuracy is achieved on the positive class, while remaining relatively consistent with the baseline on overall accuracy. It should be noted that accuracy on the positive class is the same as *recall*, or *sensitivity* (the term used in the epidemiological literature). The worst performing classifier among the 84 (disc/utterance/ S1P4) achieves 83% (+/- 1%) accuracy overall, or below the baseline by 6%, with 11% accuracy on the positive class, 100% of which is a gain over the baseline. By this standard, the best classifier of the 84 conditions (bow/dial/S4P1) matches the baseline on overall accuracy, and achieves 50% (+/- 5%) accuracy on the positive class, which far exceeds the baseline. About half of the experimental conditions meet the baseline and achieve at least 25% accuracy on the positive class.

Overall accuracy, and accuracy on the positive class, measure prediction error, but can be supplemented with additional metrics that facilitate analysis of the nature and cost of error types. As a supplementary metric, we report average F-measure, the harmonic mean of recall and precision, due to its familiarity, and because it provides a sense of how often a classifier incorrectly predicts the positive class. An F-measure close to accuracy on the positive class indicates that precision is about the same as recall, while a relatively higher F-measure indicates that the precision is even higher than the F-measure, and the converse is true when the F-measure is lower than accuracy on the positive class. Table 2 shows 32 classifiers with the highest measures of accuracy, accuracy on the positive class, and F-measure. The confidence intervals for accuracy on the positive class and F-measure are rather wide, compared to those for overall accu-

Exp.	$\overline{Baseline}$ (sd)	$\overline{Acc}$ (sd)	$\overline{AccPos(Recall)}$ (sd)	$\overline{F}$ (sd)	$>_{Acc_{pos}}$	$> F$
bow/dial/S4P1	0.89 (+/-0.010)	0.89 (+/-0.009)	0.42 (+/-0.082)	0.28 (+/-0.054)	22	11
bow/dial/S4P2	0.90 (+/-0.013)	0.89 (+/-0.010)	0.39 (+/-0.071)	0.26 (+/-0.064)	22	3
bow/utterance/S1P0	0.90 (+/-0.004)	0.90 (+/-0.005)	0.51 (+/-0.051)	0.26 (+/-0.034)	30	11
bow/utterance/S4P0	0.89 (+/-0.005)	0.88 (+/-0.006)	0.43 (+/-0.049)	0.26 (+/-0.040)	23	10
disc/dial/S2P1	0.90 (+/-0.009)	0.87 (+/-0.009)	0.32 (+/-0.059)	0.26 (+/-0.037)	4	10
bow/utterance/S4P3	0.89 (+/-0.006)	0.88 (+/-0.005)	0.41 (+/-0.050)	0.25 (+/-0.027)	22	11
combo/dial/S3P2	0.89 (+/-0.011)	0.86 (+/-0.010)	0.31 (+/-0.048)	0.25 (+/-0.031)	7	10
disc/dial/S4P3	0.90 (+/-0.008)	0.86 (+/-0.009)	0.30 (+/-0.041)	0.25 (+/-0.030)	4	10
combo/dial/S3P1	0.89 (+/-0.010)	0.86 (+/-0.011)	0.31 (+/-0.059)	0.25 (+/-0.038)	3	10
combo/dial/S4P2	0.89 (+/-0.013)	0.86 (+/-0.012)	0.30 (+/-0.044)	0.25 (+/-0.031)	4	10
combo/dial/S2P1	0.89 (+/-0.012)	0.87 (+/-0.010)	0.32 (+/-0.054)	0.25 (+/-0.033)	7	10
combo/dial/S4P3	0.90 (+/-0.007)	0.87 (+/-0.008)	0.29 (+/-0.044)	0.25 (+/-0.035)	4	10
disc/dial/S3P2	0.90 (+/-0.008)	0.87 (+/-0.008)	0.29 (+/-0.047)	0.25 (+/-0.040)	3	10
bow/utterance/S4P1	0.90 (+/-0.005)	0.89 (+/-0.004)	0.40 (+/-0.053)	0.25 (+/-0.020)	22	10
bow/dial/S4P3	0.90 (+/-0.007)	0.89 (+/-0.009)	0.39 (+/-0.072)	0.25 (+/-0.035)	22	10
disc/dial/S4P2	0.90 (+/-0.009)	0.86 (+/-0.009)	0.28 (+/-0.042)	0.25 (+/-0.030)	0	10
bow/dial/S1P0	0.90 (+/-0.009)	0.89 (+/-0.009)	0.48 (+/-0.065)	0.24 (+/-0.045)	28	0
combo/dial/S4P1	0.90 (+/-0.010)	0.86 (+/-0.010)	0.28 (+/-0.045)	0.24 (+/-0.034)	0	9
disc/dial/S3P1	0.89 (+/-0.011)	0.86 (+/-0.010)	0.29 (+/-0.046)	0.24 (+/-0.033)	2	9
bow/dial/S4P0	0.90 (+/-0.009)	0.88 (+/-0.011)	0.37 (+/-0.031)	0.24 (+/-0.040)	22	0
disc/dial/S4P1	0.90 (+/-0.009)	0.86 (+/-0.008)	0.27 (+/-0.041)	0.23 (+/-0.032)	0	3
bow/utterance/S4P2	0.89 (+/-0.007)	0.88 (+/-0.010)	0.39 (+/-0.044)	0.23 (+/-0.033)	22	0
combo/utterance/S2P0	0.89 (+/-0.005)	0.86 (+/-0.009)	0.27 (+/-0.041)	0.21 (+/-0.029)	0	0
disc/dial/S2P0	0.89 (+/-0.010)	0.86 (+/-0.009)	0.27 (+/-0.047)	0.20 (+/-0.027)	0	0
disc/utterance/S2P0	0.90 (+/-0.006)	0.86 (+/-0.008)	0.26 (+/-0.032)	0.20 (+/-0.024)	0	0
combo/utterance/S1P0	0.89 (+/-0.005)	0.88 (+/-0.006)	0.31 (+/-0.041)	0.20 (+/-0.026)	10	0
combo/utterance/S3P0	0.90 (+/-0.005)	0.86 (+/-0.008)	0.25 (+/-0.038)	0.20 (+/-0.033)	0	0
disc/utterance/S4P3	0.89 (+/-0.006)	0.86 (+/-0.009)	0.24 (+/-0.043)	0.20 (+/-0.033)	0	0
combo/utterance/S2P1	0.89 (+/-0.006)	0.86 (+/-0.008)	0.26 (+/-0.036)	0.20 (+/-0.023)	0	0
disc/utterance/S2P1	0.89 (+/-0.005)	0.86 (+/-0.007)	0.26 (+/-0.032)	0.20 (+/-0.022)	0	0
combo/utterance/S4P1	0.89 (+/-0.006)	0.85 (+/-0.008)	0.24 (+/-0.033)	0.20 (+/-0.027)	0	0
disc/utterance/S4P0	0.89 (+/-0.006)	0.85 (+/-0.009)	0.24 (+/-0.034)	0.20 (+/-0.024)	0	0

Table 2: Classification performance (with standard deviations in parentheses) of the best 40% of 84 J48 models trained on 75% of the data and tested on the remaining 25%, with bootstrapped confidence intervals from 50 trials each.

racy. To draw comparisons among the classifiers that take into account this variance, the two rightmost columns of the table indicate for each classifier how many other classifiers in the same table the current classifier surpasses on mean accuracy of the positive class, or on mean F-measure. Here, to surpass another classifier means the lower bound of its confidence interval surpasses the upper bounds of other classifiers' confidence intervals.

Table 2 shows that there is no one classifier that surpasses all others on all measures. There are, however, some clear trends. Regarding the number of utterances spanned by each data instance, the table shows that of the 32 best performing classifiers, the majority (seventeen) have size 4 spans, and all but three have spans longer than a single utterance. This trend indicates that more context leads to better accuracy overall and better accuracy on the positive class. Regarding where the segment boundary is located relative to the span, the majority of cases (twenty-two) locate the boundary within the span, meaning that the span includes one or more of the final utterances of a segment and one or more of the initial utterances of the next segment. The remaining cases involve spans that include utterances only from the beginning of the segment. There are no cases of higher performing classifiers that use spans from segment endings. Among the classifiers in the top half of the table, the best performing bow classifiers surpass a larger number of the other classifiers on accuracy of the positive class. The best performing discourse or combination classifiers surpass a larger number of other classifiers on F-measure. This suggests that in general, the bow classifiers do better on recall and the classifiers with discourse features have higher precision.

The combination of BOW and discourse features has a performance that differs little from the discourse features alone, and does not do as well as BOW S4P1. This result was unexpected, and suggests that the bow and discourse feature sets often identify nearly the same set of discourse

Discourse, Rand Dial, S4P3		
Activity Type	TP %	FN %
Inform	7 (0.11)	56 (0.89)
Book Request	**18 (0.32)**	**40 (0.68)**
Librarian Proposal	4 (0.27)	11 (0.73)
Request-Action	0 (0.00)	6 (1.00)
Information-Request	6 (0.11)	47 (0.89)
Sidebar	1 (0.08)	11 (0.92)
Conventional	5 (0.17)	25 (0.83)
Total	37 (0.14)	230 (0.86)
BOW, Rand Dial, S4P2		
Inform	7 (0.10)	70 (0.90)
Book Request	14 (0.20)	57 (0.80)
Librarian Proposal	1 (0.05)	20 (0.95)
Request-Action	0 (0.00)	5 (1.00)
Information-Request	8 (0.16)	42 (0.84)
Sidebar	0 (0.00)	13 (1.00)
Conventional	6 (0.23)	29 (0.77)
Total	37 (0.14)	230 (0.86)

Table 3: Error Analysis of the Positive Class

boundaries. Since the initial utterances of a segment seem to have features with greater predictive power than the final utterances of a segment, and since discourse cue words tend to occur in the first utterance or so of a segment, it could be that discourse cue words explain the good performance of both sets of features. This could be tested in future work by restricting a BOW representation to words other than discourse cue words.

To pursue in more detail the factors that influence accuracy on the positive class (recall), we now turn to an error analysis of the kinds of discourse units associated with true positives versus false negatives of the classifier's confusion matrix. Table 3 presents the results of an error analysis of the two cells of the confusion matrix for a classifier's results on the positive class, the true positives and the false negatives. We looked at the breakdown of the seven kinds of discourse units to see whether there were differences in the likelihood of a correct identification of a boundary, depending on the kind of discourse unit in question. Results are drawn from classifiers learned under two conditions, S4P3 spans with discourse features randomized by dialogue (disc/dial/S4P3) and S4P3 spans with BOW features, randomized by dialogue (bow/dial/S4P3). (Results from other classifiers are very similar.) In both cases, Book-Requests have a much higher probability of being among the true positives (32% for discourse, 20% for BOW) than for the positive class overall (14%). Conventional discourse units, where the participants first make their greetings, or make their final good byes, are also correctly identified more often than the overall TP rate. Librarian Proposals are identified well by the model using the discourse features, but not by the one using the BOW features. We speculate that this is because Librarian Proposals typically present information that is new to the discourse: often, the librarian is making a suggestion to the patron based on information the librarian can see in the preference field of the patron's record, or in the patron's past borrowing behavior. We speculate that the vocabulary in Librarian Proposals may be too variable to be predictive. Information-Request units and Inform units are also relatively difficult to identify correctly.

6 Conclusion

The problem of identification of conversational activities is a difficult one for machine processing for many reasons. Like vision and speech, segmentation of the units is difficult because the units are not discrete, objective, components of perception, but instead are the result of abstraction. The experiments presented here consider a novel explanation for the difficulty of the task, which is that discourse units differ from each other regarding the manner in which they evolve in time. The results show that a data representation that includes utterances from both the end of one unit and the beginning of another improves performance. The transition between one conversational activity and another takes place over the course of several utterances, rather than occurring at an instant in time. Error analysis indicates further that discourse units that correspond to conversational activities with clear end points that can be achieved have a higher probability of being recognized correctly.

References

John L. Austin. 1962. *How to do Things with Words: The William James Lectures delivered at Harvard University in 1955*. Clarendon Press, Oxford.

Satanjeev Banerjee and Alexander I. Rudnicky. 2006. A texttiling based approach to topic boundary detection in meetings. Technical report, Department of Computer Science, Carnegie Mellon University.

David R. Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and Philosophy*, 9(1):37–61.

Bradley Efron and Robert Tibshirani. 1997. Improvements on cross-validation: The .632+ boot-

strap method. *Journal of the American Statistical Association*, 92(438):548–560, June.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pages 334–343. Association for Computational Linguistics.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 562–569, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Marti A. Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1).

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19:501–530.

Jun Hu, Rebecca J. Passonneau, and Owen Rambow. 2009. Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units. In *Proceedings of the SIGDIAL 2009 Conference*, pages 357–366, London, UK, September. Association for Computational Linguistics.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.

Andrew Y. Ng. 1997. Preventing overfitting of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 245–253.

Nicolas Obin, Volker Dellwo, Anne Lacheret, and Xavier Rodet. 2010. Expectations for discourse genre identification: aprosodic study. In *Interspeech*, pages 3070–3073.

A.S. Park and J.R. Glass. 2008. Unsupervised pattern discovery in speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):186–197, Jan.

Rebecca J. Passonneau and Diane J. Litman. 1997. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, March.

Rebecca J. Passonneau, Irene Alvarado, Phil Crone, and Simon Jerome. 2011. Paradise-style evaluation of a human-human library corpus. In *Proceedings of the SIGDIAL 2011 Conference*, pages 325–331, Portland, Oregon, June. Association for Computational Linguistics.

Matthew Purver, Thomas L. Griffiths, and Joshua B. Kording, Konrad P. andTenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics (ACL-44)*, pages 17–24. Association for Computational Linguistics.

R. Bharat Rao and Glenn Fung. 2008. On the dangers of cross-validation. an experimental evaluation. In *SDM*, pages 588–596. SIAM.

Klaus Ries, Lori Levin, Liza Valle, Alon Lavie, and Alex Waibel. 2000. Shallow discourse genre annotation in callhome spanish. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*. European Language Resources and Evaluation (ELRA).

Klaus Ries. 2002. Segmenting conversations by topic, initiative, and style. In AnniR. Coden, EricW. Brown, and Savitha Srinivasan, editors, *Information Retrieval Techniques for Speech Applications*, volume 2273 of *Lecture Notes in Computer Science*, pages 51–66. Springer Berlin Heidelberg.

J.D. Rodriguez, A. Perez, and J.A. Lozano. 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):569–575, March.

Zeno Vendler. 1957. Verbs and times. *Philosophical Review*, 66(2):143–160.

Nigel G. Ward and Karen A. Richart-Ruiz. 2013. Patterns of importance variation in spoken dialog. In *SigDial*.

Nigel G. Ward and Alejandro Vega. 2012. A bottom-up exploration of the dimensions of dialog state in spoken interaction. In *SigDial*.

Nigel G. Ward and Steven D. Werner. 2013. Using dialog-activity similarity for spoken information retrieval. In *Interspeech*.