Identifying Discourse Markers in Spoken Dialog

Peter A. Heeman[†] and **Donna Byron**‡ and **James F. Allen**‡

[†]Computer Science and Engineering
Oregon Graduate Institute
PO Box 91000 Portland OR 97291
heeman@cse.ogi.edu

‡Department of Computer Science
University of Rochester
Rochester NY 14627
{dbyron,james}@cs.rochester.edu

Abstract

In this paper, we present a method for identifying discourse marker usage in spontaneous speech based on machine learning. Discourse markers are denoted by special POS tags, and thus the process of POS tagging can be used to identify discourse markers. By incorporating POS tagging into language modeling, discourse markers can be identified during speech recognition, in which the timeliness of the information can be used to help predict the following words. We contrast this approach with an alternative machine learning approach proposed by Litman (1996). This paper also argues that discourse markers can be used to help the hearer predict the role that the upcoming utterance plays in the dialog. Thus discourse markers should provide valuable evidence for automatic dialog act prediction.

Introduction

Discourse markers are a linguistic devise that speakers use to signal how the upcoming unit of speech or text relates to the current discourse state (Schiffrin 1987). Previous work in computational linguistics has emphasized their role in marking changes in the global discourse structure (e.g. (Grosz & Sidner 1986; Reichman 1985; Cohen 1984)). For instance, "by the way" is used to mark the start of a digression, "anyway" to mark the return from one, and "now" to shift to a new topic. Schiffrin's work in social dialogue (1987) took a much wider scope, and examined how discourse markers in general are used. She found that they are used to mark the information status in an utterance and how it relates to the previous discourse state. For instance, when someone is about to disagree with information in the discourse state, they might introduce the utterance with "well".

In human-human task-oriented dialogs, discourse markers abound. In the Trains corpus of spontaneous speech (Heeman & Allen 1995), 44.1% of the turns (other than acknowledgments) are introduced with a discourse marker. Because discourse markers are so prominent in task-oriented dialogs, they could be a valuable source of information for understanding the utterances that they introduce. This striking feature of task-oriented dialog has been largely ignored by other researchers in building spoken dialog systems, which simply regard them as noise (cf. (Dahlbäck & Jönsson 1992)). Task-oriented dialogs manifest a considerably different surface form than either monologues, social dialog or written text (Brown & Yule

1983), so it is not clear whether discourse markers are playing the same role in task-oriented dialogs as in other forms of discourse.

One problem with discourse markers, however, is that there is ambiguity as to whether lexical items are functioning as discourse markers. Consider the lexical item "so". Not only can it be used as a discourse marker to introduce an utterance, but it can also be used sententially to indicate a subordinating clause as illustrated by the following example from the Trains corpus.

Example 1 (d93-15.2 utt9)
it takes an hour to load them
just so you know

Discourse markers can also be used inside an utterance to mark a *speech repair*, where the speaker goes back and repeats or corrects something she just said. Here, the discourse markers play a much more internal role, as the following example with "well" illustrates.

Example 2 (d93-26.3 utt12)

can I have engine $\underbrace{\text{well}}_{et}$ if I take engine one and pick up a boxcar
$\underbrace{\text{can I have engine}}_{reparandum}$ $\underset{ip}{\uparrow}$

Due to these difficulties, an effective algorithm for identifying discourse markers in spontaneous speech needs to also address the problem of segmenting speech into utterance units and identifying speech repairs (Heeman & Allen 1997b).

In the rest of this paper, we first review the Trains corpus and the manner in which the discourse markers were annotated by using special part-of-speech (POS) tags to denote them. We then examine the role that discourse markers play in task-oriented dialogs. We then present our speech recognition language model, which incorporates POS tagging, and thus discourse marker identification. We show that distinguishing discourse marker usages results in improved language modeling. We also show that discourse marker identification is improved by modeling interactions with utterance segmentation and resolving speech repairs. From this, we conclude that discourse markers can be used by hearers to set up expectations of the role that the upcoming utterance plays in the dialog. Due to the ability to automatically identify discourse markers during the speech recognition process, we argue that they can be exploited in the task of dialog act identification,

which is currently receiving much attention in spontaneous speech research (e.g. (Taylor *et al.* 1997; Chu-Carroll 1998; Stolcke *et al.* 1998)). We conclude with a comparison to the method proposed by Litman (1996) for identifying discourse markers.

Trains Corpus

As part of the Trains project (Allen *et al.* 1995), which is a long term research project to build a conversationally proficient planning assistant, we have collected a corpus of problem solving dialogs (Heeman & Allen 1995). The dialogs involve two human participants, one who is playing the role of a user and has a certain task to accomplish, and another who is playing the role of the system by acting as a planning assistant. The collection methodology was designed to make the setting as close to human-computer interaction as possible, but was not a *wizard* scenario, where one person pretends to be a computer; rather, the user knows that he is talking to another person. The Trains corpus consists of approximately six and a half hours of speech. Table 1 gives some general statistics about the corpus, including the number of dialogs, speakers, words, speaker turns, and occurrences of discourse markers.

Dialogs	98
Speakers	34
Words	58298
Turns	6163
Discourse Markers	8278

Table 1: Size of the Trains Corpus

Our strategy for annotating discourse markers is to mark such usages with special POS tags. Four special POS tags were added to the Penn Treebank tagset (Marcus, Santorini, & Marcinkiewicz 1993) to denote discourse marker usage. These tags are defined in Table 2.[1] Verbs used as discourse

AC: Single word acknowledgments, such as "okay", "right", "mm-hm", "yeah", "yes", "alright", "no", and "yep".

UH_D: Interjections with discourse purpose, such as "oh"", "well", "hm", "mm", and "like".

CC_D: Co-ordinating conjuncts used as discourse markers, such as "and", "so", "but", "oh", and "because".

RB_D: Adverbials used as discourse markers, such as "then", "now", "actually", "first", and "anyway".

Table 2: POS tags for Discourse Markers

markers, such as "wait", and "see", are not given special markers, but are annotated as verbs. Also, no attempt has

[1] Other additions to the tagset are described in Heeman (1997).

been made at analyzing multi-word discourse markers, such as "by the way" and "you know". However, phrases such as "oh really" and "and then" are treated as two individual discourse markers. Lastly, filled pause words, namely "uh", "um" and "er", are marked with **UH_FP**; but these are not considered as discourse markers.

POS-Based Language Model

The traditional goal of speech recognition is to find the sequence of words $\hat{W}$ that is maximal given the acoustic signal A. In earlier work (Heeman & Allen 1997a; Heeman 1997), we argue that this view is too limiting. In a spoken dialog system, word recognition is just the first step in understanding the speaker's turn. Furthermore, speech recognition is difficult especially without the use of higher level information. Hence, we propose as a first step to incorporate POS tagging into the speech recognition process.

Previous approaches that have made use of POS tags in speech recognition view the POS tags as intermediate objects by summing over the POS tag sequences (Jelinek 1985). Instead, we take the approach of redefining the goal of the speech recognition process so that it finds the best word (W) and POS tag (P) sequence given the acoustic signal. The derivation of the acoustic model and language model is now as follows.

$$
\begin{aligned}
\hat{W}\hat{P} &= \arg\max_{W,P} \Pr(WP|A) \\
&= \arg\max_{WP} \frac{\Pr(A|WP)\Pr(WP)}{\Pr(A)} \\
&= \arg\max_{WP} \Pr(A|WP)\Pr(WP)
\end{aligned}
$$

The first term $\Pr(A|WP)$ is the factor due to the acoustic model, which we can approximate by $\Pr(A|W)$. The second term $\Pr(WP)$ is the factor due to the language model. We rewrite $\Pr(WP)$ as $\Pr(W_{1,N}P_{1,N})$, where N is the number of words in the sequence. We now rewrite the language model probability as follows.

$$
\begin{aligned}
&\Pr(W_{1,N}P_{1,N}) \\
&= \prod_{i=1,N} \Pr(W_iP_i|W_{1,i\text{-}1}P_{1,i\text{-}1}) \\
&= \prod_{i=1,N} \Pr(W_i|W_{1,i\text{-}1}P_{1,i})\Pr(P_i|W_{1,i\text{-}1}P_{1,i\text{-}1})
\end{aligned}
$$

The final probability distributions are similar to those used by previous attempts to use POS tags in language modeling (Jelinek 1985) and those used for POS tagging of written text (Charniak *et al.* 1993; Church 1988; DeRose 1988). However, these approaches simplify the probability distributions as shown by the approximations below.

$$
\begin{aligned}
\Pr(W_i|W_{1,i\text{-}1}P_{1,i}) &\approx \Pr(W_i|P_i) \\
\Pr(P_i|W_{1,i\text{-}1}P_{1,i\text{-}1}) &\approx \Pr(P_i|P_{1,i\text{-}1})
\end{aligned}
$$

However, as we have shown in earlier work (Heeman & Allen 1997a; Heeman 1997), such simplifications lead to poor language models.

Probability Distributions

We have two probability distributions that need to be estimated. The simplest approach for estimating the probability of an event given a context is to use the relative frequency that the event occurs given the context according to a training corpus. However, no matter how large the training corpus is, there will always be event-context pairs that have not been seen or that have been seen too rarely to accurately estimate the probability. To alleviate this problem, one can partition the contexts into a smaller number of equivalence classes and use these equivalence classes to compute the relative frequencies.

We use a decision tree learning algorithm (Bahl *et al.* 1989; Black *et al.* 1992; Breiman *et al.* 1984), which uses information theoretic measures to construct equivalence classes of the context in order to cope with sparseness of data. The decision tree algorithm starts with all of the training data in a single leaf node. For each leaf node, it looks for the question to ask of the context such that splitting the node into two leaf nodes results in the biggest decrease in *impurity*, where the impurity measures how well each leaf predicts the events in the node. Heldout data is used to decide when to stop growing the tree: a split is rejected if the split does not result in a decrease in impurity with respect to the heldout data. After the tree is grown, the heldout dataset is used to smooth the probabilities of each node with its parent (Bahl *et al.* 1989).

To allow the decision tree to ask questions about the words and POS tags in the context such that the questions can generalize about words and POS tags that behave similarly, we cluster the words and POS tags using the algorithm of Brown *et al.* (1992) into a binary classification tree. The algorithm starts with each word (or POS tag) in a separate class, and successively merges classes that result in the smallest lost in mutual information in terms of the co-occurrences of these classes. By keeping track of the order that classes were merged, we can construct a hierarchical classification of the classes. Figure 1 shows a POS classification tree, which was automatically built from the training data. Note that the classification algorithm has clustered the discourse marker POS tags close to each other in the classification tree.

The binary classification tree gives an implicit binary encoding for each POS tag, which is determined by the sequence of top and bottom edges that leads from the root node to the node for the POS tag. The binary encoding allows the decision tree to ask about the words and POS tags using simple binary questions, such as 'is the third bit of the POS tag encoding equal to one?' the POS tag.
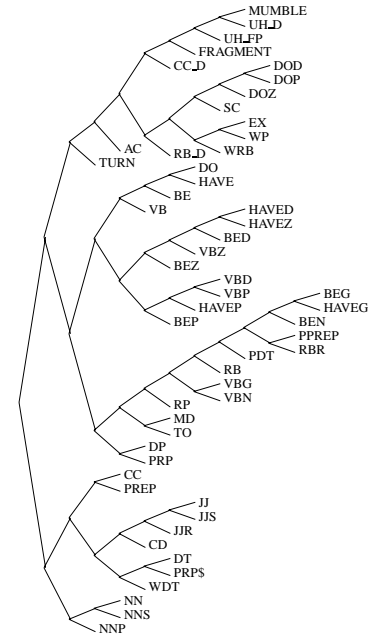

Figure 1: POS Classification Tree

Unlike other work (e.g. (Black *et al.* 1992; Magerman 1995)), we treat the word identities as a further refinement of the POS tags; thus we build a word classification tree for each POS tag. We grow the classification tree by starting with a unique class for each word and each POS tag that it takes on. When we merge classes to form the hierarchy, we only allow merges if all of the words in both classes have the same POS tag. The result is a word classification tree for each POS tag. This approach of building a word classification tree for each POS tag has the advantage that it better deals with words that can take on multiple senses, such as the word "loads", which can be a plural noun (**NNS**) or a present tense third-person verb (**VBZ**). As well, it constrains the task of building the word classification trees since the major distinctions are captured by the POS classification tree, thus allowing us to build classification trees even for small corpora. Figure 2 gives the classification tree for the acknowledgments (**AC**). For each word, we give the number of times that it occurred in the training data. Words that only occurred once in the training corpus have been grouped together in the class '**!unknown**'. Although the clustering algorithm was able to group some of the similar acknowledgments with each other, such as the group of "mm-hm" and "uh-huh", the group of "good", "great", and "fine", other similar words were not grouped together, such as "yep" with "yes" and "yeah", and "no" with "nope". Word adjacency information is insufficient for capturing such semantic information.

```
                        !unknown 1
                      fine 4
                  exactly 6
                  good 13
                  great 14
              sorry 14
          alright 155
      okay 1700
          hello 71
          hi 8
                  yeah 185
                  yes 194
          no 128
                  nope 5
                  sure 9
              correct 13
          yep 108
          mm-hm 246
          uh-huh 30
      right 434
```
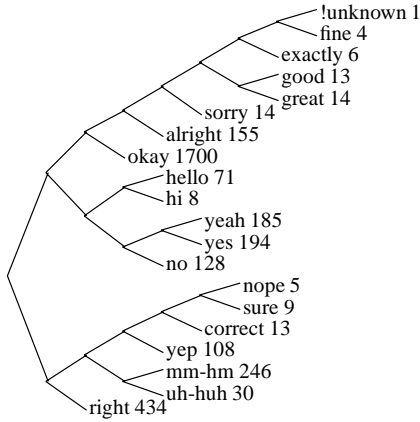
Figure 2: AC Classification Tree

Results

To demonstrate our model, we use a 6-fold cross validation procedure, in which we use each sixth of the corpus for testing data, and the rest for training data. We start with the word transcriptions of the Trains corpus, thus allowing us to get a clearer indication of the performance of our model without having to take into account the poor performance of speech recognizers on spontaneous speech.

Table 3 reports the results of explicitly modeling discourse markers with special POS tags. The second column, "No DM", reports the results of collapsing the discourse marker usages with the sentential usages. Thus, the discourse conjunct **CC_D** is collapsed into **CC**, the discourse adverbial **RB_D** is collapsed into **RB**, and the acknowledgment **AC** and discourse interjection **UH_D** are collapsed into **UH_FP**. The third column gives the results of the model that does distinguish discourse marker usages, but ignoring POS errors due to miscategorizing words as being discourse markers or not. We see that modeling discourse markers results in a reduction of POS errors from 1219 to 1189, giving a POS error rate of 2.04%. We also see a small decrease in perplexity from 24.20 to 24.04. Perplexity of a test set of N words $w_{1,N}$ is calculated as follows.

$$2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 \Pr(w_i|w_{1,i-1})}$$

In previous work (Heeman & Allen 1997b; Heeman 1997), we argued that discourse marker identification is

	No DM	DM
POS Errors	1219	1189
POS Error Rate	2.09	2.04
Perplexity	24.20	24.04

Table 3: Discourse Markers and Perplexity

	Base Model	Tones Repairs Corrections	Tones Repairs Corrections Silences
POS Tagging			
Errors	1711	1652	1572
Error Rate	2.93	2.83	2.69
Perplexity	24.04	22.96	22.35
Discourse Markers			
Errors	630	611	533
Recall	96.75	96.67	97.26
Precision	95.68	95.97	96.32

Table 4: POS Tagging and Perplexity Results

tightly intertwined with the problems of intonational phrase identification and resolving speech repairs. These three tasks, we claim, are necessary in order to understand the user's contributions. In Table 4, we show how discourse marker identification, POS tagging and perplexity benefit by modeling the speaker's utterance. The second column gives the results of the POS-based model, which was used in the third column of Table 3, the third column gives the results of incorporating the detection and correction of speech repairs and detection of intonational phrase boundary tones, and the fourth column gives the results of adding in silence information to give further evidence as to whether a speech repair or boundary tone occurred. As can be seen, modeling the user's utterances improves POS tagging and word perplexity; adding in silence information to help detect speech repairs and intonational boundaries further improves these two rates.[2] Of concern to this paper, we also see an improvement in the identification of discourse markers, improving from 630 to 533 errors. This gives a final recall rate of 97.26% and a precision of 96.32.[3] In Heeman (1997), we also show that modeling discourse markers improves the detection of speech repairs and intonational boundaries.

Comparison to Other Work

Hirschberg and Litman (1993) examined how intonational information can distinguish between the discourse and sentential interpretation for a set of ambiguous lexical items. This work was based on hand-transcribed intonational features and examined discourse markers that were one word long. In an initial study of the discourse marker "now", they found that discourse usages of the word "now" were either an intermediate phrase by themselves (or in a phrase consisting entirely of ambiguous tokens), or they are first in

---

[2]Note the POS results include errors due to miscategorizing discourse markers, which were excluded from the POS results reported in Table 3.

[3]The recall rate is the number of discourse markers that were correctly identified over the actual number of discourse markers. The precision rate is the number of correctly identified discourse markers over the total number of discourse markers guessed.

an intermediate phrase (or preceded by other ambiguous tokens) and are either de-accented or have a low accent ($\mathbf{L^*}$). Sentential uses were either non-initial in a phrase or, if first, bore a high ($\mathbf{H^*}$) or complex accent (i.e. not a $\mathbf{L^*}$ accent). In a second study, Hirschberg and Litman used a speech consisting of approximately 12,500 words. They found that the intonational model that they had proposed for the discourse marker "now" achieved a recall rate of 63.1% of the discourse markers with a precision of 88.3%.[4]

Hirschberg and Litman also looked at the effect of orthographic markers and POS tags. For the orthographic markings, they looked at how well discourse markers can be predicted based on whether they follow or precede a hand-annotated punctuation mark. They also examined correlations with POS tags. For this experiment, rather than define special POS tags as we have done, they choose discourse marker interpretation versus sentential interpretation based on whichever is more likely for that POS tag, where the POS tags were automatically computed using Church's part-of-speech tagger (1988). This gives them a recall rate of 39.0% and a precision of 55.2%.

Litman (1996) explored using machine learning techniques to automatically learn classification rules for discourse markers. She contrasted the performance of CGRENDEL (Cohen 1992; Cohen 1993) with C4.5 (Quinlan 1993). CGRENDEL is a learning algorithm that learns an ordered set of if-then rules that map a condition to its most-likely event (in this case discourse or sentential interpretation of potential discourse marker). C4.5 is a decision tree growing algorithm that learns a hierarchical set of if-then rules in which the leaf nodes specify the mapping to the most-likely event. She found that machine learning techniques could be used to learn a classification algorithm that was as good as the algorithm manually built by Hirschberg and Litman (1993). Further improvements were obtained when different sets of features about the context were explored, such as the identity of the token under consideration. The best results (although the differences between this version and some of the others might not be significant) were obtained by using CGRENDEL and letting it choose conditions from the following set: length of intonational phrase, position of token in intonational phrase, length of intermediate phrase, position of token in intermediate phrase, composition of intermediate phrase (token is alone in intermediate phrase, phrase consists entirely of potential discourse markers, or otherwise), and identity of potential discourse marker. The automatically derived classification algorithm achieved a success rate of 85.5%, which translates into a discourse marker error rate of 37.3%, in comparison to the error rate of 45.3% for the algorithm of Hirschberg and Litman (1993). Hence, machine learning

---

[4] See Heeman (1997) for a derivation of the recall and precision rates.

techniques are an effective way in which a number of different sources of information can be combined to identify discourse markers.

Direct comparisons with our results are problematic since our corpus is approximately five times as large. Also we use task-oriented human-human dialogs, rather than a monologue, and hence our corpus includes a lot of turn-initial discourse markers for co-ordinating mutual belief. However, our results are based on automatically identifying intonational boundaries, rather than including these as part of the input. In any event, the work of Litman and the earlier work with Hirschberg indicate that our results can be further improved by also modeling intermediate phrase boundaries (phrase accents), and word accents, and by improving our modeling of these events, perhaps by using more acoustic cues. Conversely, we feel that our approach, which integrates discourse marker identification with speech recognition along with POS tagging, boundary tone identification and the resolution of speech repairs, allows different interpretations to be explored in parallel, rather than forcing individual decisions to be made about each ambiguous token. This allows interactions between these problems to be modeled, which we feel accounts for some of the improvement between our results and the results reported by Litman.

Predicting Speech Acts

Discourse markers are a prominent feature of human-human task-oriented dialogs. In this section, we examine the role that discourse markers, other than acknowledgments, play at the beginning of speaker turns and show that discourse markers can be used by the hearer to set up expectations of the role that the upcoming utterance plays in the dialog. Table 5 gives the number of occurrences of discourse markers in turn initial position in the Trains corpus. From column two, we see that discourse markers start 4202 of the 6163 utterances in the corpus, or 68.2%. If we exclude turn-initial filled pauses and acknowledgments and exclude turns that consist of only filled pauses and discourse markers, we see that 44.1% of the speaker turns are marked with a non-acknowledgment discourse marker.

In earlier work (Byron & Heeman 1997a; Byron & Hee-

Turns that start with	Number	Excluding initial AC's and UH_FP's
AC	3040	n.a.
CC_D	824	1414
RB_D	63	154
UH_D	275	302
UH_FP	462	n.a.
Other	1499	2373
Total	6163	4243

Table 5: Discourse markers in turn-initial position

Restate A restatement of either the plan or facts in the world that have been explicitly stated before.

Summarize Plan A restatement of the current working plan where this plan has been previously built up in pieces but has not been previously stated in its entirety.

Request for summary Typically questions about the total time the plan will take, such as "what's the total on that."

Conclude Explicit conclusion about the planning state that has not been stated previously, e.g. 'So that's not enough time' or 'So we have thirteen hours'

Elaborate Plan Adding new plan steps onto the plan, e.g. "How about if we bring engine two and two boxcars from Elmira to Corning"

Correction Correcting either the plan or a misconception of the other speaker.

Respond to new info Explicit acknowledgment of new information, such as "oh really" or "then let's do that".

Table 6: Conversational move categories

Acknowledge Backchannel 'Okay' or 'mm-hm'.

Check Restating old information to elicit a positive response from the partner (e.g. That was three hours to Bath?).

Confirm Restating old information, with no apparent intention of partner agreement.

Filled Pause A turn containing no information such as 'hm'.

Inform Information not previously made explicit.

Request Request for information.

Respond Respond to a Request.

Y/N Question Questions requiring a yes/no answer. Differ from Check because the speaker displays no bias toward which answer he expects.

Y/N Answer Answering 'yes', 'no', 'right', etc.

Table 8: Speech Act annotations

Conversational Move	Turns beginning with			
	And	**Oh**	**So**	**Well**
Restate	0	0	6	0
Summarize Plan	5	0	4	0
Request for summary	1	0	3	0
Conclude	0	0	15	0
Elaborate Plan	22	0	0	0
Correction	0	0	0	7
Respond to new info	0	17	0	0

Table 7: Correlations with conversational move

	Total Turns	Turn begins with				DM Turns % of Total
		And	**Oh**	**So**	**Well**	
Prior speech act initiates adjacency pair						
Check	23	0	0	0	1	4%
Request Info	45	0	0	1	0	2%
Y/N Question	8	0	0	0	0	0%
Prior speech act concludes adjacency pair						
Respond	38	3	2	5	1	30%
Y/N Answer	26	1	1	1	0	12%
Acknowledge	107	21	4	16	2	40%
Prior speech act not in adjacency pair						
Confirm	42	2	0	0	1	7%
Inform	96	1	10	5	2	19%
Filled Pause	6	0	0	0	0	0%

Table 9: Prior speech act of DM-initial turns

man 1997b), we investigated the role that discourse markers play in task-oriented human-human dialogs. We investigated Shriffin's claim that discourse markers can be used to express the relationship between the information in the upcoming utterance to the information in the discourse state (Schiffrin 1987). For each turn that began with a discourse marker, we coded the type of conversational move that the discourse marker introduced. The conversational move annotations, described in Table 6, attempt to capture speaker intent rather than the surface form of the utterance. We annotated five of the Trains dialogs, containing a total of 401 speaker turns and 24.5 minutes of speech.

In accordance with Schiffrin, we found that utterances that summarize information are likely to be introduced with "so", utterances that add on to the speakers prior contribution (and perhaps ignore the other conversants intervening contribution) are likely to be introduced with "and", and utterances that express dissent with the information in the discourse state are likely to be introduced with "well". Table 7 summarizes the co-occurrence of turn-initial discourse markers with the conversational moves that they introduce.

The table shows that different discourse markers strongly correlated with particular conversational moves. Because discourse markers are found in turn-initial position, they can be used as a timely indicator of the conversational move about to be made.

A more traditional method for analyzing the function of turns in a dialog is to focus on their surface form by categorizing them into speech acts, so we wanted to see if this sort of analysis would reveal anything interesting about discourse marker usage in the Trains dialogs. Table 8 defines the speech acts that were used to annotate the dialogs. We found that discourse markers on the whole do not correlate strongly with particular speech acts, as they did with conversational moves. This is corroborated by Schiffrin's (1987) corpus analysis, in which she concluded that turn-initiators reveal little about the construction of the upcoming turn. Although not correlating with syntactic construction, discourse markers do interact with the local discourse

structure property of adjacency pairs (Schegloff & Sacks 1973). In an adjacency pair, such as Question/Answer or Greeting/Greeting, the utterance of the first speech act of the pair sets up an obligation for the partner to produce the second speech act of the pair. After the first part of an adjacency pair has been produced, there is a very strong expectation about how the next turn will relate to the preceding discourse, e.g. it will provide an answer to the question just asked.

Since discourse markers help speakers signal how the current turn relates to prior talk, we decided to investigate what speech acts discourse markers tend to follow and how they correlate with adjacency pairs. Table 9 shows the *prior* speech act of turns beginning with discourse markers. The speech acts have been organized into those that form the first part of an adjacency pair (Request Info, Y/N Question, and Check), those that form second-pair-parts (Respond, Y/N/ Answer, and Acknowledge), and those that are not part of an adjacency pair sequence (Confirm, Inform, and Filled Pause). The table reveals the very low frequency of discourse marker initial turns after the initiation of an adjacency pair. After an adjacency pair has been initiated, the next turn almost never begins with a discourse marker, because the turn following the initiation of an adjacency pair is expected to be the completion of the pair. Since the role of that turn is not ambiguous, it does not need to begin with a discourse marker to mark its relationship to preceding talk. It would indeed be odd if after a direct question such as "so how many hours is it from Avon to Dansville" the system responded "and 6" or "so 6". A possible exception would be to begin with "well" if the upcoming utterance is a correction rather than an answer. There is one "so" turn in the annotated dialogs after a Request act, but it is a request for clarification of the question.

After a turn that is not the initiation of an adjacency pair, such as Acknowledge, Respond, or Inform, the next turn has a much higher probability of beginning with a discourse marker. Also when the prior speech act concludes an adjacency pair, the role of the next statement is ambiguous, so a discourse marker is used to mark its relationship to prior discourse.

In this section, we demonstrated that the choice of discourse marker gives evidence as to the type of conversational move that the speaker is about to make. Furthermore, discourse markers are more likely to be used where there are not strong expectations about the utterance that the speaker is about to make. Thus, discourse markers provide hearers with timely information as to how the upcoming speech should be interpreted.

Usefulness of Discourse Markers

We have also shown that discourse markers can be reliably identified in task-oriented spontaneous speech. The results given in the previous section show that knowledge of the discourse marker leads to strong expectations of the speech that will follow. However, none of the work in using machine learning techniques to predict the speech act of the users speech has used the presence of a discourse marker. Chu-Carroll (1998) examined syntactic type of the utterance and turn-taking information, but not the presence of a discourse marker. The work of Taylor et al. (1997) on using prosody to identify discourse act type also ignores the presence of discourse markers. Work of Stolcke et al. (1998) also ignores them. As Dahlbäck and Jönsson observed (1992), it might be that speakers drop the usage of discourse markers in talking with computer systems, but this might be more of an effect of the current abilities of such systems and user perceptions of them, rather than that people will not want to use these as their perception of computer dialogue systems increases. A first step in this direction is to make use of these markers in dialogue comprehension. Machine learning algorithms of discourse acts are ideally suited for this task.

Conclusion

In this paper, we have shown that discourse markers can be identified very reliably in spoken dialogue by viewing the identification task as part of the process of part-of-speech tagging and using a Markov model approach to identify them. The identification process can be incorporated into speech recognition, and this leads to a small reduction in both the word perplexity and POS tagging error rate. Incorporating other aspects of spontaneous speech, namely speech repair resolution and identification of intonation phrase boundary tones, leads to further improvements in our ability to identify discourse markers.

Our method for identifying discourse markers views this task as part of the speech recognition problem along with POS tagging. As such, rather than classifying each potential word independently as to whether it is a discourse marker or not (cf. (Litman 1996)), we find the best interpretation for the acoustic signal, which includes identifying the discourse markers. Using this approach means that the probability distributions that need to be estimated are more complicated than those traditionally used in speech recognition language modeling. Hence, we make use of a decision tree algorithm to partition the training data into equivalence classes from which the probability distributions can be computed.

Automatically identifying discourse markers early in the processing stream means that we can take advantage of their presence to help predict the following speech. In fact, we have shown that discourse markers not only can be used to help predict how the speaker's subsequent speech will build on to the discourse state, but also are often used when there are not already strong expectations, in terms of adjacency

pairs. However, most current spoken dialogue systems ignore their presence, even though they can be easily incorporated into existing machine learning algorithms that predict discourse act types.

Acknowledgments

References

[Allen *et al.* 1995] Allen, J. F.; Schubert, L.; Ferguson, G.; Heeman, P.; Hwang, C.; Kato, T.; Light, M.; Martin, N.; Miller, B.; Poesio, M.; and Traum, D. 1995. The Trains project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI* 7:7–48.

[Bahl *et al.* 1989] Bahl, L. R.; Brown, P. F.; deSouza, P. V.; and Mercer, R. L. 1989. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36(7):1001–1008.

[Black *et al.* 1992] Black, E.; Jelinek, F.; Lafferty, J.; Mercer, R.; and Roukos, S. 1992. Decision tree models applied to the labeling of text with parts-of-speech. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 117–121. Morgan Kaufman.

[Breiman *et al.* 1984] Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees*. Monterrey, CA: Wadsworth & Brooks.

[Brown & Yule 1983] Brown, G., and Yule, G. 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

[Brown *et al.* 1992] Brown, P. F.; Della Pietra, V. J.; deSouza, P. V.; Lai, J. C.; and Mercer, R. L. 1992. Class-based n-gram models of natural language. *Computational Linguistics* 18(4):467–479.

[Byron & Heeman 1997a] Byron, D. K., and Heeman, P. A. 1997a. Discourse marker use in task-oriented spoken dialog. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*.

[Byron & Heeman 1997b] Byron, D. K., and Heeman, P. A. 1997b. Discourse marker use in task-oriented spoken dialog. Technical report, Department of Computer Science, University of Rochester, 664.

[Charniak *et al.* 1993] Charniak, E.; Hendrickson, C.; Jacobson, N.; and Perkowitz, M. 1993. Equations for part-of-speech tagging. In *Proceedings of the National Conference on Artificial Intelligence (AAAI '93)*.

[Chu-Carroll 1998] Chu-Carroll, J. 1998. Statistical model for discourse act recognition in dialogue interactions. In *Proceedings of the AAAI Workshop on Applying Machine Learning to Discourse Processing*.

[Church 1988] Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, 136–143.

[Cohen 1984] Cohen, R. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING)*, 251–255.

[Cohen 1992] Cohen, W. W. 1992. Compiling knowledge into an explicit bias. In *Proceedings of the Ninth International Conference on Machine Learning*.

[Cohen 1993] Cohen, W. W. 1993. Efficient pruning methods for separate-and-conquer rule learning systems. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI '93)*.

[Dahlbäck & Jönsson 1992] Dahlbäck, N., and Jönsson, A. 1992. An empirically based computationally tractable dialogue model. In *Program of the 14th Annual Conference of the Cognitive Science Society*.

[DeRose 1988] DeRose, S. J. 1988. Grammatical category disambiguation by statistical optimization. *Computational Linguistics* 14(1):31–39.

[Grosz & Sidner 1986] Grosz, B. J., and Sidner, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175–204.

[Heeman & Allen 1995] Heeman, P. A., and Allen, J. F. 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.

[Heeman & Allen 1997a] Heeman, P. A., and Allen, J. F. 1997a. Incorporating POS tagging into language modeling. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)*.

[Heeman & Allen 1997b] Heeman, P. A., and Allen, J. F. 1997b. Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 254–261.

[Heeman 1997] Heeman, P. A. 1997. Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialog. Technical report, Department of Computer Science, University of Rochester. Doctoral dissertation.

[Hirschberg & Litman 1993] Hirschberg, J., and Litman, D. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3):501–530.

[Jelinek 1985] Jelinek, F. 1985. Self-organized language modeling for speech recognition. Technical report, IBM T.J. Watson Research Center, Continuous Speech Recognition Group, Yorktown Heights, NY.

[Litman 1996] Litman, D. J. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research* 5:53–94.

[Magerman 1995] Magerman, D. M. 1995. Statistical decision trees for parsing. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics*, 7–14.

[Marcus, Santorini, & Marcinkiewicz 1993] Marcus, M. P.; Santorini, B.; and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

[Quinlan 1993] Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufman.

[Reichman 1985] Reichman, R. 1985. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics (An ATN Model)*. Cambridge, MA: MIT Press.

[Schegloff & Sacks 1973] Schegloff, E. A., and Sacks, H. 1973. Opening up closings. *Semiotica* 7:289–327.

[Schiffrin 1987] Schiffrin, D. 1987. *Discourse Markers*. New York: Cambridge University Press.

[Stolcke *et al.* 1998] Stolcke, A.; Shriberg, E.; Bates, R.; Coccaro, N.; Jurafsky, D.; Martin, R.; Meteer, M.; Ries, K.; Taylor, P.; and Ess-Dykema, C. V. 1998. Dialog act modeling for conversational speech. In *Proceedings of the AAAI Workshop on Applying Machine Learning to Discourse Processing*.

[Taylor *et al.* 1997] Taylor, P.; King, S.; Isard, S.; Wright, H.; and Kowtko, J. 1997. Using intonation to constrain language models in speech recognition. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech).*