Training a Dialogue Act Tagger For Human-Human and Human-Computer Travel Dialogues

Rashmi Prasad and **Marilyn Walker**
AT&T Labs - Research
180 Park Avenue, Florham Park, NJ 07932, U.S.A.
rjprasad, walker@research.att.com

Abstract

While dialogue acts provide a useful schema for characterizing dialogue behaviors in human-computer and human-human dialogues, their utility is limited by the huge effort involved in hand-labelling dialogues with a dialogue act labelling scheme. In this work, we examine whether it is possible to fully automate the tagging task with the goal of enabling rapid creation of corpora for evaluating spoken dialogue systems and comparing them to human-human dialogues. We report results for training and testing an automatic classifier to label the information provider's utterances in spoken human-computer and human-human dialogues with DATE (Dialogue Act Tagging for Evaluation) dialogue act tags. We train and test the DATE tagger on various combinations of the DARPA Communicator June-2000 and October-2001 human-computer corpora, and the CMU human-human corpus in the travel planning domain. Our results show that we can achieve high accuracies on the human-computer data, and surprisingly, that the human-computer data improves accuracy on the human-human data, when only small amounts of human-human training data are available.

1 Introduction

Recent research on dialogue is based on the assumption that dialogue acts provide a useful way of characterizing dialogue behaviors in both human-human (HH) and human-computer (HC) dialogue (Isard and Carletta, 1995; Shriberg et al., 2000; Di Eugenio et al., 1998; Cattoni et al., 2001). Previous research has used dialogue act tagging for tasks such as improving recognition performance (Shriberg et al., 2000), identifying important parts of a dialogue (Finke et al., 1998), evaluating and comparing spoken dialogue systems (Walker et al., 2001c; Cattoni et al., 2001; Hastie et al., 2002), as a constraint on nominal expression generation (Jordan, 2000), and for comparing HH to HC dialogues (Doran et al., 2001).

Our work builds directly on the previous application of the DATE (Dialogue Act Tagging for Evaluation) tagging scheme to the evaluation and comparison of DARPA Communicator dialogues. The hypothesis underlying the use of dialogue act tagging in spoken dialogue evaluation is that a system's dialogue behaviors have a strong effect on its usability. Because Communicator systems have unique dialogue strategies, and a unique way of representing and achieving particular communicative goals, DATE was developed to consistently label dialogue behaviors across systems so that the potential utility of dialogue act tagging could be explored. In previous work, Walker and Passonneau defined the DATE scheme, and labelled the system utterances in the June 2000 data collection of 663 dialogues from nine participating Communicator systems (Walker et al., 2001c; Walker et al., 2001a). They then derived dialogue act metrics from the DATE tags and showed that when these metrics were used in the PARADISE evaluation framework (Walker et al., 1997) that they improved models of user satisfaction by an absolute 5%, and that the new metrics could be used to understand which system's dialogue strategies were most effective.

A major part of evaluation effort using dialogue act tagging, however, is to actually label the dialogues with the dialogue act tags. In previous work (Walker et al., 2001c), the DATE labelling of the June-2000 corpus was done using a semi-automatic method that involved collection of a large number of *utterance patterns* from the different sites participating in the collection and subsequent hand labelling of these patterns. The 100% coverage and accuracy achieved by the pattern matcher that was implemented for labelling the system utterances was crucially at the cost of maintaining a large labelled pattern database.[1] Furthermore, since the collected patterns were very specific and often exact duplicates of the system utterances in the dialogues, slight variations in the system utterances over time led to a reduction in coverage of the pattern matcher. For example, on the October-2001 collection, the tagger was able to label only 60% of the data. Accounting for the unmatched (and thus unlabeled) utterances turned out again to be a tedious process of manually augmenting the pattern database with new utterance patterns.

We examine whether it is possible to fully automate the tagging task with the goal being to enable *rapid* creation of new spoken dialogue corpora that can be submitted to evaluation procedures that rely on tagging schemes like DATE. Our hypothesis, arising from rough qualitative studies, is that within the same domain there are large similarities not only between different HC dialogues (across system as well as over a period of time) such as the Communicator dialogues, but also between HC dialogues and HH dialogues. Figure 1 shows a HC dialogue and Figure 2 shows a HH dialogue from the CMU-corpus (U=user, A=human agent, S=computer system) in which the information providers' utterances for information gathering and information presentation across both dialogues are very similar. Thus, based on this and other similar observations, we decided to test the extent to which sufficiently general tagging rules can be learned from HC data to label other HC data and HH data in similar domains with high accuracy.

We focus on labelling only the information providers' side of the dialogues. The primary motivation for this is to be able to characterize system dialogue strategies for cross-site evaluation. System dialogue behaviors have a large impact on users' experience and on users' behavior. Furthermore, users in the HC dialogues rarely took initiative and their utterances showed very little variation (Doran et al., 2001). In addition, we believe that once the system side of the dialogues is labelled, it will be much easier to label the user side of the dialogues.

S: Leaving from Phoenix, And, what city are you flying to?
U: SEOUL KOREA
S: Flying from Phoenix to Seoul, What date would you like to fly?
U: WEDNESDAY OCTOBER ELEVENTH
S: Leaving on the 11th of October, What time woodja like to depart?
U: ANY TIME
S: From Seoul. Are you returning to Phoenix?
U: YES
S: And what date didja wanna return?
U: SUNDAY OCTOBER FIFTEENTH
S: I may have misheard You saying the 5th of October. Please be aware that the return date must be later than the departure date. Please tell me what date you would like too return.

Figure 1: Excerpt of a HC Dialogue from the June-2000 Corpus

U: hello
A: people's travel. what city do you want to fly to
U: chicago
A: on what date
U: on the twenty second of may there will be two people travelling
A: what time do you need to depart
U: as soon as possible after five p.m.
A: the first flight after five p.m. on that date is at five thiry five p.m. arriving in chicago at six oh six p.m. on u.s. air
U: is that o'hare
A: yes it is
U: okay. i'll take that
A: on what date do you want to return
U: the following sunday
A: at what time do you need to depart
U: i want to arrive no later than seven in the evening

Figure 2: Excerpt of a HH Dialogue from the CMU-Corpus

We report the results of applying a rule-induction method to train and test DATE taggers on various combinations of the DARPA Communicator June-2000 and October-2001 HC corpora, and the CMU HH corpus in the travel planning domain. The accuracy of a DATE tagger trained and tested on the June-2000 corpus is 98.5%. On the October-2001 corpus, this tagger achieves an accuracy of only 71.8%, but adding 2000 utterances from the 2001 corpus to the training data improves accuracy on the rest of the 2001 corpus to 93.8%. The accuracy of a tagger trained on the HC corpora and tested on the CMU-corpus is 36.7% (a significant improvement over the baseline of 28%). A DATE tagger trained on 305 examples of the HH data achieves

[1] It should be noted, however, that the task of labelling the utterances in the pattern database was still a considerable simplification over entire dialogues being labelled by human annotators, which is the usual practice in tagging tasks.

an accuracy of 48.75%, but the addition of the HC training data improves accuracy to 55.5% (majority class baseline=28%). This pair of results demonstrates quantitatively that the HC data can be used to improve performance of a tagger for HH data. However, a larger training corpus of HH data improves performance to 76.6% accuracy, as estimated by 20-fold cross-validation on the CMU-corpus.

Section 2 describes the corpora, the DATE dialogue act tagging scheme, methods for tagging the corpora for the experiments, and the features used to train a DATE dialogue act tagger for DATE labelling of the corpora. Section 3 presents our results. We postpone discussion and comparison with related work till Section 4.

2 Corpus, Data, Methods

Our experiments apply the rule learning program RIPPER (Cohen, 1996) to train a DATE dialogue act tagger for the utterances of the information provider in HC and HH travel planning dialogues. Like other automatic classifiers, RIPPER takes as input the names of a set of *classes* to be learned, the names and ranges of values of a fixed set of *features*, and *training data* specifying the class and feature values for each example in a training set. Its output is a *classification model* for predicting the class of future examples. In RIPPER, the classification model is learned using greedy search guided by an information gain metric, and is expressed as an ordered set of if-then rules. Although any of several automatic classifiers could be used to train an automatic DATE tagger, RIPPER supports textual features, which are important for this problem, and outputs if-then rules that are easy to understand and which make clear which features are useful to the DATE tagger when classifying utterances.

To apply RIPPER, the utterances in the corpus must be encoded in terms of a set of classes (the output classification) and a set of input features that are used as predictors for the classes. Below we describe the corpora, the classes derived from the DATE tagging scheme, the methods used for tagging the corpora using the DATE scheme, and the features that are extracted from the dialogue in which each utterance occurs.

2.1 Travel Planning Corpora

Our experiments utilize both HC and HH dialogues in the travel planning domain. The DARPA Communicator HC dialogue corpus consists of the June-2000 corpus and the October-2001 corpus. The June-2000 corpus contains 663 experimental dialogues collected during a three week period in June of 2000 for conversations between human users and 9 different Communicator travel planning systems. The October-2001 corpus contains 1252 experimental dialogues collected between April and October of 2001 for conversations between human users and 8 different COMMUNICATOR travel planning systems. The dialogues were quite complex, ranging between simple one way trips requiring no ground arrangements to multileg trips to international or domestic destinations that required car and hotel arrangements. The dialogues typically lasted between 2 and 10 minutes. There was a great deal of variation in the dialogue strategies implemented by the different systems, both between the sites during each collection as well as within a single site across the different collections, from 2000 to 2001. There were a total of 22930 system utterances in the June-2000 corpus and a total of 69766 utterances in the October-2001 corpus. Each dialogue interaction was logged by each system using a shared logfile standard. We were primarily interested in three logged features: (1) the text of each system utterance; (2) what the recognizer understood for each user utterance; and (3) the transcription that each site provided for what the user actually said. We describe below in Section 2.4 how we used these three logfile features to derive the features used to train the DATE tagger.

The HH dialogue corpus consists of the CMU-corpus (Eskenazi et al., 1999). Dialogues in the travel planning domain were collected by the Communicator group at Carnegie Mellon University (CMU), who arranged with the onsite travel agency *People's Travel* to record calls from a number of volunteer subjects who called the human travel agent to plan intended trips. These calls were then transcribed and the recordings and the transcriptions were made available to members of the Communicator community. Labellers at our site subsequently segmented the travel agent side of the conversation into utterances where each utterance realized a single dialogue act. We used this utterance level segmentation to define the unit for tagging in the experiments described below. The CMU-corpus consists of 38 dialogues with a total of 1062 travel agent utterances.

2.2 Class Assignment

The classes used to train the DATE tagger are derived directly from the DATE tagging scheme (Walker et al., 2001c). DATE classifies each utterance along three cross-cutting orthogonal dimensions of utterance classification: (1) a SPEECH ACT dimension; (2) a CONVERSATIONAL-DOMAIN dimension; and (3) a TASK-SUBTASK dimension. The SPEECH ACT and CONVERSATIONAL-DOMAIN dimensions should be general across domains, while the TASK-SUBTASK dimension involves a task model that is not only domain specific, but could vary from system to system because some systems might make finer-grained subtask distinctions.

The SPEECH ACT dimension captures distinctions between distinct communicative goals such as requesting information (REQUEST-INFO), presenting information (PRESENT-INFO) and making offers (OFFER) to act on behalf of the caller. The types of speech acts are specified and illustrated in Figure 3.

Speech-Act	Example
REQUEST-INFO	*And, what city are you flying to?*
PRESENT-INFO	*The airfare for this trip is 390 dollars.*
OFFER	*Would you like me to hold this option?*
ACKNOWLEDGMENT	*I will book this leg.*
BACKCHANNEL	*Okay.*
STATUS-REPORT	*Accessing the database; this might take a few seconds.*
EXPLICIT-CONFIRM	*You will depart on September 1st. Is that correct?*
IMPLICIT-CONFIRM	*Leaving from Dallas.*
INSTRUCTION	*Try saying a short sentence.*
APOLOGY	*Sorry, I didn't understand that.*
OPENING-CLOSING	*Hello. Welcome to the C M U Communicator.*

Figure 3: Example Speech Acts in DATE

The CONVERSATIONAL-DOMAIN dimension distinguishes between talk devoted to the task of booking airline reservations ("about-task") versus talk devoted to maintaining the verbal channel of communication ("about-communication") (Allen and Core, 1997). DATE adds a third domain called "about-situation-frame", to distinguish utterances that provide information about the interactional context, e.g. *Try saying a short sentence*, or *I know about 500 international destinations*.

The TASK-SUBTASK dimension focusses on specifying which subtask of the travel reservation task the utterance contributes to. Some examples are given in Figure 4. This dimension distinguishes among 28 subtasks, some of which can also be grouped at a level below the top level task. The TOP-LEVEL-TRIP task describes the task which contains as its subtasks the ORIGIN, DESTINATION, DATE, TIME, AIRLINE, TRIP-TYPE, RETRIEVAL and ITINERARY tasks. The GROUND task includes both the HOTEL and CAR subtasks. The HOTEL task includes both the HOTEL-NAME and HOTEL-LOCATION subtasks.

Some utterances, especially about-situation-frame utterances such as instructions and apologies are not specific to any task. For example, apologies made by the system about a misunderstanding can be made within any subtask. We give these utterances a "meta" value in the task dimension.

Task	Example
TOP-LEVEL-TRIP	*What are your travel plans?*
ORIGIN	*And, what city are you leaving from?*
DESTINATION	*And, where are you flying to?*
DATE	*What day would you like to leave?*
TIME	*Departing at what time?.*
AIRLINE	*Did you have an airline preference?*
RETRIEVAL	*Accessing the database; this might take a few seconds.*
ITINERARY	*The airfare for this trip is 390 dollars.*
GROUND	*Did you need to make any ground arrangements?.*
HOTEL	*Did you need a hotel?.*
HOTEL-NAME	*Do you have a preferred hotel chain?.*
HOTEL-LOCATION	*Would you like a hotel near downtown or near the airport?.*
CAR	*Do you need a car in San Jose?*
CAR-TYPE	*What kind of car did you want?*
CAR-RENTAL	*Do you have a preferred rental agency?*

Figure 4: Example Subtasks in DATE

It is possible to achieve very specific labelling of system utterances by applying all three dimensions simultaneously. For example, one set of output classes for the DATE tagger consists of the combination of all three classes so that an utterance such as *I found three flights that match your request* is classified as ABOUT-TASK:PRESENT-INFO:FLIGHT.[2] However, the DATE scheme also makes it possible to train and test a DATE tagger for just the SPEECH-ACT dimension or just the TASK dimension. Figure 5 shows utterances from a June-2000 dialogue fragment that are classified along each of the three DATE dimensions.

Tagging utterances along the SPEECH ACT dimension provides the most general tagging. This level of categorization is task-independent and possibly situation independent, ie. from HC to HH dialogues. One set of experiments simply tests performance of a DATE tagger for the speech-act dimension on the HC dialogue data. In addition, we also train a DATE tagger on the HC dialogues using only the speech

[2] DATE labels that are specified for all the three dimensions have the dimension values given in three fields separated by ":". The first field contains the value for the Conversational-Domain Dimension, the second for the Speech-Act Dimension, and the third for the Task-Subtask Dimension.

act dimension for the purpose of applying it to a test set of the CMU-corpus of HH dialogues.[3]

2.3 Preparation of Training and Test Data via DATE Tagging

The DATE labelling of the June-2000 data was done with a semi-automatic tagger: an utterance or utterance sequence is identified and labelled automatically by reference to a database of utterance patterns hand-labelled with DATE tags. The collection and DATE labelling of the utterance patterns was done in cooperation with site developers. As discussed above, these patterns for the 2000 data set were often quite specific, and often involved whole utterances. However, since the systems use template based generation and have only a limited number of ways of saying the same content, relatively few utterance patterns needed to be hand-labelled when compared to the actual number of utterances occurring in the corpus. Further abstraction on the patterns was done with a named-entity labeller which replaces specific tokens of city names, airports, hotels, airlines, dates, times, cars, and car rental companies with their generic type labels. For example, *what time do you want to leave <AIRPORT> on <DATE-TIME>?* is the typed utterance for *what time do you want to leave Newark International on Monday?*. For the 2000 tagging, the number of utterances in the pattern database was 1700 whereas the total number of utterances in the 663 dialogues was 22930. The named-entity labeller was also applied to the system utterances in the corpus. We collected vocabulary lists from all the sites for the named-entity labelling task. In most cases, systems had preclassified the individual tokens into generic types.

The tagger implements a simple pattern matching algorithm to do the dialogue act labelling: for each utterance pattern in the pattern database, the tagger attempts to find a match in the dialogues; if the match succeeds, the DATE label of that pattern is assigned to the matching utterance in the dialogue. The matching ignores punctuation since systems vary in the way they record punctuation.[4]

Certain utterances have different communicative functions depending on the context in which they occur. For example, phrases like *leaving in the <DATE-TIME>* are implicit confirmations when they constitute an utterance on their own, but are part of the flight information presentation when they occur embedded in utterances such as *I have one flight leaving in the <DATE-TIME>*. To prevent incorrect labelling for such ambiguous cases, the pattern database is sorted so that sub-patterns are listed later than the patterns within which they are embedded, and the pattern matcher is forced to match patterns in their order of occurrence in the database.

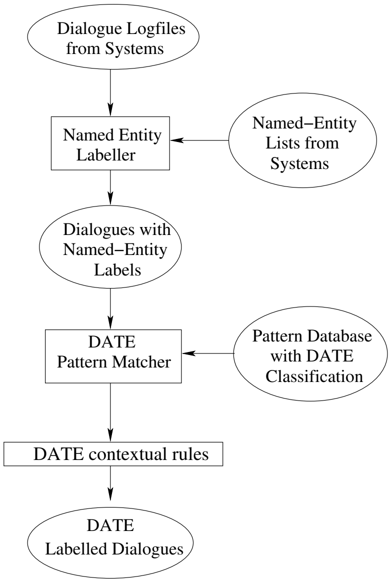
Dialogue Logfiles from Systems
Named Entity Labeller
Named–Entity Lists from Systems
Dialogues with Named–Entity Labels
DATE Pattern Matcher
Pattern Database with DATE Classification
DATE contextual rules
DATE Labelled Dialogues

Figure 6: The DATE Dialogue Act Tagger

While this tagger achieved 100% accuracy for the 2000 data by using many specific patterns, when applied to the 2001 corpus it was able to label only 60% of the data. On examination of the unlabelled utterances, we found that many systems had augmented their inventory of named-entity items as well as system utterances from the 2000 to the 2001 data collection. As a result, there were many new patterns unaccounted for in the existing named-entity lists as well as in the pattern database. In an attempt to cover the remaining 40% of the data, we therefore augmented the named-entity lists by obtaining a new set of preclassified vocabulary items from the sites, and added 800 hand-labelled patterns to the pattern database. For the labelling of any additional unaccounted-for patterns, we implemented a contextual rule-based postprocessor that looks at the surrounding dialogue acts of an unmatched utterance within a turn and attempts to label it. The contextual rules are intended to capture

[3] Tagging utterances along the TASK dimension may provide a rough notion of discourse segmentation in that utterances about the same task may be grouped together. Due to lack of space, however, we do not present results for task tagging.

[4] Ignoring punctuation does not, however, create an utterance segmentation problem for the tagger. We assume that the utterances in the pattern database provide the reference points for utterance boundaries.

rigid system dialogue behaviors that are reflected in the DATE sequences within a turn.[5] For example, one very frequently occurring DATE sequence within system turns is about_task:present_info:flight, about_task:present_info:price, about_task:offer:flight. The rule using this contextual information can be informally stated as follows: if in a turn, the first two utterances are labelled as about_task:present_info:flight and about_task:present_info:price, and the third utterance is unlabelled, assign the third utterance the label about_task:offer:flight. Not all turn-internal DATE sequences are used as contextual rules, however, because many of them are highly ambiguous. For example, about_communicaton:apology:meta_slu_reject can be followed by a system instruction as well as any kind of request for information (typically) repeated from the previous system utterance. Figure 6 shows the current DATE tagging system, augmented with the DATE rule-based postprocessor.

With the 2000 tagger augmented with the additional named-entity items, utterance patterns, and the postprocessor, we were able to label 98.4% of the (69766) utterances in the 2001 corpus.

We conducted a hand evaluation of 10 dialogues which we selected randomly from each system. The evaluation of the total 80 dialogues shows that we achieved 96% accuracy on the 2001 tagging.

In order to label the HH corpus of 1062 utterances, we started with 10 dialogues (305 utterances) labelled with the CSTAR dialogue act tagging scheme (Finke et al., 1998; Doran et al., 2001). We automatically converted the labels to DATE, and then hand-corrected them. We labelled the rest of the HH data by training a DATE tagger, applying it to the remainder of the corpus, and hand-correcting the results.

2.4 Feature Extraction

The corpus is used to construct the machine learning features as follows. In RIPPER, feature values are continuous (numeric), set-valued (textual), or symbolic. We encoded each utterance in terms of a set of 19 features that were either derived from the logfiles, derived from human transcription of the user utterances, or represent aspects of the dialogue context in which each utterance occurs.

The complete feature set used by the machine learner is described in Figure 7. The features fall into three categories: (1) **target utterance features** ; (2) **context features** ; and (3) **whole dialogue features**.

- **target utterance features**: *utt-string*, *contains-word-FLIGHT-or-AIRLINE*, *contains-word-HOTEL-or-ROOM*, *contains-word-RENTAL-or-CAR*, *contains-word-CITY-or-AIRPORT*, *contains-word-DATE-TIME*, *pattern-length*.

- **context features**: *left-sys-utt-string*, *right-sys-utt-string*, *da-num*, *position-in-turn*, *left–dacontext1*, *left-da-context2*, *usr-orig-string*, *usr-typed-string*, *rec-orig-string*, *rec-typed-string usr-rec-string-identity*.

- **whole dialogue features**: *system-name*, *turn-number*.

Figure 7: Features used by the Machine Learner

The **target utterance** features include the target utterance string for which the dialogue act is to be predicted (*utt-string*), and a set of features derived from the named-entity labelling about what semantic types are instantiated in the target string. For example the feature *contains-word-FLIGHT-or-AIRLINE* is represented by a boolean variable specifying whether the utterance string contains the words FLIGHT or AIRLINE. Similar features are *contains-word-HOTEL-or-ROOM*, *contains-word-RENTAL-or-CAR*, *contains-word-CITY-or-AIRPORT*, and *contains-word-DATE-TIME*. The *pattern-length* feature encodes the character length of the target utterance. The motivation for these features is to represent basic aspects of the target utterance, e.g. its length, and the lexical items and semantic types that appear in the utterance.

The **context features** encode simple aspects about the context in which the target utterance occurs. Two of these represent the system utterance strings to the left and right of the target utterance (*left-sys-utt-string* and *right-sys-utt-string*). The *left-da-context1* and *left-da-context2* features represent the left unigram and bigram dialogue act context of the target utterance; this goes beyond the target turn to only the last dialogue act in the previous system turn. The *da-num* feature encodes the number of dialogue acts in the target turn and the *position-in-turn* feature encodes the position of the target utterance in its turn. In addition, the user's previous utterance is represented as part of the context, both in terms of automatically extractable features like what the automatic speech recognizer thought the user said (*rec-orig-string*), and a version of this on which the named-entity labeller has been run (*rec-typed-string*), as well as in terms of human generated transcriptions of the user's utterance. Features based on the transcriptions include the original human transcription (*usr-orig-string*) and the transcription after named-entity tagging (*usr-typed-string*). The *usr-rec-string-identity* feature is a

[5] The logfile standard distinguishes system and user turns within the dialogues.

Training Data	Test Data	Dim	Maj. Cl. Baseline	Acc.	(SE)
JUNE-2000	4fold Xval JUNE-2000	All	6.45%	98.5%	0.11%
JUNE-2000	OCTOBER-2001	All	9.52%	71.82%	0.17%
JUNE-2000 & 2000 examples of October-2001	October-2001 w/out 2000	All	10.18%	93.82%	0.09%

Table 1: Results for Identifying Three-Way DATE Tags in the October-2001 Communicator Corpus, (Dim = Dimension of Date used for output classification (Maj. Cl. = Majority Class, Acc = Accuracy, SE = Standard Error)

boolean feature based on comparing the user's transcribed utterance with the recognizer's hypothesis of what the user said, using simple string-identity. Some applications of DATE tagging would not use features derived from human generated transcriptions so the experiments below report accuracy figures for DATE taggers which ignore these features. The motivation for the context features is to represent aspects of the context in which the utterance occurs in terms of a window of surrounding lexical items and dialogue acts.

The **whole dialogue features** are the name of the site whose system generated the dialogue (*system-name*), and the turn number of the target utterance within the whole dialogue (*turn-number*). For HH dialogues the *system-name* has the value "human". The motivation for including the *system-name* feature is to see whether there are any aspects of the dialogue act realizations that are specific to particular systems. The motivation for the *turn-number* feature is that particular types of dialogue acts are more likely to occur in particular phases of the dialogue.

3 Results

Given the corpora and features described above, we constructed a set of training and test files for use with the RIPPER engine. Each spoken dialogue utterance by the system or by the human travel agent in the corpora are represented in terms of the features and class values described above. One of the primary goals in these experiments is to test the ability of the trained DATE tagger to learn and apply general rules for dialogue act tagging. In the HC data, we examine how a DATE tagger trained on the June-2000 corpus performs on the October-2001 corpus, with and without 2000 labelled examples of October-2001 training data. For the HH data, we examine how a DATE tagger trained on the two HC corpora (June-2000 and October-2001) performs on the CMU-corpus, with and without 305 utterances of HH labelled training data. We first report accuracy results for a DATE tagger trained and tested on the HC June-2000 and October-2001 corpora and then report results for the HH CMU-corpus.

Human-Computer Results: Table 1 shows that the reported accuracies for the HC experiments are signifcantly better than the baseline in each case and the differences between the rows are also statistically significant. The first row shows that the accuracy of a DATE tagger trained and tested using four-fold cross-validation on the June-2000 data is 98.5% with a standard error of only 0.11%. This indicates that after training on 75% of the data, there are few unexpected utterances in the remaining 25%. However, the second row shows that a DATE tagger trained on the 9 systems represented in the June-2000 corpus and tested on the (subset) 8 systems represented in the October-2001 corpus only achieves 71.82% accuracy. This roughly matches our earlier finding in Section 2.3 that during the interval from June-2000 to April-2001 when the 2001 data collection began, many changes had been made to the Communicator systems and that the learned rules from the June-2000 data were not able to generalize as well to the October-2001 corpus.. The third row shows that the overall variation in the data is still low: when 2000 labelled examples of the October-2001 data are added to the June-2000 data for training, the accuracy increases to 93.82%. This suggests that adding a small amount of new labelled training data for successive versions of a system would support high accuracy DATE tagging for the new version of the system.

Some of the rules that RIPPER learned from the HC corpora for predicting the DATE tag for utterances requesting information about the origin city, e.g. *What city are you departing from?*, and requesting information about the destination city, e.g. *Where are you traveling to?*, are shown in Figure 8. The figure shows that all of the rules for both *about_task:request_info:orig_city* and small *about_task:request_info:dest_city* utilize the utterance string feature. This suggests that single words in utterances can be regarded as reliable indicators

Training Data	Test Data	Dim	Maj. Cl. Baseline	Acc.	(SE)
JUNE-2000	4fold Xval JUNE-2000	SPA	31.28%	99.1%	.09%
JUNE-2000	OCTOBER-2001	SPA	31.28%	82.57%	0.14%
JUNE-2000 & 2000 examples of October-2001	October-2001 w/out 2000	SPA	30.88%	95.68%	0.08%

Table 2: Results for Identifying Speech-Act DATE tags in the October-2001 Communicator Corpus, (Dim = Dimension of Date used for output classification (SPA = Speech Act, Maj. Cl. = Majority Class, Acc = Accuracy, SE = Standard Error)

of DATE tags. More interestingly, the words utilized are intuitively plausible for the travel planning domain. For example, the learned question words such as *which, where* and *would* are significant for utterances that have *request_info* as their SPEECH-ACT dimension. The words *city, airport, from, destination* and *departing* are significant predictors of utterances that have *orig_city* and *dest_city* as their task dimension.

```
if utt-string contains city $\wedge$ utt-string contains from $\wedge$ pattern-length $\leq 38$
or if utt-string contains airport $\wedge$ pattern-length $\leq 38$
or if utt-string contains city $\wedge$ pattern-length $\leq 17$ $\wedge$ pattern-length$\geq 15$
or if utt-string contains from $\wedge$ pattern-length $\leq 66$ $\wedge$ utt-string contains Where
or if utt-string contains city $\wedge$ utt-string contains say
or if utt-string contains DEPARTING
or if utt-string contains which $\wedge$ utt-string contains From
or if utt-string contains city $\wedge$ system-name=IBM $\wedge$ utt-string contains departure
or if utt-string contains fly $\wedge$ utt-string contains which $\wedge$ left-sys-utt-string contains city
or if utt-string contains fly $\wedge$ utt-string contains O
then about_task:request_info:orig_city
```

```
if utt-string contains where $\wedge$ utt-string contains must
or if utt-string contains city $\wedge$ pattern-length $\leq 35$
or if utt-string contains Where
or if utt-string contains destination
or if utt-string contains DESTINATION
or if utt-string contains which $\wedge$ utt-string contains city
or if utt-string contains where
or if utt-string contains WOULD
then about_task:request_info:dest_city
```

Figure 8: Rules for DATE tags *about_task:request_info:orig_city* and *about_task:request_info:dest_city* for Training on the June-2000 Corpus and 2000 Examples of October-2001 Corpus.

Human-Computer Speech-Act Results: Because the DATE scheme describes utterances in terms of SPEECH-ACT, CONVERSATIONAL-DOMAIN and TASK dimensions, it is also possible to extract from the composite labels and examine the DATE tagger performance for the individual dimensions. Here we focus on the SPEECH-ACT dimension since, as mentioned above, it is more likely to generalize to HH travel dialogues and to other task domains. Table 2 shows the results for a DATE tagger trained and tested on only the SPEECH-ACT dimension. The reported accuracies are signifcantly better than the baseline in each case and the differences between the rows are also statistically significant. The results support our original hypothesis, showing that the June-2000 SPEECH-ACT DATE tagger generalizes more readily to the October-2001 corpus, with an accuracy of 82.57% (Row 2). Furthermore, as before, even a small amount of training data from the 2001 corpus makes a significant improvement in accuracy to 95.68% (Row 3), which is close to the 99.1% accuracy (Row 1) reported for training and testing on the June-2000 corpus as estimated by 4-fold cross-validation.

Human-Human Results: In order to examine whether there is any generalization from labelled HC data to HH data for the same task, we apply a DATE tagger trained on only the SPEECH-ACT dimension. The first row of Table 3 shows that when a DATE tagger is trained on only the HC corpus and tested on the HH corpus that the accuracy is 36.72% (a significant improvement over the baseline). This result demonstrates quantitatively that the HC data can be used to improve performance of a tagger for HH data.

Now, let us consider a situation where we only have 305 HH labelled utterances from 10 of the HH dialogues to train a DATE tagger. Row 2 shows that we achieve 48.75% accuracy when testing on the remainder of the HH corpus. However if we add the HC data to the training set, the accuracy improves significantly to 55.48% (Row 3). Again this result demonstrates quantitatively that the HC data can improve performance of a tagger for HH data.

Row 4 shows that the utility of the HC corpus decreases if larger amounts of HH labelled data are available; using 95% of the data to train and test-

Training Data	Test Data	Maj. Cl. Baseline	Acc.	(SE)
JUNE-2000 & OCTOBER-2001	CMU-CORPUS	28.07%	36.72%	2.76%
305 CMU-CORPUS	CMU-CORPUS - 305	43.93%	48.75%	1.82 %
JUNE-2000, OCTOBER-2001 & 305 CMU-CORPUS	CMU-CORPUS - 305	28.04%	55.48%	1.81 %
CMU-CORPUS	20fold Xval CMU-CORPUS	54.14%	76.56%	1.03 %

Table 3: Results for Identifying DATE Speech-Act Tags in the CMU Human-Human Corpus (Maj. Cl. = Majority Class, Acc. = Accuracy, SE = Standard Error)

ing on 5% with 20-fold cross-validation achieves an accuracy of 76.56%.

Examination of the errors that the tagger makes indicates both similarities and differences between HH and HC dialogues. For example, information is presented in small installments in the HH dialogues whereas information presentation utterances in the HC dialogues tend to be very long. The information presentation utterances in HH dialogues then appear to be syntactically similar to the implicit confirmations in the HC data. Finally, some utterance types that are very frequent in the HC data such as instructions rarely occur in the HH dialogues.

The rules that are learned for a DATE tagger trained on the HC corpora and the HH CMU-corpus for the offer SPEECH-ACT are in Figures 9 and 10. There are two main conclusions that can be drawn from these figures about the generalization from HC to HH corpora in the SPEECH-ACT dimension. First, in general, a larger number of rules are learned for the HH data, suggesting that there is greater variation for the same speech act in HH dialogues. While this is not surprising, there is also significant overlap in the features and values used in the rules. For example, the utterance string feature utilizes words such as *select, flight, do, okay, fine, these* in both rule sets.

4 Discussion and Future Work

In summary our results show that: (1) It is possible to assign DATE dialogue act tags to system utterances in HC dialogues from many different systems for the same domain with high accuracy; (2) A DATE tagger trained on data from an earlier version of the system only achieves moderate accuracy on a later version of the system without a small amount labelled training data from that later version; (3) Labelled training data from HC dialogues can improve the performance of a DATE tagger for HH dialogue when only a small amount of HH training data is available.

Previous work has also reported results for dialogue act taggers, using similar features to those we use, with accuracies ranging from 62% to 75% (Reithinger and Klesen, 1997; Shriberg et al., 2000; Samuel et al., 1998). Our best accuracy for the HC data is 98%. The best performance for the HH corpus is 76% accuracy for the cross-validation study using only HH data. However, accuracies reported for previous work are not directly comparable to ours for several reasons. First, some of our results concern labelling the system side of utterances in HC dialogues for the purpose of automatic evaluation of system performance. It is much easier to develop a high accuracy tagger for HC dialogue than it is for HH dialogue.

We also applied the DATE tagger to HH dialogue, and focused on the travel agent side of the dialogue. Here the accuracies that we report are more comparable with that of other researchers, but large differences should nevertheless be expected due to differences in the types of corpora, dialogue act tagging schemes, and features used.

We considered the possibility of generating dialogue acts automatically in the logfiles. This idea was attractive because it is possible to easily implement the generation of dialogue acts tags in the logfiles. Large amounts of human-computer data would then be available for the human-human labelling task or for evaluation efforts. However, this turned out to be impractical because we found it difficult to get dialogue designers across the different participating sites to agree on a labelling standard. We therefore believe that machine learning methods for classification such as the one discussed here might still be necessary to automate the tagging task for rapid evaluation and labelling efforts.

As part of the ISLE NSF/EU project, the labelled corpus that we developed for this work will soon be released by the LDC, and other researchers will then be able to utilize it to improve upon our results. In addition, we believe this corpus could be useful as a training resource for spoken response generation in dialogue systems. For example, the dialogue act representation can be used to provide a broad range of text-planning inputs for a stochastic sentence planner in the travel domain (Walker et al., 2001b), or to represent the systems' dialogue strategies for reinforcement learning (Walker, 2000; Scheffler and Young, 2002). In future work, we hope to demonstrate that features derived from the labelling of the system side of the dialogue can also improve performance of a dialogue act tagger for the human utterances in the dialogue, and to conduct additional analyses demonstrating the utility of this representation for cross-site evaluation.

5 Acknowledgments

Thanks to John Aberdeen and Christy Doran for their contribution of a labelled set of 10 HH dialogues from the CMU-corpus. The work reported in this paper was partially funded by DARPA contract MDA972-99-3-0003.

References

J. Allen and M. Core. 1997. Draft of DAMSL: Dialog act markup in several layers. Coding scheme developed by the MultiParty group, 1st Discourse Tagging Workshop, Univ. of Penn, March 1996.

R. Cattoni, M. Danieli, A. Panizza, V. Sandrini, and C. Soria. 2001. Building a corpus of annotated dialogues: the ADAM experience. In *Proc. of the Conference Corpus-Linguistics-2001, Lancaster, U.K.*

W. Cohen. 1996. Learning trees and rules with set-valued features. In *14th Conference of AAAI.*

B. Di Eugenio, P. W. Jordan, J. D. Moore, and R. H. Thomason. 1998. An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proc. of the 36th ACL Conference*.

C. Doran, J. Aberdeen, L. Damianos, and L. Hirschman. 2001. Comparing several aspects of human-computer and human-human dialogues. In *SIGDIAL Workshop in conjuction with Eurospeech 2001.*

M. Eskenazi, A. Rudnicky, K. Gregory, P. Constantinides, R. Brennan, K. Bennett, and J. Allen. 1999. Data collection and processing in the carnegie mellon communicator. In *Proc. of Eurospeech-99*, pages 2695–2698.

M. Finke, M. Lapata, A. Lavie, L. Levin, L. Mayfield Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. 1998. Clarity: Inferring discourse structure from speech. In *AAAI Symposium on Applying Machine Learning to Discourse Processing Proceedings, Stanford, California.*

H. Hastie, R. Prasad, and M. A. Walker. 2002. Automatic evaluation: Using a date dialogue act tagger for user satisfaction and task completion prediction. In *LREC 2002*.

A. Isard and J. C. Carletta. 1995. Replicability of transaction and action coding in the map task corpus. In M. A. Walker and J. Moore, eds., *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pages 60–67.

P. W. Jordan. 2000. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

N. Reithinger and M. Klesen. 1997. Dialogue act classification using language models. In *Proc. of Eurospeech '97*, pages 2235–2238, Rhodes, Greece.

K. Samuel, S. Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. In *Proc. of COLING-ACL*, pages 1150–1156.

K. Scheffler and S. Young. 2002. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *HLT Conference*.

E. Shriberg, P. Taylor, R. Bates, A. Stolcke, K. Ries, D. Jurafsky, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. 2000. Can prosody aid the automatic classification of dialog acts in conversational speech. *Language and Speech: Special Issue on Prosody and Conversation*.

M. A. Walker, D. Litman, C. A. Kamm, and A. Abella. 1997. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proc. of ACL/EACL 97*, pages 271–280.

M. A. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. 2001a. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Proc. of EUROSPEECH 2001.*

M. A. Walker, O. Rambow, and M. Rogati. 2001b. Spot: A trainable sentence planner. In *Proc. of NAACL.*

M. A. Walker, R. Passonneau, and J. E. Boland. 2001c. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In *Proc. of ACL 2001.*

M. A. Walker. 2000. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416.

Speaker	Utterance	C-Domain	Speech Act	Task
SYS	Welcome.	about-task	opening/closing	meta
	You are logged in as a guest user.	about-frame	instruction	meta
	You may say repeat, help me out, start over, or, that's wrong,	about-frame	instruction	meta
	you can also correct and interrupt the system at any time.	about-frame	instruction	meta
	What airport woodja like to fly out of?	about-task	request-info	orig-city
USER	MIAMI FLORIDA			
SYS	Leaving from Miami,	about-comm	implicit-confirm	orig-city
	And, what city are you flying to?	about-task	request-info	dest-city
USER	MINNEAPOLIS MINNESOTA			
SYS	Flying from Miami to Minneapolis,	about-comm	implicit-confirm	orig-city, dest-city
	What date would you like to fly?	about-task	request-info	depart-date
USER	UH MONDAY OCTOBER TWENTY THIRD			
SYS	Leaving on the 23rd of October,	about-comm	implicit-conf	date
	And what time didja wanna leave?	about-task	request-info	time
USER	UH LATE MORNING			
SYS	From Minneapolis.	about-comm	implicit-conf	origin
	Are you returning to Miami?	about-task	request-info	trip-type
USER	UM			
SYS	Are you going back to Miami?	about-task	request-info	trip-type
	Please say yes or no.	about-frame	instruct	meta
USER	NO			
SYS	Okay, it will just take a moment while I look.	about-task	status-report	retrieval
SYS	I found 3 flights from Miami to Minneapolis on October 23rd . For option one I have a flight with United departing at 10 A M and arriving at 3 P M with one stopover and	about-task	present-info	itinerary
	Ticket price is 390 dollars.	about-task	present-info	price
	Please say next option, or, flight details, or, I'll take it.	about-task	offer	itinerary
USER	I'LL TAKE IT			
SYS	Great! I am adding this flight to your itinerary.	about-task	acknowledge	booking
	Okay, Is Minneapolis your final destination?	about-task	request-info	trip-type
USER	YES IT IS			
SYS	Would you like to make any ground arrangements?	about-task	offer	ground
USER	NO			
SYS	Now, would you like to hear a summary of your itinerary?	about-task	offer	itinerary
USER	YES PLEASE			
SYS	Okay, Here's your itinerary. On October 23rd Northwest Flight 123 departs Miami at 2 P M, arrives Pittsburgh at 4 30 P M connecting to Northwest Flight 146 that departs Pittsburgh at 5 P M and arrives Minneapolis at 7 P M.	about-task	present-info	itinerary

Figure 5: Dialogue Illustrating the Conversational-Domain, Speech-Act, and Task-Subtask Dimensions of DATE

if left-context1=SPA-present-info $\wedge$ pattern-length$\leq$25 $\wedge$ position-in-turn$\geq$2 $\wedge$ right-sys-utt-string *contains* none

or if left-context2=SPA-present-info-SPA-present-info $\wedge$ system-name=ATT

or if right-sys-utt-string *contains* also $\wedge$ left-sys-utt-string *contains* at

or if utt-string *contains* select

or if utt-string *contains* confirm

or if utt-string *contains* fine

or if right-sys-utt-string *contains* locations $\wedge$ utt-string *contains* If

or if left-context2=SPA-implicit-confirm-SPA-instruction $\wedge$ utt-string *contains* Which

or if utt-string *contains* Okay $\wedge$ utt-string *contains* flight

or if left-context2=SPA-explicit-confirm-SPA-acknowledgement $\wedge$ utt-string *contains* flight

or if utt-string *contains* these $\wedge$ utt-string *contains* Are

or if rec-orig-string *contains* sixteenth $\wedge$ utt-string *contains* Do

then *offer*

Figure 9: Rules learned for DATE SPEECH-ACT *offer* using June-2000 plus 2000 Examples of October-2001 as Training

if left-sys-utt-string *contains* 'NUMBER' $\wedge$ pattern-length$\leq$25 $\wedge$ right-sys-utt-string *contains* none $\wedge$ utt-string *contains* OK

or if position-in-turn$\geq$2 $\wedge$ left-sys-utt-string *contains* dollars $\wedge$ pattern-length$\leq$55 $\wedge$ contains-word-CITY-or-AIRPORT=false

or if utt-string *contains* this $\wedge$ pattern-length$\leq$37 $\wedge$ contains-word-FLIGHT-or-AIRLINE=true

or if left-sys-utt-string *contains* per $\wedge$ da-num$\leq$2

or if right-sys-utt-string *contains* rate

or if utt-string *contains* these

or if utt-string *contains* itinerary $\wedge$ pattern-length$\leq$41

or if utt-string *contains* reservation

or if utt-string *contains* select

or if utt-string *contains* book $\wedge$ utt-string *contains* it

or if utt-string *contains* whether

or if utt-string *contains* OK $\wedge$ utt-string *contains* Is

or if utt-string *contains* MAKE

or if utt-string *contains* one $\wedge$ right-sys-utt-string *contains* none

or if utt-string *contains* fine

or if utt-string *contains* Kay

or if right-sys-utt-string *contains* locations

or if utt-string *contains* THE $\wedge$ utt-string *contains* LIKE

or if left-sys-utt-string *contains* over $\wedge$ utt-string *contains* flight $\wedge$ utt-string *contains* would

or if utt-string *contains* take $\wedge$ utt-string *contains* Do

or if left-sys-utt-string *contains* yes $\wedge$ utt-string *contains* what $\wedge$ utt-string *contains* flight

then *offer*

Figure 10: Rules learned for DATE SPEECH-ACT *offer* using 305 CMU-Corpus Utterances as Training