Journal of Artificial Intelligence Research 9 (1998) 247-293 Submitted 5/98; published 11/98

An Empirical Approach to Temporal Reference Resolution

Janyce M. Wiebe WIEBE@CS.NMSU.EDU
Thomas P. O'Hara TOMOHARA@CS.NMSU.EDU
Thorsten Öhrström-Sandgren SANDGREN@LUCENT.COM
Kenneth J. McKeever KMCKEEVE@REDWOOD.DN.HAC.COM
Department of Computer Science and the Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003

Abstract

Scheduling dialogs, during which people negotiate the times of appointments, are common in everyday life. This paper reports the results of an in-depth empirical investigation of resolving explicit temporal references in scheduling dialogs. There are four phases of this work: data annotation and evaluation, model development, system implementation and evaluation, and model evaluation and analysis. The system and model were developed primarily on one set of data, and then applied later to a much more complex data set, to assess the generalizability of the model for the task being performed. Many different types of empirical methods are applied to pinpoint the strengths and weaknesses of the approach. Detailed annotation instructions were developed and an intercoder reliability study was performed, showing that naive annotators can reliably perform the targeted annotations. A fully automatic system has been developed and evaluated on unseen test data, with good results on both data sets. We adopt a pure realization of a recency-based focus model to identify precisely when it is and is not adequate for the task being addressed. In addition to system results, an in-depth evaluation of the model itself is presented, based on detailed manual annotations. The results are that few errors occur specifically due to the model of focus being used, and the set of anaphoric relations defined in the model are low in ambiguity for both data sets.

1. Introduction

Temporal information is often a significant part of the meaning communicated in dialogs and texts, but is often left implicit, to be recovered by the listener or reader from the surrounding context. When scheduling a meeting, for example, a speaker may ask, "How about 2?" expecting the listener to determine which day is being specified. Recovering temporal information implicitly communicated in the discourse is important for many natural language processing applications. For example, consider extracting information from memos and reports for entry into a database. It would be desirable to enter completely resolved dates and times, rather than incomplete components such as the day or time alone. A specific application for which temporal reference resolution is important is appointment scheduling in natural language between human and machine agents (Busemann, Declerck, Diagne, Dini, Klein, & Schmeier, 1997). To fully participate, the machine agent must be able to understand the many references to times that occur in scheduling dialogs.

Maintaining the temporal context can aid in other aspects of understanding. For example, Levin et al. (1995) and Rosé et al. (1995) found that the temporal context, as part of the larger discourse context, can be exploited to improve various kinds of disambiguation, including speech act ambiguity, type of sentence ambiguity, and type of event ambiguity.

This paper presents the results of an in-depth empirical investigation of temporal reference resolution. Temporal reference resolution involves identifying temporal information that is missing due to anaphora, and resolving deictic expressions, which must be interpreted with respect to the current date. The genre addressed is scheduling dialogs, in which participants schedule meetings with one another. Such strongly task-oriented dialogs would arise in many useful applications, such as automated information providers and phone operators.

A model of temporal reference resolution in scheduling dialogs was developed through an analysis of a corpus of scheduling dialogs. A critical component of any method for anaphora resolution is the focus model used. It appeared from our initial observations that a recency-based model might be adequate. To test this hypothesis, we made the strategic decision to limit ourselves to a local, recency-based model of focus, and to analyze the adequacy of such a model for temporal reference resolution in this genre. We also limit the complexity of our algorithm in other ways. For example, there are no facilities for centering within a discourse segment (Sidner, 1979; Grosz, Joshi, & Weinstein, 1995), and only very limited ones for performing tense and aspect interpretation. Even so, the methods investigated in this work go a long way toward solving the problem.

From a practical point of view, the method is reproducible and relatively straightforward to implement. System results and the detailed algorithm are presented in this paper. The model and the implemented system were developed primarily on one data set, and then applied later to a much more complex data set to assess the generalizability of the model for the task being performed. Both data sets are challenging, in that they both include negotiation, contain many disfluencies, and show a great deal of variation in how dates and times are discussed. However, only in the more complex data set do the participants discuss their real life commitments or stray significantly from the scheduling task.

To support the computational work, the temporal references in the corpus were manually annotated. We developed explicit annotation instructions and performed an intercoder reliability study involving naive subjects, with excellent results. To support analysis of the problem and our approach, additional manual annotations were performed, including anaphoric chain annotations.

The system's performance on unseen test data from both data sets is evaluated. On both, the system achieves a large improvement over the baseline accuracy. In addition, ablation (degradation) experiments were performed, to identify the most significant aspects of the algorithm. The system is also evaluated on unambiguous input, to help isolate the contribution of the model itself to overall performance.

The system is an important aspect of this work, but does not enable direct evaluation of the model, due to errors committed by the system in other areas of processing. Thus, we evaluate the model itself based on detailed manual annotations of the data. Important questions addressed are how many errors are attributable specifically to the model of focus and what kinds of errors they are, and how good is the coverage of the set of anaphoric relations defined in the model and how much ambiguity do the relations introduce. The analysis shows that few errors occur specifically due to the model of focus, and the relations are low in ambiguity for the data sets.

The remainder of this paper is organized as follows. The data sets are described in Section 2. The problem is defined and the results of an intercoder reliability study are presented in Section 3. An abstract model of temporal reference resolution is presented in Section 4 and the high-level algorithm is presented in Section 5. Detailed results of the implemented system are included in Section 6, and other approaches to temporal reference resolution are discussed in Section 7. In the final part of the paper, we analyze the challenges presented by the dialogs to an algorithm that does not include a model of global focus (in Section 8.1), evaluate the coverage, ambiguity, and correctness of the set of anaphoric relations defined in the model (in Section 8.2), and assess the importance of the architectural components of the algorithm (in Section 8.3). Section 9 is the conclusion.

There are three online appendices. Online Appendix 1 contains a detailed specification of the temporal reference resolution rules that form the basis of the algorithm. Online Appendix 2 gives a specification of the input to the algorithm. Online Appendix 3 contains a BNF grammar describing the core set of the temporal expressions handled by the system. In addition, the annotation instructions, sample dialogs, and manual annotations of the dialogs are available on the project web site, http://www.cs.nmsu.edu/~wiebe/projects.

2. The Corpora

The algorithm was primarily developed on a sample of a corpus of Spanish dialogs collected under the JANUS project at Carnegie Mellon University (Shum, Levin, Coccaro, Carbonell, Horiguchi, Isotani, Lavie, Mayfield, Rosé, Van Ess-Dykema & Waibel, 1994). These dialogs are referred to here as the "CMU dialogs." The algorithm was later tested on a corpus of Spanish dialogs collected under the Artwork project at New Mexico State University by Daniel Villa and his students (Wiebe, Farwell, Villa, Chen, Sinclair, Sandgren, Stein, Zarazua, & O'Hara, 1996). These are referred to here as the "NMSU dialogs." In both cases, subjects were asked to set up a meeting based on schedules given to them detailing their commitments. The NMSU dialogs are face-to-face, while the CMU dialogs are like telephone conversations. The participants in the CMU dialogs rarely discuss anything from their real lives, and almost exclusively stay on task. The participants in the NMSU dialogs embellish the schedule given to them with some of their real life commitments, and often stray from the task, discussing topics other than the meeting being planned.

3. The Temporal Annotations and Intercoder Reliability Study

Consider the passage shown in Figure 1, which is from the CMU corpus (translated into English). An example of temporal reference resolution is that utterance (2) refers to 2-4pm, Thursday 30 September.

Because the dialogs are centrally concerned with negotiating an interval of time in which to hold a meeting, our representations are geared toward such intervals. The basic representational unit is given in Figure 2. It is referred to throughout as a *Temporal Unit* (*TU*).

Temporal context: Tuesday 28 September		
s1	1	On Thursday I can only meet after two pm
	2	From two to four
	3	Or two thirty to four thirty
	4	Or three to five
s2	5	Then how does from two thirty to four thirty seem to you
	6	On Thursday
s1	7	Thursday the thirtieth of September

Figure 1: Corpus Example

((start-month,	start-date,	start-day-of-week,	start-hour&minute,	start-time-of-day)
(end-month,	end-date,	end-day-of-week,	end-hour&minute,	end-time-of-day))

Figure 2: The Temporal Unit Representation

For example, the time specified[1] in "From 2 to 4, on Wednesday the 19th of August" is represented as follows:

((August,	19,	Wednesday,	2,	pm)
(August,	19,	Wednesday,	4,	pm))

Thus, the information from multiple noun phrases is often merged into a single representation of the underlying interval specified by the utterance.

Temporal references to times in utterances such as "The meeting starts at 2" are also represented in terms of intervals. An issue this kind of utterance raises is whether or not a speculated end time of the interval should be filled in, using knowledge of how long meetings usually last. In the CMU data, the meetings all last two hours, by design. However, our annotation instructions are conservative with respect to filling in an end time given a starting time (or vice versa), specifying that it should be left open unless something in the dialog explicitly suggests otherwise. This policy makes the instructions applicable to a wider class of dialogs.

Weeks, months, and years are represented as intervals starting with the first day of the interval (for example, the first day of the week), and ending with the last day of the interval (for example, the last day of the week).

Some times are treated as points in time (for example, the time specified in "It is now 3pm"). These are represented as Temporal Units with the same starting and end times (as

1. Many terms have been used in the literature for the relation between anaphoric expressions and discourse entities. For example, Sidner (1983) and Webber (1983) argue that "refer" should be reserved for something people do with words, rather than something words do. Webber uses the term "evoke" for first references to an entity and "access" for subsequent references. Sidner uses the term "specify" for the relation between a noun phrase and a discourse entity. We primarily use Sidner's term, but use "refer" in a few contexts in which it seems more natural.

in Allen, 1984). If just the starting or end time is specified, all the fields of the other end of the interval are null. And, of course, all fields are null for utterances that do not contain any temporal information. In the case of an utterance that specifies multiple, distinct intervals, the representation is a list of Temporal Units (for further details of the coding scheme, see O'Hara, Wiebe, & Payne, 1997).

Temporal Units are also the representations used in the evaluation of the system. That is, the system's answers are mapped from its more complex internal representation (an *ILT*, see Section 5.2) into this simpler vector representation before evaluation is performed.

The evaluation Temporal Units used to assess the system's performance were annotated by personnel working on the project. The training data were annotated by the second author of this paper, who also worked on developing the rules and other knowledge used in the system. However, the test data were annotated by another project member, Karen Payne, who contributed to the annotation instructions and to the integration of the system with the Enthusiast system (see below in Section 5.2), but did not contribute to developing the rules and other knowledge used in the system.

As in much recent empirical work in discourse processing (see, for example, Arhenberg, Dahlbäck, & Jönsson, 1995; Isard & Carletta, 1995; Litman & Passonneau, 1995; Moser & Moore, 1995; Hirschberg & Nakatani, 1996), we performed an intercoder reliability study investigating agreement in annotating the times. The main goal in developing annotation instructions is to make them precise but intuitive so that they can be used reliably by non-experts after a reasonable amount of training (see Passonneau & Litman, 1993; Condon & Cech, 1995; Hirschberg & Nakatani, 1996). Reliability is measured in terms of the amount of agreement among annotators; high reliability indicates that the encoding scheme is reproducible given multiple annotators. In addition, the instructions also serve to document the annotations.

The subjects were three people with no previous involvement in the project. They were given the original Spanish and the English translations. However, as they have limited knowledge of Spanish, in essence they annotated the English translations.

The subjects annotated two training dialogs according to the instructions. After receiving feedback, they annotated four unseen test dialogs. Intercoder reliability was assessed using Cohen's Kappa statistic (κ) (Siegel & Castellan, 1988; Carletta, 1996). Agreement for each Temporal Unit field (for example, *start-month*) was assessed independently.

κ is calculated as follows:

$$\kappa = \frac{Pa - Pe}{1 - Pe}$$

The numerator is the average percentage agreement among the annotators (Pa) less a term for expected chance agreement (Pe), and the denominator is 100% agreement less the same term for chance agreement (Pe).

Pa and Pe are calculated as follows (Siegel & Castellan, 1988). Suppose that there are N objects, M classes, and K taggers. We have the following definitions.

- n_{ij} is the number of assignments of object i to category j. Thus, for each i, $\sum_{j=1}^{M} n_{ij} = K$.
- $C_j = \sum_{i=1}^{N} n_{ij}$, the total number of assignments of objects to category j.

- $p_j = \frac{C_j}{N \times K}$, the percentage of assignments to category j (note that $N \times K$ is the total number of assignments).

We can now define Pe:

$$Pe = \sum_{j=1}^{M} p_j^2$$

The extent of agreement among the taggers concerning the *ith* object is S_i, defined as follows. It is the total number of actual agreements for object i, over the maximum possible agreement for one object:

$$S_i = \frac{\sum_{j=1}^{M} \binom{n_{ij}}{2}}{\binom{K}{2}}$$

Finally, Pa is the average agreement over objects:

$$Pa = \frac{1}{N} \sum_{i=1}^{N} S_i$$

κ is 0.0 when the agreement is what one would expect under independence, and it is 1.0 when the agreement is exact (Hays, 1988). A κ value of 0.8 or greater indicates a high level of reliability among raters, with values between 0.67 and 0.8 indicating only moderate agreement (Hirschberg & Nakatani, 1996; Carletta, 1996).

In addition to measuring intercoder reliability, we compared each coder's annotations to the gold standard annotations used to assess the system's performance. Results for both types of agreement are shown in Table 1. The agreement among coders is shown in the column labeled κ, and the average pairwise κ values for the coders and the expert who performed the gold standard annotations are shown in the column labeled κ_{avg}. This was calculated by averaging the individual κ scores (which are not shown).

There is a high level of agreement among annotators in all cases except the *end time of day* field, a weakness we are investigating. There is also good agreement between the evaluation annotations and the naive coders' evaluations: with the exception of the *time of day* fields, κ_{avg} indicates high average pairwise agreement between the expert and the naive subjects.

Busemann et al. (1997) also annotate temporal information in a corpus of scheduling dialogs. However, their annotations are at the level of individual expressions rather than at the level of Temporal Units, and they do not present the results of an intercoder reliability study.

4. Model

This section presents our model of temporal reference resolution in scheduling dialogs. Section 4.1 describes the cases of deictic reference covered and Section 4.2 presents the

Field	Pa	Pe	κ	κ_{avg}
start				
Month	.96	.51	.93	.94
Date	.95	.50	.91	.93
DayofWeek	.96	.52	.91	.92
HourMin	.98	.82	.89	.92
TimeDay	.97	.74	.87	.74
end				
Month	.97	.51	.93	.94
Date	.96	.50	.92	.94
DayofWeek	.96	.52	.92	.92
HourMin	.99	.89	.90	.88
TimeDay	.95	.85	.65	.52

Table 1: Agreement Among Coders

$$
\begin{tikzcd}
 & \textit{day of week} \arrow[dr] & & \\
\textit{month} \arrow[ur] \arrow[dr] & & \textit{time of day} \arrow[r] & \textit{hour\&minute} \\
 & \textit{date} \arrow[ur] & &
\end{tikzcd}
$$

Figure 3: Specificity Ordering

anaphoric relations defined. Section 4.3 gives some background information about focus models, and then describes the focus model used in this work.

Anaphora is treated in this paper as a relationship between a Temporal Unit representing a time specified in the current utterance ($TU_{current}$) and one representing a time specified in a previous utterance ($TU_{previous}$). The resolution of the anaphor is a new Temporal Unit representing the interpretation, in context, of the contributing words in the current utterance.

Fields of Temporal Units are partially ordered as in Figure 3, from least to most specific. The month has the lowest specificity value.

In all cases of deictic reference listed in Section 4.1 and all cases of anaphoric reference listed in Section 4.2, after the resolvent has been formed, it is subjected to highly accurate, obvious inference to produce the final interpretation. Examples are filling in the day of the week given the month and the date; filling in *pm* for modifiers such as "afternoon"; and filling in the duration of an interval from the starting and end points.

In developing the rules, we found domain knowledge and task-specific linguistic conventions to be most useful. However, we observed some cases in the NMSU data for which syntactic information could be exploited (Grosz et al., 1995; Sidner, 1979). For example, "until" in the following suggests that the first utterance specifies an end time.

"... could it be until around twelve?"
"12:30 there"

A preference for parallel syntactic roles might be used to recognize that the second utterance specifies an end time too. We intend to pursue such preferences in future work.

4.1 Deictic References

The deictic expressions addressed in this work are those interpreted with respect to the dialog date (i.e., "today" in the context of the dialog).

4.1.1 Simple deictic relation

A deictic expression such as "tomorrow" or "last week" is interpreted with respect to the dialog date. (See rule D-simple in Section 5.3.)

4.1.2 Frame of reference deictic relation

A forward time reference is calculated using the dialog date as a frame of reference. Let F be the most specific field in $TU_{current}$ less specific than *time of day* (e.g., the *date* field). The resolvent is the next F after the dialog date, augmented with the fillers of the fields in $TU_{current}$ that are at least as specific as *time of day*. (See rule D-frame-of-reference in Section 5.3.)

Following is an example. Assume that the dialog date is *Monday 19 August*.

Utterance	Interpretation
How about Wednesday at 2?	*2 pm, Wednesday 21 August*

For both this and the *frame of reference* anaphoric relation, there are subcases for whether the starting and/or end times are involved.

4.2 Anaphoric Relations

Generally speaking, many different kinds of relationships can be established between an anaphor and its antecedent. Examples are co-reference ("John saw Mary. He..."), part-whole ("John bought a car. The engine..."), and individual-class ("John bought a truck. They are good for hauling...") (see, for example, Webber, 1983). The latter two involve *bridging descriptions* (see, for example, Clark, 1977; Heim, 1982; Poesio, Vieira, & Teufel, 1997): some reasoning is required to infer the correct interpretation. This section presents a set of anaphoric relations that have good coverage for temporal expressions in scheduling dialogs (see Section 8.2 for an evaluation). Many temporal references involve bridging inferences, in the sense that times are calculated by using the antecedent as a frame of reference or by modifying a previous temporal interpretation.

4.2.1 Co-reference anaphoric relation

The same times are specified, or $TU_{current}$ is more specific than $TU_{previous}$. The resolvent contains the union of the information in the two Temporal Units. (See rule A-co-reference in Section 5.3.)

For example (see also (1)-(2) of the corpus example in Figure 1):

Utterance	Interpretation
How is Tuesday, January 30th?	
How about 2?	*2pm, Tuesday 30 January*

4.2.2 Less-specific anaphoric relation

$TU_{current}$ includes $TU_{previous}$, and $TU_{current}$ is less specific than $TU_{previous}$. Let F be the most specific field in $TU_{current}$. The resolvent contains all of the information in $TU_{previous}$ of the same or lower specificity than F. (See rule A-less-specific in Section 5.3.)

For example (see also (5)-(6) of the corpus example in Figure 1):

Utterance	Interpretation
How about Monday at 2?	*Assume: 2pm, Monday 19 August*
Ok, well, Monday sounds good.	*Monday 19 August*

4.2.3 Frame of reference anaphoric relation

This is the same as the *frame of reference* deictic relation above, but the new time is calculated with respect to $TU_{previous}$ instead of the dialog date. (See rule A-frame-of-reference in Section 5.3.)

Following are two examples:

Utterance	Interpretation
Would you like to meet Wednesday, August 2nd?	
No, how about Friday at 2.	*2pm, Friday 4 August*

Utterance	Interpretation
How about the 3rd week of August?	
Let's see, Tuesday sounds good.	*Tuesday of the 3rd week in August*

In the first example, the day specified in the first utterance is used as the frame of reference. In the second example, the beginning day of the interval representing the 3rd week of August is used as the frame of reference.

Note that tense can influence the choice of whether to calculate a forward or backward time from a frame of reference (Kamp & Reyle, 1993), but this is not accounted for because there is not much tense variation in the CMU corpus on which the algorithm was developed. However, errors can occur because backward calculations are not covered. For example, one might mention "Friday" and then "Thursday," intending "Thursday" to be calculated as the day **before** that Friday, rather than the Thursday of the week following that Friday. We are investigating creating a new anaphoric relation to cover these cases.

4.2.4 Modify anaphoric relation

$TU_{current}$ is calculated by modifying the interpretation of the previous temporal reference. The times differ in the filler of a field F, where F is at least as specific as *time of day*, but are consistent in all fields less specific than F. The resolvent contains the information in $TU_{previous}$ that is less specific than F together with the information in $TU_{current}$ that is of the same or greater specificity as F. (See rule A-modify in Section 5.3.)

For example (see also (3)-(5) of the corpus example in Figure 1):

Utterance	Interpretation
Monday looks good.	*Assume: Monday 19 August*
How about 2?	*(co-reference relation) 2pm, Monday 19 August*
Hmm, how about 4?	*(modify relation) 4pm, Monday 19 August*

4.3 Focus Models

The focus model, or model of attentional state (Grosz & Sidner, 1986), is a model of which entities the dialog is most centrally about at each point in the dialog. It determines which previously mentioned entities are the candidate antecedents of anaphoric references. As such, it represents the role that the structure of the discourse plays in reference resolution.

We consider three models of attentional state in this paper: (1) the linear-recency model (see, for example, the work by Hobbs (1978) and Walker[2] (1996)), (2) Grosz and Sidner's (1986) stack-based model, and (3) the *graph structured stack* model introduced by Rosé, Di Eugenio, Levin, and Van Ess-Dykema (1995). Ordered from (1) to (3), the models are successively more complex, accounting for increasingly more complex structures in the discourse.

In a linear-recency based model, entities mentioned in the discourse are stored on a focus list, ordered by recency. The corresponding structure in the dialog is shown in Figure 4a: a simple progression of references, uninterrupted by subdialogs.

In Grosz and Sidner's stack-based model, the entities in focus in a particular discourse segment are stored together in a *focus space* associated with that segment. To handle anaphoric references across discourse segments, focus spaces are pushed on and popped off the stack as appropriate to mirror the structure of the discourse. As each new segment is recognized, a focus space is created and pushed onto the stack. To interpret an anaphoric reference, the entities in the focus space on the top of the stack are considered first. However, if the current utterance resumes a previous discourse segment, the intervening focus spaces are popped off. This allows anaphoric reference to an earlier entity, even if more recently mentioned entities are possible antecedents (for more details, see Grosz & Sidner, 1986). Figure 4b illustrates a discourse structure that the stack-based model is designed to handle. Suppose that both TU_1 and TU_2 are possible antecedents of TU_3 (for example, suppose they are specified by pronouns that agree in number and gender), but TU_2 is in a subsegment and is not a correct antecedent of TU_3, even though it is mentioned more recently than TU_1. In the stack-based model, the focus space containing TU_2 is popped off the stack when the end of its segment is recognized, thus removing TU_2 as a competitor for understanding TU_3. Following is an example from the NMSU corpus (this is the dialog segment labeled 09-09, in row 7, in Figure 10 presented later).

2. Note that Walker's model is a cache-based model for which recency is a very important but not unique criterion for determining which entities are in the cache.

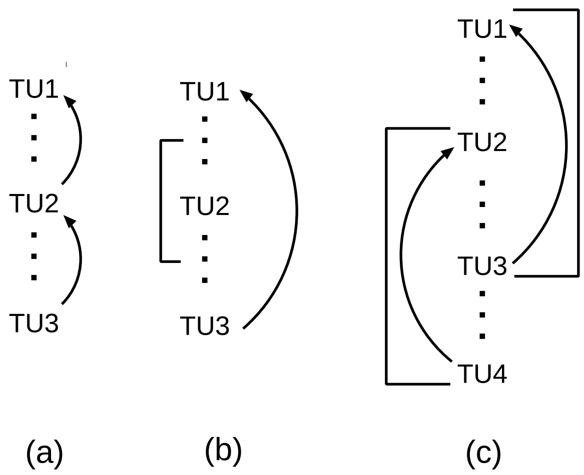

Figure 4: Discourse Structures Targeted by Different Focus Models

			Dialog Date: Monday 10 May
	1	S1	Listen, daughter, I was thinking of inviting you to a demonstration on interior things, ornaments for decorating your house.
	2		Uh, I would like to do it at two p.m. Wednesday,
	3		But I don't know if you are free at that time or ...
TU_1	4	S2	Uh, Wednesday, Mom, well *Resolved to Wednesday, May 12*
	5		You know that,
$TU_{2,1}$	6		last week uh, I got a job and uh, a full-time job *Unambiguous deictic; resolved to the week before the dialog date*
$TU_{2,2}$	7		I go in from seven in the morning to five in the afternoon *Habitual*
	8	S1	Oh, maybe it would be better
TU_3	9	S2	Well, I have lunch from twelve to one *Utterance (4) is needed for the correct interpretation: 12-1, Wednesday 12 May*

In this passage, utterances (6)-(7) are in a subdialog about S2's job. To interpret "twelve to one" in utterance (9) correctly, one must go back to utterance (4). Incorrect interpretations involving the temporal references in (6) and (7) are possible (using the *co-reference* relation with (6) and the *modify* relation with (7)), so those utterances must be skipped.

Rosé et al.'s graph structured stack is designed to handle the more complex structure depicted in Figure 4c. We will return to this structure later in Section 8.1, when the adequacy of our focus model is analyzed.

Once the candidate antecedents are determined, various criteria can be used to choose among them. Syntactic and semantic constraints are common.

4.3.1 Our Focus Model for Temporal Reference Resolution

As mentioned earlier, our algorithm for temporal reference resolution is recency based. Specifically, the focus model is structured as a linear list of all times mentioned so far in the current dialog. The list is ordered by recency, and no entries are deleted from the list.

The candidate antecedents are as follows. For each type of anaphoric relation, the most recent Temporal Unit on the focus list that satisfies that relation, if there is one, is a candidate antecedent.

The antecedent is chosen from among the candidate antecedents based on a combined score reflecting a priori preferences for the type of anaphoric relation established, how recently the time was mentioned, and how plausible the resulting temporal interpretation would be (see Section 5). These numerical heuristics contribute to some extent to the success of the implementation, but are not critical components of the model, as shown in Section 8.3.

4.4 The Need for Explicit Identification of Relations

As mentioned in the introduction, one goal of this work is to assess the adequacy of a recency-based focus model for this task and genre. To be well founded, such evaluations must be made with respect to a particular set of relations. For example, the *modify* relation supports a recency-based approach. Consider the following example, reproduced from Section 4.2:

Utterance	Interpretation
(1) Monday looks good.	*Assume: Monday 19 August*
(2) How about 2?	*(co-reference relation) 2pm, Monday 19 August*
(3) Hmm, how about 4?	*(modify relation) 4pm, Monday 19 August*

Because our model includes the *modify* anaphoric relation, the Temporal Unit in (2) is an appropriate antecedent for the one in (3). A model without this relation might require (3)'s antecedent to be provided by (1).

5. Algorithm

This section presents our high-level algorithm for temporal reference resolution. After an overview in Section 5.1, the rule application architecture is described in Section 5.2, and the main rules composing the algorithm are given in Section 5.3. The complete set of rules is given in detail in Online Appendix 1.

5.1 Overview

An important feature of our approach is that the system is forced to choose among possibilities only if the resulting interpretations would be inconsistent. If the results for two possibilities are consistent, the system merges the results together.

At a high level, the algorithm operates as follows. There is a set of rules for each of the relations presented in Section 4.2. The rules include constraints involving the current utterance and another Temporal Unit. In the anaphoric cases, the other Temporal Unit is a potential antecedent from the focus list. In the deictic cases, it is the dialog date or a

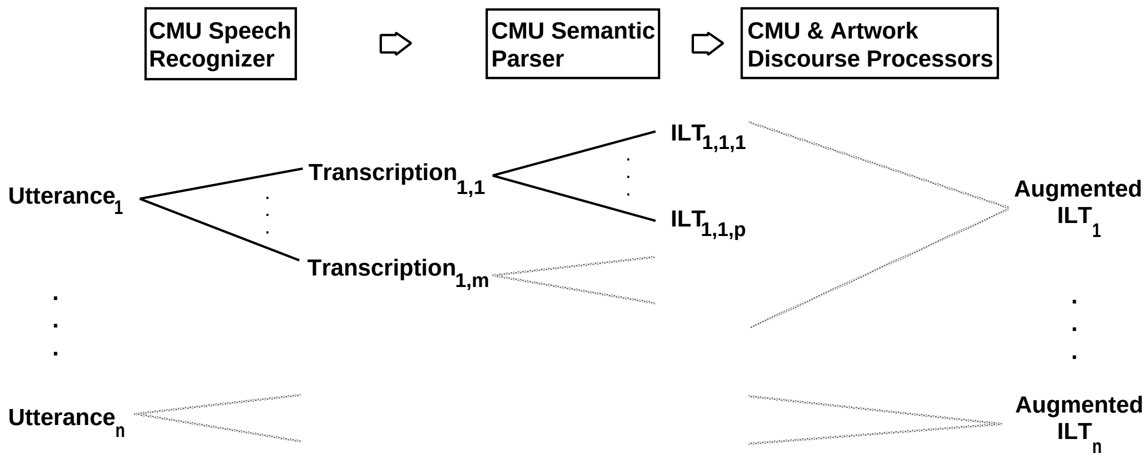

Figure 5: The Enthusiast System

later time. For the current temporal expression to be resolved, each rule is applied. For the anaphoric rules, the antecedent considered is the most recent one satisfying the constraints. All consistent maximal mergings of the results are formed, and the one with the highest score is the chosen interpretation.

5.2 Architecture

Our system was developed to be integrated into the Enthusiast system developed at Carnegie Mellon University (see Qu, Eugenio, Lavie, Levin, & Rosé, 1996; Levin et al., 1995; Rosé et al., 1995; Lavie & Tomita, 1993). Enthusiast is a speech-to-speech machine translation system from Spanish into English. The aspects of the system needed for this paper are shown in Figure 5. The system processes all the utterances of a single speaker turn together (utterances 1 through n in the figure). Each spoken Spanish utterance is input to the speech recognizer, which produces one or more transcriptions of the utterance. The output of the speech recognition system is the input to a semantic parser (Lavie & Tomita, 1993; Levin et al., 1995), which produces a representation of the literal meaning of the sentence. This representation is called an *Interlingual Text* (*ILT*). The output of the semantic parser is ambiguous, consisting of multiple ILT representations of the input transcription. All of the ILT representations produced for an utterance are input to the discourse processor, which produces the final, unambiguous representation of that utterance. This representation is called an *augmented ILT*.

The discourse processor can be configured to be our system alone, a plan-based discourse processor developed at CMU (Rosé et al., 1995), or the two working together in integrated mode. The main results, presented in Tables 2 and 3 in Section 6, are for our system working alone, taking as input the ambiguous output of the semantic parser. For the CMU dialogs, the input to the semantic parser is the output of the speech recognition system. The NMSU dialogs were input to the semantic parser directly in the form of transcriptions.[3]

To produce one ILT, the semantic parser maps the main event and its participants into one of a small set of case frames (for example, a *meet* frame or an *is busy* frame). It also produces a surface representation of the temporal information in the utterance, which mirrors the form of the input utterance. Although the events and states discussed in the NMSU data are often outside the coverage of this parser, the temporal information generally is not. Thus, the parser provides a sufficient input representation for our purposes on both sets of data.

As the Enthusiast system is configured, the input is presented to our discourse processor in the form of alternative sequences of ILTs. Each sequence contains one ILT for each utterance. For example, using the notation in Figure 5, a sequence might consist of $ILT_{1,2,3}$, $ILT_{2,1,1}$, ..., $ILT_{n,2,1}$. Our system resolves the ambiguity in batches. Specifically, it produces a sequence of Augmented ILTs for each input sequence, and then chooses the best sequence as its final interpretation of the corresponding utterances. In this way, the input ambiguity is resolved as a function of finding the best temporal interpretations of the utterance sequences in context (as suggested by Qu et al., 1996). However, the number of alternative sequences of ILTs for a set of utterances can be prohibitively large for our system. The total number of sequences considered by the system is limited to the top 125, where the sequences are ordered using statistical rankings provided by the Enthusiast system.

Our method for performing semantic disambiguation is appropriate for this project, because the focus is on temporal reference resolution and not on semantic disambiguation. However, much semantic ambiguity cannot be resolved on the basis of the temporal discourse context alone, so this represents a potential area for improvement in the system performance results presented in Section 6. In fact, the Enthusiast researchers have already developed better techniques for resolving the semantic ambiguity in these dialogs (Shum et al., 1994).

Because the ILT representation was designed to support various projects in discourse, semantic interpretation, and machine translation, the representation produced by the semantic parser is much richer than is required for our temporal reference resolution algorithm. We recommend that others who implement our algorithm for their application build an input parser to produce only the necessary temporal information. The specification of our input is available in Online Appendix 2.

As described in Section 4.3, a focus list records the Temporal Units that have been discussed so far in the dialog. After a final Augmented ILT has been created for the current utterance, the Augmented ILT and the utterance are placed together on the focus list. In the case of utterances that specify more than one Temporal Unit, a separate entity is added for each to the focus list, in order of mention. Otherwise, the system architecture is similar to a standard production system, with one major exception: rather than choosing the results of just one of the rules that fires, multiple results can be merged. This is a flexible architecture that accommodates sets of rules targeting different aspects of the interpretation.

3. The semantic parser but not the speech recognizer was available for us to process the NMSU data. Presumably, the speech recognizer would not perform as well on the NMSU dialogs as it does on the CMU dialogs, since it was trained on the latter.

Following are the basic steps in processing a single ILT.

Step 1. The input ILT is *normalized*. In producing the ILTs that serve as input to our system, the semantic parser often represents pieces of information about the same time separately, mirroring the surface form of the utterance. This is done in order to capture relationships, such as topic-comment relationships, among clauses. Our system needs to know which pieces of information are about the same time, but does not need to know about the additional relationships. Thus, the system maps the input representation into a normalized form, to shield the reasoning component from the idiosyncracies of the input representation. A specification of the normalized form is given in Online Appendix 2.

The goal of the normalization process is to produce one Temporal Unit per distinct time specified in the utterance. The normalization program is quite detailed (since it must account for the various structures possible in the CMU input ILT), but the core strategy is straightforward: it merges information provided by separate noun phrases into one Temporal Unit, if it is consistent to do so. Thus, new Temporal Units are created only if necessary. Interestingly, few errors result from this process. Following are some examples.

I can meet Wednesday or Thursday.	*Represented as two disjoint TUs.*
I can meet from 2:00 until 4:00 on the 14th.	*Represented as one TU.*
I can meet Thursday the 11th of August.	*Represented as one TU.*

After the normalization process, highly accurate, obvious inferences are made and added to the representation.

Step 2. All of the rules are applied to the normalized input. The result of a rule application is a *Partial Augmented ILT*—information this rule will contribute to the interpretation of the utterance, if it is chosen. This information includes a certainty factor representing an a priori preference for the type of anaphoric or deictic relation being established. In the case of anaphoric relations, this factor is adjusted by a term representing how far back on the focus list the antecedent is (in the anaphoric rules in Section 5.3, the adjustment is represented by *distance factor* in the calculation of the certainty factor *CF*). The result of this step is the set of Partial Augmented ILTs produced by the rules that fired (i.e., those that succeeded).

In the case of multiple Temporal Units in the input ILT, each rule is applied as follows. If the rule does not access the focus list, the rule is applied to each Temporal Unit. A list of Partial Augmented ILTs is produced, containing one entry for each successful match, retaining the order of the Temporal Units in the original input. If the rule does access the focus list, the process is the same, but with one important difference. The rule is applied to the first Temporal Unit. If it is successful, then the same focus list entity used to apply the rule to this Temporal Unit is used to interpret the remaining Temporal Units in the list. Thus, all the anaphoric temporal references in a single utterance are understood with respect to the same focus list element. So, for example, the anaphoric interpretations of the temporal expressions in "I can meet Monday or Tuesday" both have to be understood with respect to the same entity in the focus list.

When accessing entities on the focus list, an entry for an utterance that specifies multiple Temporal Units may be encountered. In this case, the Temporal Units are simply accessed in order of mention (from most to least recent).

Step 3. All maximal mergings of the Partial Augmented ILTs are created. Consider a graph in which the Partial Augmented ILTs are the vertices, and there is an edge between two Partial Augmented ILTs if they are compatible. Then, the maximal cliques of the graph (i.e., the maximal complete subgraphs) correspond to the maximal mergings. Each maximal merging is then merged with the normalized input ILT, resulting in a set of Augmented ILTs.

Step 4. The Augmented ILT chosen is the one with the highest certainty factor. The certainty factor of an Augmented ILT is calculated as follows. First, the certainty factors of the constituent Partial Augmented ILTs are summed. Then, critics are applied to the resulting Augmented ILT, lowering the certainty factor if the information is judged to be incompatible with the dialog state.

The merging process might have yielded additional opportunities for making obvious inferences, so this process is performed again, to produce the final Augmented ILT.

To process the alternative input sequences, a separate invocation to the core system is made for each sequence, with the sequence of ILTs and the current focus list as input. The result of each call is a sequence of Augmented ILTs, which are the system's best interpretations of the input ILTs, and a new focus list, representing the updated discourse context corresponding to that sequence of interpretations. The system assigns a certainty factor to each sequence of Augmented ILTs, specifically, the sum of the certainty factors of the constituents. It chooses the sequence with the highest certainty factor, and updates the focus list to the focus list calculated for that sequence.

5.3 Temporal Reference Resolution Rules

Figure 6 presents the main temporal resolution rules, one for each of the cases described in Sections 4.1 and 4.2. In the complete set of rules, given in Online Appendix 1, many are broken down into subcases involving, for example, the end times or starting times.

The rules apply to individual Temporal Units. They return a certainty factor, and either a more fully specified Temporal Unit or an empty structure indicating failure.

Many of the rules calculate temporal information with respect to a frame of reference, using a separate calendar utility. Following are functions and conventions used in Figure 6.

1. **next**($TimeValue$, RF): returns the next $timeValue$ that follows reference frame RF. For example, next(Monday, [...Friday, 19th,...]) = Monday, 22nd.

2. **resolve_deictic**(DT, RF): resolves the deictic term DT with respect to the reference frame RF.

3. **merge**(TU_1, TU_2): if Temporal Units TU_1 and TU_2 contain no conflicting fields, returns a Temporal Unit containing all of the information in the two units; otherwise returns $\{\}$.

4. **merge_upper**(TU_1, TU_2): similar to the previous function, except that the only fields from TU_1 that are included are those that are of the same or less specificity as the most specific field in TU_2.

5. **specificity**(TU): returns the specificity of the most specific field in TU.

6. **most_specific**(TU): returns the most specific field in TU.

7. **starting_fields**(TU): returns a list of starting field names for those in TU having non-null values.

8. **structure→component:** returns the named component of the structure.

9. **conventions:** Values are in **bold face** and variables are in *italics*. TU is the current Temporal Unit being resolved. $TodaysDate$ is a representation of the dialog date. $FocusList$ is the list of discourse entities from all previous utterances.

The algorithm does not cover some subcases of relations concerning the end times. For instance, rule D-frame-of-reference covers only the starting-time case of the *frame of reference* deictic relation. An example of an end-time case that is not handled is the utterance "Let's meet until Thursday," under the meaning that they should meet from today through Thursday. This is an area for future work.

6. Results

As mentioned in Section 3, the main results are based on comparisons against human annotation of the held out test data. The results are based on straight field-by-field comparisons of the Temporal Unit representations introduced in Section 3. To be considered correct, information must not only be right, but it also has to be in the right place. Thus, for example, "Monday" correctly resolved to *Monday 19 August*, but incorrectly treated as a starting rather than an end time, contributes 3 errors of omission and 3 errors of commission (and receives no credit for the correct date).

Detailed results for the test sets are presented in this section, starting with results for the CMU data (see Table 2). *Accuracy* measures the extent to which the system produces the correct answer, while *precision* measures the extent to which the system's answers are correct (see the formulas in Table 2). For each component of the extracted temporal structure, the system's correct and incorrect answers were counted. Since null values occur quite often, these counts exclude cases in which the system's answer, the correct answer, or both answers are null. Those cases were counted separately. Note that each test set contains three complete dialogs with an average of 72 utterances per dialog.

These results show that the system achieves an overall accuracy of 81%, which is significantly better than the baseline accuracy (defined below) of 43%. In addition, the results show a high precision of 92%. In some of the individual cases, however, the results could be higher due to several factors. For example, our system development was inevitably focused more on some fields than others. An obvious area for improvement is the system's processing of the *time of day* fields. Also, note that the values in the *Mis* column are higher than those in the *Ext* column. This reflects the conservative coding convention, mentioned in Section 3, for filling in unspecified end points.

The accuracy and precision figures for the *hour & minute* and *time of day* fields are very high because a large proportion of them are null. We include null correct answers in our

Rules for deictic relations

Rule D-simple: All cases of the *simple* deictic relation.
if there is a deictic term, DT, in TU then
    return $\langle$0.9, merge(TU, resolve_deictic(DT, $TodaysDate$))$\rangle$

Rule D-frame-of-reference: The starting-time cases of the *frame of reference* deictic relation.
if (most_specific(starting_fields(TU)) $<$ **time_of_day**) then
    Let f be the most specific field in starting_fields(TU)
    return $\langle$0.4, merge(TU, next($TU \rightarrow f$, $TodaysDate$))$\rangle$

Rules for anaphoric relations

Rule A-co-reference: All cases of the *co-reference* anaphoric relation.
for each non-empty Temporal Unit TU_{fl} from $FocusList$ (starting with most recent)
    if specificity(TU_{fl}) $\leq$ specificity(TU) and not empty merge(TU_{fl}, TU) then
        $CF = 0.8 -$ distance_factor(TU_{fl}, $FocusList$)
        return $\langle CF$, merge(TU_{fl}, TU)$\rangle$

Rule A-less-specific: All cases of the *less-specific* anaphoric relation.
for each non-empty Temporal Unit TU_{fl} from $FocusList$ (starting with most recent)
    if specificity(TU_{fl}) $>$ specificity(TU) and not empty merge_upper(TU_{fl}, TU) then
        $CF = 0.5 -$ distance_factor(TU_{fl}, $FocusList$)
        return $\langle CF$, merge_upper(TU_{fl}, TU)$\rangle$

Rule A-frame-of-reference: Starting-time case of the *frame of reference* anaphoric relation.
if (most_specific(starting_fields(TU)) $<$ **time_of_day**) then
    for each non-empty Temporal Unit TU_{fl} from $FocusList$ (starting with most recent)
        if specificity(TU) $\geq$ specificity(TU_{fl}) then
            Let f be the most specific field in starting_fields(TU)
            $CF = 0.6 -$ distance_factor(TU_{fl}, $FocusList$)
            return $\langle CF$, merge(TU, next($TU \rightarrow f$, $TU_{fl} \rightarrow$start_date))$\rangle$

Rule A-modify: All cases of the *modify* anaphoric relation.
if (specificity(TU) $\geq$ **time_of_day**) then
    for each non-empty Temporal Unit TU_{fl} from $FocusList$ (starting with most recent)
        if specificity(TU) $\geq$ specificity(TU_{fl}) and specificity(TU_{fl}) $\geq$ **time_of_day** then
            if not empty merge_upper(TU_{fl}, TU) then
                $CF = 0.5 -$ distance_factor(TU_{fl}, $FocusList$)
                return $\langle CF$, merge_upper(TU_{fl}, TU)$\rangle$

Figure 6: Main Temporal Resolution Rules

Label	Cor	Inc	Mis	Ext	Nul	Poss	Act	BaseAcc	**Acc**	**Prec**
start										
Month	49	3	7	3	0	59	55	0.338	0.831	0.891
Date	48	4	7	3	0	59	55	0.403	0.814	0.873
DayofWeek	46	6	7	3	0	59	55	0.242	0.780	0.836
HourMin	18	0	7	0	37	62	55	0.859	0.887	1.000
TimeDay	9	0	18	0	35	62	44	0.615	0.710	1.000
end										
Month	48	3	7	1	3	61	55	0.077	0.836	0.927
Date	47	5	6	3	1	59	56	0.048	0.814	0.857
DayofWeek	45	7	6	3	1	59	56	0.077	0.780	0.821
HourMin	9	0	9	0	44	62	53	0.862	0.855	1.000
TimeDay	4	0	13	1	44	61	49	0.738	0.787	0.980
Overall	323	28	87	17	165	534	604	0.428	**0.809**	**0.916**

Legend

Cor(rect):	System and key agree on non-null value
Inc(orrect):	System and key differ on non-null value
Mis(sing):	System has null value for non-null key
Ext(ra):	System has non-null value for null key
Nul(l):	Both system and key give null answer
Poss(ible):	Correct + Incorrect + Missing + Null
Act(ual):	Correct + Incorrect + Extra + Null
Base(line)Acc(uracy):	Baseline accuracy (input used as is)
Acc(uracy):	% Key values matched correctly ((Correct + Null)/Possible)
Prec(ision):	% System answers matching the key ((Correct + Null)/Actual)

Table 2: Evaluation of System on CMU Test Data

Label	Cor	Inc	Mis	Ext	Nul	Poss	Act	BaseAcc	**Acc**	**Prec**
start										
TimeDay	9	0	18	0	35	62	44	0.615	0.710	1.000
Month	55	0	23	5	3	63	81	0.060	0.716	0.921
Date	49	6	23	5	3	63	81	0.060	0.642	0.825
DayofWeek	52	3	23	5	3	63	81	0.085	0.679	0.873
HourMin	34	3	7	6	36	79	80	0.852	0.875	0.886
TimeDay	18	8	31	2	27	55	84	0.354	0.536	0.818
end										
Month	55	0	23	5	3	63	81	0.060	0.716	0.921
Date	49	6	23	5	3	63	81	0.060	0.642	0.825
DayofWeek	52	3	23	5	3	63	81	0.060	0.679	0.873
HourMin	28	2	13	1	42	73	85	0.795	0.824	0.959
TimeDay	9	2	32	5	38	54	81	0.482	0.580	0.870
Overall	401	33	221	44	161	639	816	0.286	**0.689**	**0.879**

Table 3: Evaluation of System on NMSU Test Data

figures because such answers often reflect valid decisions not to fill in explicit values from previous Temporal Units.

Table 3 contains the results for the system on the NMSU data. It shows that the system performs respectably, with 69% accuracy and 88% precision, on the more complex set of data. The precision is still comparable, but the accuracy is lower, since more of the entries are left unspecified (that is, the figures in the *Mis* column in Table 3 are higher than in Table 2). Furthermore, the baseline accuracy (29%) is almost 15% lower than the one for the CMU data (43%), supporting the claim that this data set is more challenging.

The baseline accuracies for the test data sets are shown in Table 4. These values were derived by disabling all the rules and evaluating the input itself (after performing normalization, so that the evaluation software could be applied). Since null values are the most frequent for all fields, this is equivalent to using a naive algorithm that selects the most frequent value for each field. Note that in Tables 2 and 3, the baseline accuracies for the end *month*, *date*, and *day of week* fields are quite low because the coding convention calls for filling in these fields, even though they are not usually explicitly specified. In this case, an alternative baseline would have been to use the corresponding starting field. This has not been calculated, but the results can be approximated by using the baseline figures for the starting fields.

The rightmost column of Table 4 shows that there is a small amount of error in the input representation. This figure is 1 minus the precision of the input representation (after normalization). Note, however, that this is a close but not exact measure of the error in the input, because there are a few cases of the normalization process committing errors and a few of it correcting errors. Recall that the input is ambiguous; the figures in Table 4 are based on the system selecting the first ILT in each case. Since the parser orders the

Set	Cor	Inc	Mis	Ext	Nul	Act	Poss	**Acc**	Input Error
cmu	84	6	360	10	190	290	640	**0.428**	0.055
nmsu	65	3	587	4	171	243	826	**0.286**	0.029

Table 4: Baseline Figures for Both Test Sets

seen/ unseen	cmu/ nmsu	ambiguous, uncorrected/ unambiguous, partially corrected	Dialogs	Utterances	**Acc**	**Prec**
seen	cmu	ambiguous, uncorrected	12	659	0.883	0.918
seen	cmu	unambiguous, partially corrected	12	659	0.914	0.957
unseen	cmu	ambiguous, uncorrected	3	193	0.809	0.916
seen	nmsu	ambiguous, uncorrected	4	358	0.679	0.746
seen	nmsu	unambiguous, partially corrected	4	358	0.779	0.850
unseen	nmsu	ambiguous, uncorrected	3	236	0.689	0.879

Table 5: Overall Results

ILTs based on a measure of acceptability, this choice is likely to have the relevant temporal information.

The above results are for the system taking ambiguous semantic representations as input. To help isolate errors due to our model, the system was also evaluated on unambiguous, partially corrected input for all the seen data (the test sets were retained as unseen test data). The input is only partially corrected because some errors are not feasible to correct manually, given the complexity of the input representation.

The overall results are shown in the Table 5. The table includes the results presented earlier in Tables 2 and 3, to facilitate comparison. In the CMU data set, there are twelve dialogs in the training data and three dialogs in a held out test set. The average length of each dialog is approximately 65 utterances. In the NMSU data set, there are four training dialogs and three test dialogs.

In both data sets, there are noticeable gains in performance on the seen data going from ambiguous to unambiguous input, especially for the NMSU data. Therefore, the semantic ambiguity and input errors contribute significantly to the system's errors.

Some challenging characteristics of the seen NMSU data are vast semantic ambiguity, numbers mistaken by the input parser for dates (for example, phone numbers are treated as dates), and the occurrences of subdialogs.

Most of the the system's errors on the unambiguous data are due to parser error, errors in applying the rules, errors in mistaking anaphoric references for deictic references (and vice versa), and errors in choosing the wrong anaphoric relation. As will be shown in Section 8.1, our approach handles focus effectively, so few errors can be attributed to the wrong entities being in focus.

7. Other Work on Temporal Reference Resolution

To our knowledge, there are no other published results on unseen test data of systems performing similar temporal reference resolution tasks. Rosé et al. (1995, Enthusiast), Alexandersson et al. (1997, Verbmobil), and Busemann et al. (1997, Cosma) describe other recent natural language processing systems that resolve temporal expressions in scheduling dialogs. Rosé et al. also address focus issues; we compare our work to theirs in detail in Section 8.1. All of the systems share certain features, such as the use of a calendar utility to calculate dates, a specificity ordering of temporal components (such as in Figure 3), and a record of the temporal context.

However, all of the other systems perform temporal reference resolution as part of their overall processing, in service of solving another problem such as speech act resolution. None of them lays out a detailed approach or model for temporal reference resolution, and none gives results of system performance on any temporal interpretation tasks.

Kamp and Reyle (1993) address representational and processing issues in the interpretation of temporal expressions. However, they do not implement their ideas or present the results of a working system. They do not attempt coverage of a data set, or present a comprehensive set of relations, as we do, but consider only specific cases that are interesting for their Discourse Representation Theory. In addition, they do not address the issues of discourse structure and attentional state focused on here. For example, they recognize that references such as "on Sunday" may have to be understood with respect to a frame of reference. But they do not address how the frame of reference is chosen in context, so do not address the question of what type of focus model is required.

Note that temporal reference resolution is a different problem from tense and aspect interpretation in discourse (as addressed in, for example, Webber, 1988; Song & Cohen, 1991; Hwang & Schubert, 1992; Lascarides, Asher, & Oberlander, 1992; Kameyama, Passonneau, & Poesio, 1993). These tasks are briefly reviewed here to clarify the differences. Temporal reference resolution is determining what time is being explicitly specified by noun phrases that are temporal referring expressions (e.g., "Monday" resolved to *Monday 19 August*). Tense and aspect interpretation involves determining implicit information about the states and events specified by verb phrases (e.g., that the kissing event specified in "He had kissed her" happened before some reference time in the past). While it could aid in performing temporal reference resolution, we are not addressing tense and aspect interpretation itself.

Scheduling dialogs, or scheduling subdialogs of other kinds of dialogs, predominantly employ the present and future tenses, due to the nature of the task. As discussed further below in Section 8.1, a primary way that tracking the tense and aspect would aid in temporal reference resolution would be to recognize discourse segments that depart from the scheduling dialog or subdialog. In addition, Kamp and Reyle (1993) address some cases in which tense and aspect, temporal nouns, and temporal adverbs interact to affect the temporal interpretation. We intend to pursue these ideas in future work.

8. Analysis

The implementation is an important proof of concept. However, as discussed in Section 6, various kinds of errors are reflected in the results; many are not directly related to discourse

	# TUs	# TUs specified anaphorically
CMU	196	167
NMSU	96	71
Total	292	238

Figure 7: Counts of Temporal Unit References in the Training Data

processing or temporal reference resolution. Examples are (1) completely null inputs, when the semantic parser or speech recognizer fails, (2) numbers mistaken as dates, and (3) failures to recognize that a relation can be established, due to a lack of specific domain knowledge.

To evaluate the algorithm itself, in this section we separately evaluate the components of our method for temporal reference resolution. Sections 8.1 and 8.2 assess the key contributions of this work: the focus model (in Section 8.1) and the deictic and anaphoric relations (in Section 8.2). These evaluations required us to perform extensive additional manual annotation of the data. In order to preserve the test dialogs as unseen test data, these annotations were performed on the training data only. In Section 8.3, we isolate the architectural components of our algorithm, such as the certainty factor calculation and the critics, to assess the effects they have on performance.

8.1 Evaluation of the Focus Model

The algorithm presented here does not include a mechanism for recognizing the global structure of the discourse, such as in the work of Grosz and Sidner (1986), Mann and Thompson (1988), Allen and Perrault (1980), and in descendent work. Recently in the literature, Walker (1996) argues for a more linear-recency based model of attentional state (though not that discourse structure need not be recognized), while Rosé et al. (1995) argue for a more complex model of attentional state than is represented in most current computational theories of discourse.

Many theories that address how attentional state should be modeled have the goal of performing intention recognition as well. We investigate performing temporal reference resolution directly, without also attempting to recognize discourse structure or intentions. We assess the challenges the data present to our model when only this task is attempted.

The total number of Temporal Units and the number of them specified by anaphoric noun phrases in the two training data sets are given in Figure 7.[4] There are different units that could be counted, from the number of temporal noun phrases to the number of distinct times referred to in the dialog. Here, we count the entities that must be resolved by a temporal reference resolution algorithm, i.e., the number of distinct temporal units specified in each sentence, summed over all sentences. Operationally, this is a count of Temporal Units after the normalization phase, i.e., after Step 1 in Section 5.2. This is the unit considered in the remainder of this paper.

4. The anaphoric counts include the cases in which both deictic and anaphoric interpretations yield the correct result.

To support the evaluation presented in this section, antecedent information was manually annotated in the training data. For each Temporal Unit specified by an anaphoric noun phrase, all of the antecedents that yield the correct interpretation under one of the anaphoric relations were identified, except that, if both TU_i and TU_j are appropriate antecedents, and one is an antecedent of the other, only the more recent one is included. Thus, only the heads of the anaphoric chains existing at that point in the dialog are included. In addition, *competitor* discourse entities were also identified, i.e., previously mentioned Temporal Units for which some relation could be established, but the resulting interpretation would be incorrect. Again, only Temporal Units at the head of an anaphoric chain were considered. To illustrate these annotations, Figure 8 shows a graph depicting anaphoric chain annotations of an NMSU dialog (dialog 9). In the figure, solid lines link the correct antecedents, dotted lines show competitors, and edges to nowhere indicate deictics.

8.1.1 Cases in which the immediately preceding time is not an appropriate antecedent

The main purpose of a focus model is to make an appropriate set of discourse entities available as candidate antecedents at each point in the discourse. As described above in Section 4.3, Grosz and Sidner's model captures situations in which entities should not be available as candidate antecedents, and Rosé et al. identify situations in which Grosz and Sidner's model may incorrectly eliminate entities from consideration (i.e., dialogs with multiple threads). The potential challenge for a recency-based model like ours is that entities may be available as candidate antecedents that should not be. An entity E may occur to which an anaphoric relation could be established, but an entity mentioned before E is needed for the correct interpretation. (From another perspective, E yields the wrong interpretation but cannot be ruled out as a possible antecedent.) To assess the magnitude of this problem for our method, in this section we characterize the cases in which the most recent entity is not an appropriate antecedent.

Before proceeding, we note that there is only one situation in which our model incorrectly makes a needed entity *un*available. Recall from Section 4.3 that, for a particular relation R, only the most recent Temporal Unit for which R can be established is a candidate (call it C). The problem arises when the correct interpretation requires that that same relation R be established with an entity mentioned earlier than C. This is a problem because the earlier time is not a candidate. If such cases were to occur in the training data, they would have been found by the analysis presented below. However, none were found.

Based on the anaphoric chain annotations, we identified how far back on the focus list one must go to find an antecedent that is appropriate according to the model. An antecedent is considered to be appropriate according to the model if there exists a relation defined in the model such that, when established between the current utterance and the antecedent, it yields the correct interpretation. Note that we allow antecedents for which the anaphoric relation would be a trivial extension of one of the relations explicitly defined in the model. For example, phrases such as "after lunch" should be treated as if they are simple times of day under the *co-reference* and *modify* anaphoric relations, but, as explicitly defined, those relations do not cover such phrases. For example, given *Wednesday 14 April*, the reference "after lunch" should be interpreted as *after lunch, Wednesday 14 April* under the

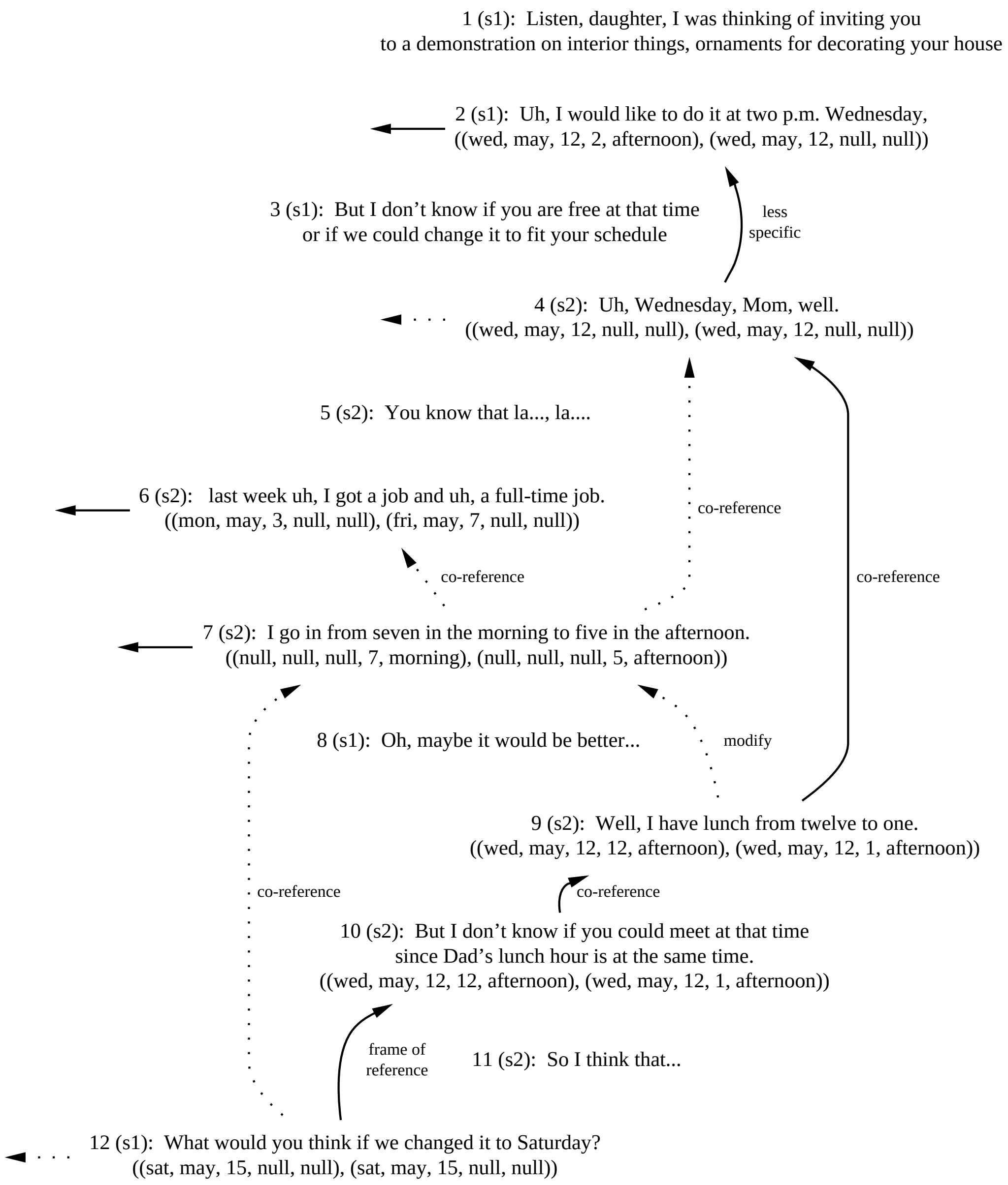

Figure 8: Anaphoric Annotations of Part of NMSU Dialog 9.

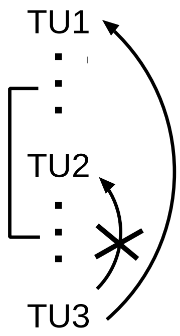

Figure 9: Structure Challenging the Recency Model.

co-reference relation. Similarly, given *10am, Wednesday, 14 April*, "After lunch" in "After lunch would be better" should be interpreted as *after lunch, Wednesday 14 April* under the *modify* anaphoric relation.

The results are striking. Between the two sets of training data, there are only nine anaphoric temporal references for which the immediately preceding Temporal Unit is not an appropriate antecedent, $3/167 = 1.8\%$ in the CMU data, and $6/71 = 8.4\%$ in the NMSU data.

Figure 9 depicts the structure involved in all nine cases. TU_3 represents the anaphoric reference for which the immediately preceding Temporal Unit is not an appropriate antecedent. TU_1 represents the most recent appropriate antecedent, and TU_2 represents the intervening Temporal Unit or Units. The ellipses represent any intervening non-temporal utterances.

Figure 10 characterizes the nine cases along a number of dimensions. To isolate the issues addressed, it was assumed in deriving these figures that the dialog is correctly interpreted up to and including TU_1.

In three of the cases (rows 2, 4, and 9, labeled 07-63, 08-57, 10-55, respectively), there is a correct deictic interpretation of TU_3 under our model, in addition to the correct (with antecedent TU_1) and incorrect (with antecedent TU_2) anaphoric interpretations.

Column 1 of Figure 10 shows that, in all three cases in the CMU data and in two cases in the NMSU data, the second most recently mentioned Temporal Unit is an appropriate antecedent. In the remaining four cases, the third most recently mentioned time is appropriate.

In three of the cases, the references represented by TU_2 in Figure 9 are in subdialogs off the main topic and scheduling task (indicated as "Yes" in column 2). All of these subdialogs are in the NMSU data. In four cases, the TU_2 references are in subsegments that are directly in service of the main task (indicated as "No" in column 2), and in two cases, we judged them to be borderline.

Column 3 characterizes the type of reference the TU_2 references are. The two marked "Anaphoric, main task" are specific references to times that involve the main scheduling

	1	2	3	4	5	6
	Distance to most recent appropriate antecedent	Subdialog?	Type of TU_2	TU_2 Correct?	TU_2 a Competitor?	Potential Cumulative Errors
1 (07-37) CMU	2	No	Anaphoric, main task	Yes	Yes	21
2 (07-63) CMU	2	No	Habitual	No	Yes	0
3 (15-31) CMU	2	No	Anaphoric, main task	Yes	Yes	4
4 (08-57) NMSU	2	Yes	Reference outside dialog	No	Yes	2 minor
5 (08-66) NMSU	3	Yes	1 deictic 1 habitual	Yes No	Yes Yes	10 (worst case)
6 (09-39) NMSU	2	No	habitual	No	No	0
7 (09-09) NMSU	3	Yes	1 deictic 1 habitual	Yes No	Yes	4 (worst case)
8 (09-45) NMSU	3	Borderline	both habitual	No	Yes	6
9 (10-55) NMSU	3	Borderline	both habitual	No	Yes	3

Figure 10: Summary of Cases in Which Most Recent TU is not an Appropriate Antecedent

	Dialog Date: Monday 10 May
TU_1:	It's just that ... this Thursday *[Thursday May 13]* is our second wedding anniversary and I don't know what to do.
	⟨ 31 non-temporal utterances about what to cook ⟩
	Did you go with my mother?
TU_2:	With my mother? Yes. I went at around six in the morning.
	Did you and Maura go for a walk?
	No, no we didn't.
	Hmmmmm. We got lazy.
	Ah Claudia.
TU_3	Well, yes. Listen Lily. What do you think if we see each other on, on Thursday at six and I, at six?

Figure 11: Dialog Segment of the Case in Row 4 in Figure 10

task. The subdialog marked "Reference outside dialog" (row 4, label 8-57) is shown in Figure 11.

The main topic of this dialog is a party for the anniversary mentioned in TU_1. The TU_2 reference, "around six in the morning," involves the participants' shared knowledge of an event that is not related to the scheduling task. The only interpretation possible in our model is six in the morning on the day specified in the TU_1 reference, while in fact the participants are referring to six in the morning on the dialog date. (There is currently no coverage in our model for deictic references that mention only a time of day.) Thus, the interpretation of the TU_2 reference is incorrect, as indicated in column 4.

Many of the TU_2 references are habitual (marked "habitual" in column 3 of Figure 10). For example, the participants discuss their usual work schedules, using utterances such as "during the week I work from 3 to 6." Since there is no coverage of habituals in our model, the interpretations of all of the TU_2 habitual references are incorrect, as indicated in column 4.

We now turn to column 5, which asks a key question: is TU_2 a competitor? TU_2 is a competitor if there is some relation in the model that can be established between TU_3 and TU_2. In the cases in which TU_2 represents multiple utterances (namely, the fifth, seventh, eighth, and ninth rows of Figure 10), "yes" is indicated in column 5 if an interpretation of the segment involving both of the TU_2 references is possible. Cumulative error (column 6) can be non-zero only if the entry in column 5 is "Yes": if the TU_2 references are not competitors, they cannot be antecedents under our model, so they cannot prevent TU_3 from being recognized as a correct antecedent.

It is important to note that the incorrect interpretation of TU_3 and the cumulative errors indicated in column 6 are only potential errors. In all cases in Figure 10, the correct interpretation of TU_3 involving TU_1 is available as a possible interpretation. What is shown in column 6 is the number of cumulative errors that would result if an interpretation involving TU_2 were chosen over a correct interpretation involving TU_1. In many cases, the system's answer is correct because the (correct) TU_3–TU_1 interpretation involves the *co-reference*

Correct Interpretation of the TU_1 reference: Monday 22nd November
TU_2: of December?
TU_3: of November.

Figure 12: Dialog Segment of the Case in Row 1 in Figure 10

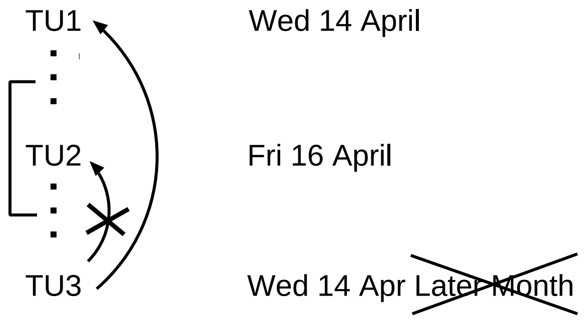

Figure 13: Structure of the Case in Row 3 of Figure 10

anaphoric relation, while the (incorrect) TU_3–TU_2 interpretation involves the *frame of reference* anaphoric relation; the certainty factor of the former is sufficiently larger than that of the latter to overcome the distance-factor penalty. In addition, such interpretations often involve large jumps forward in time, which are penalized by the critics.

The worst case of cumulative error, row 1, is an example. The segment is depicted in Figure 12. The incorrect interpretation involving TU_2 is November of the following year, calculated under the *frame of reference* anaphoric relation. The participants do not discuss the year, so the system cannot recover. Thus, a large amount of cumulative error would result if that interpretation were chosen.

The segment corresponding to row 3 is similar. Its structure is depicted in Figure 13. In this passage, two days are mentioned in sequence, *Wednesday 14 April* (the TU_1 reference) and *Friday 16 April* (the TU_2 reference). Then, the day mentioned first—*Wednesday 14 April*—is referred to again as "Wednesday the 14th" (the TU_3 reference). There is no relation in our model that enables the correct interpretation of TU_3 to be obtained from TU_2. If TU_2 were taken to be the antecedent of TU_3, the resulting incorrect interpretation would be the next possible *Wednesday 14*, in a later month (possibly in a later year), under the *frame of reference* anaphoric relation. What is required for the correct interpretation is the *co-reference* anaphoric relation to be established between TU_1 and TU_3. We saw exactly the same pattern above for the row 1 discourse segment, depicted in Figure 12, except that in that case a later month was calculated, rather than a later date.

It should be noted that, if times rather than days or months were being discussed, the correct interpretation for TU_3 could be obtained from TU_2 under the *modify* anaphoric relation. A good example of this occurs in the corpus example in Figure 1, repeated here

Temporal context: Tuesday 28 September			
	s1	1	On Thursday I can only meet after two pm
		2	From two to four
TU_1		3	Or two thirty to four thirty
TU_2		4	Or three to five
TU_3	s2	5	Then how does from two thirty to four thirty seem to you
		6	On Thursday
	s1	7	Thursday the thirtieth of September

Figure 14: Corpus Example from Figure 1

as Figure 14. The *modify* anaphoric relation enables TU_2 to be the antecedent of TU_3. The same would be true in the simpler case of "Two? Or Three? How about Two?". A promising future extension would be to develop a new *modify* anaphoric relation for these cases.

Returning to column 6 of Figure 10, note that two of the cumulative error figures are listed as "worst case." These are cases in which there are two TU_2 references and there are many different possible interpretations of the passage.

Notice that the second and fourth rows correspond to cases in which TU_2 is a competitor, yet no significant potential cumulative error results (the minor errors listed for row 4 are due to the relation not fitting exactly, rather than an error from choosing the wrong antecedent: *six in the morning* rather than *in the morning* is placed into the high specificity fields). In both of these cases, the error corrects itself: TU_1 is incorrectly taken to be the antecedent of TU_2, which is in turn incorrectly taken to be the antecedent of TU_3. But TU_2 in effect copies over the information from TU_1 that is needed to interpret TU_3. As a result, the interpretation of TU_3 is correct.

In the cases for which there are only a few potential cumulative errors, either a new, unambiguous time is soon introduced, or a time being discussed before the offending TU_2 reference is soon *re*introduced, getting things back on track.

An important discourse feature of the dialogs is the degree of redundancy of the times mentioned (Walker, 1996). This limits the ambiguity of the times specified, and it also leads to a higher level of robustness, since additional Temporal Units with the same time are placed on the focus list and previously mentioned times are reintroduced. Table 6 presents measures of redundancy. The redundancy is broken down into the case where redundant plus additional information is provided (*Redundant*) versus the case where the temporal information is just repeated (*Reiteration*). This shows that roughly 27% of the CMU utterances with temporal information contain redundant temporal references, while 20% of the NMSU ones do.

In considering how the model could be improved, in addition to adding a new *modify* anaphoric relation for cases such as those in Figures 12 and 13, habituals are clearly an area for investigation. Many of the offending references are habitual, and all but one of the subdialogs and borderline subdialogs involve habituals. In a departure from the algorithm,

Dialog Set	Temporal Utterances	Redundant	Reiteration	%
cmu	210	36	20	26.7
nmsu	122	11	13	19.7

Table 6: Redundancy in the Training Dialogs

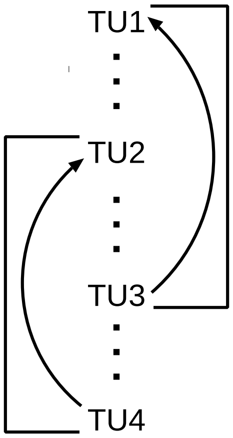

Figure 15: Temporal Multiple Thread Structure

the system uses a simple heuristic for ignoring subdialogs: a time is ignored if the utterance evoking it is in the simple past or past perfect. This prevents some of the potential errors and suggests that changes in tense, aspect, and modality are promising clues to explore for recognizing subsegments in this kind of data (see, for example, Grosz & Sidner, 1986; Nakhimovsky, 1988).

8.1.2 Multiple threads

Rosé et al. (1995, p. 31) describe dialogs composed of multiple threads as "negotiation dialogues in which multiple propositions are negotiated in parallel." According to Rosé et al., dialogs with such multiple threads pose challenges to a stack-based discourse model on both the intentional and attentional levels. They posit a more complex representation of attentional state to meet these challenges, and improve their results on speech act resolution in a corpus of scheduling dialogs by using their model of attentional state.[5]

As discussed above, in this work we address only the attentional level. The relevant structure for temporal reference resolution, abstracting from the examples given by Rosé et al., is shown in Figure 15. There are four Temporal Units mentioned in the order TU_1, TU_2, TU_3, and TU_4 (other times could be mentioned in between). The (attentional) multiple thread case is when TU_1 is required to be an antecedent of TU_3, but TU_2 is also needed to

5. They do not report how many multiple thread instances appear in their data.

Assumed Dialog Date: Friday 11 April		
(1)	S1:	We need to set up a schedule for the meeting.
(2)		How does your schedule look for next week?
(3)	S2:	Well, Monday and Tuesday both mornings are good.
(4)		Wednesday afternoon is good also.
(5)	S1:	It looks like it will have to be Thursday then.
(6)		Or Friday would also possibly work.
(7)		Do you have time between twelve and two on Thursday?
(8)		Or do you think sometime Friday afternoon you could meet?
(9)	S2:	No.
(10)		Thursday I have a class.
(11)		And Friday is really tight for me.
(12)		How is the next week?
(13)		If all else fails there is always video conferencing.
(14)	S1:	Monday, Tuesday, and Wednesday I am out of town.
(15)		But Thursday and Friday are both good.
(16)		How about Thursday at twelve?
(17)	S2:	Sounds good.
(18)		See you then.

Figure 16: Example of Deliberating Over A Meeting Time
(Rosé et al., 1995, p. 32)

interpret TU_4. There are no realizations of this structure, in terms of our model, in either the NMSU or CMU training data set.

The case represented by row three in Figure 10, whose structure is depicted above in 13, is the instance in our data that is most closely related to the situations addressed by Rosé et al. This is a type of structure that Grosz and Sidner's model addresses, but it is not a multiple thread case, since TU_2 is not needed to interpret a Temporal Unit mentioned after TU_3.

Rosé et al.'s examples of dialogs containing multiple threads are shown in Figures 16 and 17, which are Rosé et al.'s Figures 1 and 2, respectively. Figure 16 is an extended example, and Figure 17 contains a simplified example which they analyze in greater detail.

The passage in Figure 16 would be processed by our algorithm as follows. The dialog date is not given in (Rosé et al., 1995). For concreteness, let us suppose that the dialog date is *Friday 11 April*. Then, *next week* is *Monday 14 April* through *Friday 18 April* (the dialog does not mention weekend days, so we exclude them for ease of discussion). Utterance 2 is deictic, introducing *next week* into the discourse. Utterances 3-6 all have both deictic and anaphoric readings, all of which yield the correct results.

The deictic relation for all of them is the *frame of reference* deictic relation, under which the interpretations are forward references from the dialog date:

Utterance	Deictic Interpretation
3	Monday 14 April & Tuesday 15 April
4	Wednesday 16 April
5	Thursday 17 April
6	Friday 18 April

The correct interpretations of (3)-(6) are also established with the *co-reference* anaphoric relation, with antecedent *next week* in utterance 2: they each can be interpreted as specifying a more specific time than *next week*, that is, as a particular day of *next week*.

Finally, the *frame of reference* anaphoric relation yields the correct result for "Tuesday" in (3)[6] and for the times specified in utterances (4)-(6). The interpretation is the day calculated forward from the most recently mentioned Temporal Unit:

Utterance	Antecedent	Interpretation
3	Monday 14 April, Utterance 3	Tuesday 15 April
4	Tuesday 15 April, Utterance 3	Wednesday 16 April
5	Wednesday 16 April, Utterance 4	Thursday 17 April
6	Thursday 17 April, Utterance 5	Friday 18 April

Utterances (7) and (10) are potential challenges for our algorithm, representing instances of the situation depicted in Figure 13: *Thursday 24 April* is a possible incorrect interpretation of "Thursday" in these utterances, yielded by the *frame of reference* anaphoric relation. The correct interpretation is also a candidate, yielded by multiple relations: the *frame of reference* deictic relation and the *co-reference* anaphoric relation, with *Thursday 17 April* in utterance (5) as antecedent. The relative magnitude of the certainty factors of the *co-reference* and *frame of reference* anaphoric relations means that the correct interpretation is likely to be chosen in practice, as mentioned in Section 8.1.1. If the incorrect interpretation were chosen for utterances (7) and (10), then incorrect interpretations of "Friday" in each of (8) and (11) would be possible: the Friday after the incorrect date of *Thursday 24 April*, yielded by the *frame of reference* anaphoric relation. However, the correct interpretations would be possible too, yielded by the *frame of reference* deictic relation and the *co-reference* anaphoric relation.

Utterances (12) through (16) have analogous interpretations, except that the deictic interpretations yield incorrect results (that is, due to utterance 12, "How is the next week?", the days are actually of the week *Monday 21 April* through *Friday 25 April*; the deictic interpretations are of the week *Monday 14 April* through *Friday 18 April*). Thus, there are one correct and two incorrect interpretations for some of the utterances, making it less likely in practice that the correct interpretation would be chosen. Note that, generally speaking, which focus model is used does not directly address the deictic/anaphoric ambiguity, so, for the purposes of this section, the two parts of the dialog pose the same challenge to the focus model.

The dialog in Figure 17 is analogous. However, "The other day" in (5) brings up other issues. There is a special case of the *co-reference* anaphoric relation for such expressions

6. Recall that multiple Temporal Units specified in a single utterance are added to the focus list in order of mention and treated as separate discourse entities.

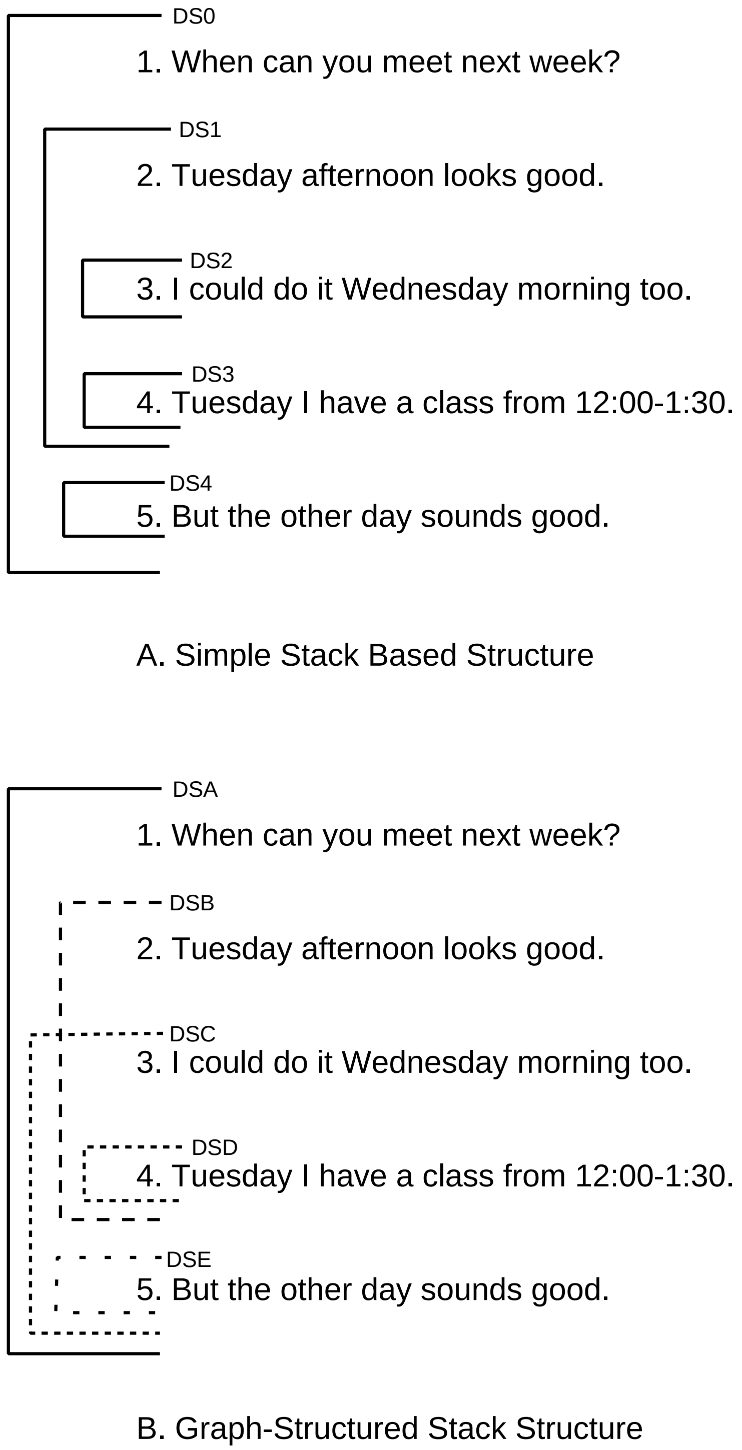

Figure 17: Sample Analysis
(Rosé et al., 1995, p. 33)

(i.e., "the other" "day"|"month"|"year"; see Anaphoric Rule 7 in Online Appendix 1). In this case, the second most recent day, month, or year, as appropriate, is the candidate antecedent. Presumably, neither the most recently mentioned day nor a day mentioned before two or more others would be referred to as "the other day"; thus, we anticipate that this is a good heuristic. Nevertheless, if (5) explicitly mentioned "Wednesday," our algorithm would have a correct and an incorrect interpretation to choose between.

In summary, there were no instances of temporal multiple threads of the type addressed by Rosé et al., either in the CMU training data upon which the algorithm was developed, or in the NMSU training data to which the algorithm was later applied. If segments such as those illustrated in Rosé et al. were to appear, an incorrect interpretation by our algorithm would be possible, but, under our model, the correct antecedent would also be available. For the examples they present, the algorithm faces the same choice: establish a *co-reference* relation to a time before the last one (the correct interpretation), or establish a *frame of reference* relation with the immediately preceding time (an incorrect interpretation). If performing temporal reference resolution is the goal, and if one is faced with an application in which such temporal multiple threads do occur, our investigation of the problem suggests that this specific situation should be investigated before assuming that a more complex focus model is needed. Adding a new *modify* anaphoric relation could be investigated. Or, as implemented in our system, a specific preference could be defined for the *co-reference* relation over the *frame of reference* relation when both are possible in a local context. Statistical techniques could be used to establish preferences appropriate for the particular application.

The different findings between Rosé et al. and our work might be due to the fact that different problems are being addressed. Having no intentional state, our model does not distinguish between times being negotiated and other times. It is possible that another structure is relevant for the intentional level. Rosé et al. do not specify whether or not this is so. The different findings may also be due to differences in the data: their protocol is like a radio conversation in which a button must be pressed in order to transmit a message, and the other participant cannot transmit a message until the speaker releases the button. This results in less dynamic interaction and longer turns (Villa, 1994). In the dialogs used here, the participants have free control over turn-taking.

8.2 Coverage and Ambiguity of the Relations Defined in the Model

A question naturally arises from the evaluation presented in the previous section: in using a less complex focus model, have we merely "pushed aside" the ambiguity into the set of deictic and anaphoric relations? In this section, we assess the ambiguity of the anaphoric relations for the NMSU and CMU training sets. This section also presents other evaluations of the relations, including an assessment of their coverage, redundancy, how often they are correct, and how often they are applicable.

The evaluations presented in this section required detailed, time-consuming manual annotations. The system's annotations would not suffice, because the implementation does not perfectly recognize when a rule is applicable. A sample of four randomly selected dialogs in the CMU training set and the four dialogs in the NMSU training set were annotated.

The counts derived from the manual annotations for this section are defined below. Because this section focuses on the relations, we consider them at the more specific level of the deictic and anaphoric rules presented in Online Appendix 1. In addition, we do not allow trivial extensions of the relations, as we did in the evaluation of the focus model (Section 8.1). The criterion for correctness in this section is the same as for the evaluation of the system: a field-by-field exact match with the manually annotated correct interpretations. There is one exception. The starting and end *time of day* fields are ignored, since these are known weaknesses of the rules, and they represent a relatively minor proportion of the overall temporal interpretation.

The following were derived from manual annotations.

- *TimeRefs:* the number of distinct times referred to in each sentence, summed over all sentences.
- *TimeRefsC:* The number of *TimeRefs* for which a correct interpretation is available under our model (whether or not an incorrect interpretation is also possible).
- *Interp:* The number of interpretations possible under the model. For the current Temporal Unit, there is one *Interp* for every rule that can be applied.
- *CorrI:* The number of *Interps* that are correct, where correctness is defined as an exact match with the manually annotated correct interpretation, except that the starting and end *time of day* fields are ignored.
- *IncI:* The number of incorrect *Interps* (i.e., *Interp* = *IncI* + *CorrI*).
- *DiffI:* The number of different interpretations
- *DiffICorr:* The number of different interpretations, excluding interpretations of Temporal Units for which there is not a correct interpretation under our model.

The values for each data set, together with coverage and ambiguity evaluations, are presented in Table 7.

The ambiguity for both data sets is very low. The *Ambiguity* figure in Table 7 represents the average number of interpretations per temporal reference, considering only those for which the correct interpretation is possible (i.e., it is $DiffICorr$ / $TimeRefsC$). The table also shows the ambiguity when all temporal references are included (i.e., $DiffI$ / $TimeRefs$). As can be seen from the table, the average ambiguity in both data sets is much less than two interpretations per utterance.

The coverage of the relations can be evaluated as ($TimeRefsC$ / $TimeRefs$), the percentage of temporal references for which at least one rule yields the correct interpretation. While the coverage of the NMSU data set, 85%, is not perfect, it is good, considering that the system was not developed on the NMSU data.

The data also show that there is often more than one way to achieve the correct interpretation. This is another type of redundancy: redundancy of the data with respect to the model. It is calculated in Table 7 as ($CorrI$ / $TimeRefsC$), that is, the number of correct interpretations over the number of temporal references that have a correct interpretation.

CMU Training Set 4 randomly selected dialogs						
TimeRefs	TimeRefsC	Interp	CorrI	IncI	DiffI	DiffICorr
78	74	165	142	23	91	85

Coverage (TimeRefsC / TimeRefs) = **95%**
Ambiguity (DiffICorr / TimeRefsC) = **1.15**
Overall Ambiguity (DiffI / TimeRefs) = 1.17
Rule Redundancy (CorrI / TimeRefsC) = 142/74 = 1.92 %

NMSU Training Set 4 dialogs						
TimeRefs	TimeRefsC	Interp	CorrI	IncI	DiffI	DiffICorr
98	83	210	154	56	129	106

Coverage (TimeRefsC / TimeRefs) = **85%**
Ambiguity (DiffICorr / TimeRefsC) = **1.28**
Overall Ambiguity (DiffI / TimeRefs) = 1.32
Rule Redundancy (CorrI / TimeRefsC) = 154 / 83 = 1.86 %

Table 7: Coverage and Ambiguity

CMU Training Set 4 randomly selected dialogs			
Rule	Correct	Total	Accuracy
D1	4	4	1.00
D2i	0	0	0.00
D2ii	**35**	**40**	**0.88**
a frame-of-reference deictic relation			
D3	1	2	0.50
D4	0	0	0.00
D5	0	0	0.00
D6	2	2	1.00
A1	**45**	**51**	**0.88**
a co-reference anaphoric relation			
A2	0	0	0.00
A3i	1	1	1.00
A3ii	**35**	**37**	**0.95**
a frame-of-reference anaphoric rel.			
A4	**14**	**18**	**0.78**
a modify anaphoric relation			
A5	0	0	0.00
A6i	2	2	1.00
A6ii	1	1	1.00
A7	0	1	0.00
A8	0	0	0.00

NMSU Training Set 4 dialogs			
Rule	Correct	Total	Accuracy
D1	4	4	1.00
D2i	0	0	0.00
D2ii	**24**	**36**	**0.67**
a frame-of-reference deictic relation			
D3	6	9	0.67
D4	0	1	0.00
D5	0	0	0.00
D6	0	0	0.00
A1	**57**	**68**	**0.84**
a co-reference anaphoric relation			
A2	5	5	1.00
A3i	0	0	0.00
A3ii	**21**	**32**	**0.66**
a frame-of-reference anaphoric rel.			
A4	**27**	**37**	**0.73**
a modify anaphoric relation			
A5	0	1	0.00
A6i	7	9	0.78
A6ii	0	0	0.00
A7	0	0	0.00
A8	0	0	0.00

Table 8: Rule Applicability Based on Manual Annotations

For both data sets, there are, on average, roughly two different ways to achieve the correct interpretation.

Table 8 shows the number of times each rule applies in total (column 3) and the number of times each rule is correct (column 2), according to our manual annotations. Column 4 shows the accuracies of the rules, i.e., (column 2 / column 3). The rule labels are the ones used in Online Appendix 1 to identify the rules.

The same four rules are responsible for the majority of applications in both data sets, the ones labeled *D2ii*, *A1*, *A3ii*, and *A4*. The first is an instance of the *frame of reference* deictic relation, the second is an instance of the *co-reference* anaphoric relation, the third is an instance of the *frame of reference* anaphoric relation, and the fourth is an instance of the *modify* anaphoric relation.

How often the system considers and actually uses each rule is shown in Table 9. Specifically, the column labeled *Fires* shows how often each rule applies, and the column labeled *Used* shows how often each rule is used to form the final interpretation. To help isolate the accuracies of the rules, these experiments were performed on unambiguous data. Comparing this table with Table 8, we see that the same four rules shown to be the most important by

CMU data set		
Name	Used	Fires
D1	16	16
D2i	1	3
D2ii	**78**	**90**
a frame-of-reference deictic relation		
D3	5	5
D4	9	9
D5	0	1
D6	2	2
A1	**95**	**110**
a co-reference anaphoric relation		
A2	2	24
A3i	1	1
A3ii	**72**	**86**
a frame-of-reference anaphoric rel.		
A4	**45**	**80**
a modify anaphoric relation		
A5	4	5
A6i	10	10
A6ii	0	0
A7	0	0
A8	1	1

NMSU data set		
Name	Used	Fires
D1	4	4
D2i	2	2
D2ii	**20**	**31**
a frame-of-reference deictic relation		
D3	2	3
D4	0	0
D5	0	0
D6	0	0
A1	**46**	**65**
a co-reference anaphoric relation		
A2	6	12
A3i	0	2
A3ii	**18**	**27**
a frame-of-reference anaphoric rel.		
A4	**24**	**42**
a modify anaphoric relation		
A5	3	5
A6i	6	8
A6ii	0	0
A7	0	0
A8	0	0

Table 9: Rule Activation by the System on Unambiguous Data

the manual annotations are also responsible for the majority of the system's interpretations. This holds for both the CMU and NMSU data sets.

8.3 Evaluation of the Architectural Components

In this section, we evaluate the architectural components of our algorithm using degradation (ablation) studies. We perform experiments without each component in turn, and then with none of them, to observe the impact on the system's performance. Such studies have been useful in developing practical methods for other kinds of anaphora resolution as well (see, for example, Mitkov & Stys, 1997). Specifically, an experiment was performed testing each of the following variations.

1. The certainty factors of all of the rules are set to 1.

   Recall that all rules are applied to each utterance, and each rule that matches produces a Partial-Augmented-ILT (which is assigned the certainty factor of the rule). All maximal mergings of the Partial-Augmented-ILTs are then formed, to create a set of Augmented-ILTs. Then, the final interpretation of the utterance is chosen from among the set of Augmented-ILTs. The certainty factor of each Augmented-ILT is the sum of the certainty factors of the Partial-Augmented-ILTs composing it. Thus, setting the certainty factors to 1 implements the scheme in which the more partial results are merged into an interpretation, the higher the overall certainty factor of that interpretation. In other words, this scheme favors the Augmented-ILT resulting from the greatest number of rule applications.

2. The certainty factors of all of the rules are set to 0.

   This scheme is essentially random selection among the Augmented-ILTs that make sense according to the critics. If the critics did not exist, then setting the rule certainty factors to 0 would result in random selection. With the critics, any Augmented-ILTs to which the critics apply are excluded from consideration, because the critics will lower their certainty factors to negative numbers.

3. No merging of the rule results is performed.

   That is, the Partial-Augmented-ILTs are not merged prior to selection of the final Augmented-ILT. The effect of this is that the result of one single rule is chosen to be the final interpretation.

4. The critics are not used.

5. The distance factors are not used.

   In this case, the certainty factors for rules that access the focus list are not adjusted based on how far back the chosen focus list item is.

6. All variations are applied, excluding case 2.

   Specifically, neither the critics nor the distance factors are used, no merging of partial results is performed, and the rules are all given the same certainty factor (namely, 1).

Table 10 shows the results for each variation when run over the unambiguous but uncorrected CMU training data. For comparison, the first row shows the results for the system as normally configured. As with the previous evaluations, accuracy is the percentage of the correct answers the system produces, while precision is the percentage of the system's answers that are correct.

Only two of the differences are statistically significant ($p \leq 0.05$), namely, the precision of the system's performance when the critics are not used, and the accuracy of the system's performance when all of the certainty factors are 0. The significance analysis was performed using paired t-tests comparing the results for each variation with the results for the system as normally configured.

The performance difference when the critics are not used is due to extraneous alternatives that the critics would have weeded out. The drop in accuracy when the certainty factors are all 0 shows that the certainty factors have some effect. Experimenting with statistical methods to derive them would likely lead to further improvement.

The remaining figures are all only slightly lower than those for the full system, and are all much higher than the baseline accuracies.

Variation	Cor	Inc	Mis	Ext	Nul	Act	Poss	**Acc**	**Prec**
system as is	1283	44	112	37	574	1938	2013	0.923	0.958
all CFs 1.0	1261	77	101	50	561	1949	2000	0.911	0.935
all CFs 0.0	1202	118	119	49	562	1931	2001	0.882	0.914
-critics	1228	104	107	354	667	2353	2106	0.900	0.805
-dist. factors	1265	52	122	50	591	1958	2030	0.914	0.948
-merge	1277	46	116	54	577	1954	2016	0.920	0.949
combo	1270	53	116	67	594	1984	2033	0.917	0.940

Legend

Cor(rect):	System and key agree on non-null value
Inc(orrect):	System and key differ on non-null value
Mis(sing):	System has null value for non-null key
Ext(ra):	System has non-null value for null key
Nul(l):	Both system and key give null answer
Poss(ible):	Correct + Incorrect + Missing + Null
Act(ual):	Correct + Incorrect + Extra + Null
Base(line)Acc(uracy):	Baseline accuracy (input used as is)
Acc(uracy):	% Key values matched correctly ((Correct + Null)/Possible)
Prec(ision):	% System answers matching the key ((Correct + Null)/Actual)

Table 10: Evaluation of the Variations on CMU Unambiguous/Uncorrected Data

It is interesting to note that the unimportance of the distance factors (variation 5) is consistent with the findings presented in Section 8.1 that the last mentioned time is an acceptable antecedent in the vast majority of cases. Otherwise, we might have expected to see an improvement in variation 5, since the distance factors penalize going further back on the focus list.

9. Conclusions

Scheduling dialogs, during which people negotiate the times of appointments, are common in everyday life. This paper reports the results of an in-depth empirical investigation of resolving explicit temporal references in scheduling dialogs. There are four basic phases of this work: data annotation, model development, system implementation and evaluation, and model evaluation and analysis. The system and model were developed primarily on one set of data (the CMU dialogs), and then applied later to a much more complex set of data (the NMSU dialogs), to assess the generalizability of the model for the task being performed. Many different types of empirical methods were applied to both data sets to pinpoint the strengths and weaknesses of the approach.

In the data annotation phase, detailed coding instructions were developed and an inter-coder reliability study involving naive subjects was performed. The results of the study are very good, supporting the viability of the instructions and annotations. During the model development phase, we performed an iterative process of implementing a proposed set of anaphoric and deictic relations and then refining them based on system performance (on the CMU training data), until we settled on the set presented here. We also developed our focus model during this phase. The question of what type of focus model is required for various tasks is a question of ongoing importance in the literature. It appeared from our initial observations of the data that, contrary to what we expected, a recency-based focus model might be adequate. To test this hypothesis, we made the strategic decision to limit ourselves to a recency-based model, rather than build some kind of hybrid model whose success or failure would not have told us as much.

During system implementation and evaluation, a system implementing the model was implemented and evaluated on unseen test data, using a challenging field-by-field comparison of system and human answers. To be considered the right answer, the information must not only be correct, but must also be included in the correct field of the output representation. Taking as input the ambiguous output of a semantic grammar, the system achieves an overall accuracy of 81% on unseen CMU test data, a large improvement over the baseline accuracy of 43%. On an unseen test set from the more complex NMSU data, the results are very respectable: an overall accuracy of 69%, with a much lower baseline accuracy of 29%. This also shows the robustness of the CMU semantic parser (Lavie & Tomita, 1993; Levin et al., 1995), which was given the NMSU dialogs as input without being modified in any way to handle them.

The implementation is an important proof of concept. However, it is not a direct evaluation of the model, because there are errors due to factors we do not focus on in this work. Some of the error is simply due to utterance components being outside the coverage of the CMU parser, or having high semantic ambiguity. The only information we use to perform semantic disambiguation is the temporal context. The Enthusiast researchers have already developed better techniques for resolving the semantic ambiguity in these dialogs (Shum et al., 1994), which could be used to improve performance.

Thus, in the model evaluation and analysis phase, we performed extensive additional evaluation of the algorithm itself. We focus on the relations and the focus model, because they are the main contributions of this work. Our degradation studies support this, as they show that the other aspects of the algorithm, such as the distance factors and merging process, are responsible for little of the system's success (see Section 8.3).

Our evaluations show the strength of the focus model for the task, not only for the CMU data on which it was developed, but also for the more complex NMSU data. While the NMSU data is more complex, there are few cases in which the last mentioned time is not an appropriate antecedent, highlighting the importance of recency (Walker, 1996); see Section 8.1. We characterized those cases along a number of dimensions, to identify the particular types of challenges they pose (see Figure 10).

In order to compare our work to that of others, we formally defined subdialogs and the multiple thread structures addressed by Rosé et al. (1995) with respect to our model and the specific problem of temporal reference resolution. An interesting finding is that, while subdialogs of the types addressed by Grosz and Sidner (1986) were found in the data, no cases of multiple threads were found. That is, some subdialogs, all in the NMSU data, mention times that potentially interfere with the correct antecedent. But in none of these cases would subsequent errors result if, upon exiting the subdialog, the offending information were popped off a discourse stack or otherwise made inaccessible. Changes in tense, aspect, and modality are promising clues for recognizing subdialogs in this data, which we plan to explore in future work.

To assess whether or not using a simpler focus model requires one to use a highly ambiguous set of relations, we performed a separate evaluation of the relations, based on detailed, manual annotations of a set of dialogs. The ambiguity of the relations for both data sets is very low, and the coverage is good (see Table 7). In a comparison of system and human annotations, the same four rules identified to be most important in the manual annotations are responsible for the majority of the system's interpretations for both data sets (see Tables 8 and 9), suggesting that the system is a good implementation of the model.

Recently, many in computational discourse processing have turned to empirical studies of discourse, with a goal to develop general theories by analyzing specific discourse phenomena and systems that process them (Walker & Moore, 1997). We contribute to this general enterprise. We performed many different evaluations, on the CMU data upon which the model was developed, and on the more complex NMSU data. The task and model components were explicitly specified to facilitate evaluation and comparison. Each evaluation is directed toward answering a particular question; together, the evaluations paint an overall picture of the difficulty of the task and of the success of the proposed model.

As a contribution of this work, we have made available on the project web page the coding instructions, the NMSU dialogs, and the various kinds of manual annotations we performed.

10. Acknowledgements

This research was supported in part by the Department of Defense under grant number 0-94-10. A number of people contributed to this work. We want to especially thank David Farwell, Daniel Villa, Carol Van Ess-Dykema, Karen Payne, Robert Sinclair, Rocio Guillén, David Zarazua, Rebecca Bruce, Gezina Stein, Tom Herndon, and the project members of Enthusiast at CMU, whose cooperation greatly aided our project. We wholeheartedly thank the anonymous reviewers, whose comments and criticisms were very helpful. We also thank Esther Steiner, Philip Bernick, and Julie England for participating in the intercoder reliability study, and Linda Fresques for proofreading the paper.

References

Alexandersson, J., Reithinger, N., & Elisabeth, M. (1997). Insights into the dialogue processing of Verbmobil. In *Proc. 5th Conference on Applied Natural Language Processing*, pp. 33–40. Association for Computational Linguistics.

Allen, J. (1984). Toward a general theory of action and time. *Artificial Intelligence*, *23*, 123–154.

Allen, J., & Perrault, C. (1980). Analyzing intention in utterances. *Artificial Intelligence*, *15*, 143–178.

Arhenberg, L., Dahlbäck, N., & Jönsson, A. (1995). Coding schemes for natural language dialogues. In *Working Notes of AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pp. 8–13.

Busemann, S., Declerck, T., Diagne, A. K., Dini, L., Klein, J., & Schmeier, S. (1997). Natural language dialogue service for appointment scheduling agents. In *Proc. 5th Conference on Applied Natural Language Processing*, pp. 25–32. Association for Computational Linguistics.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, *22(2)*, 249–254.

Clark, H. H. (1977). Bridging. In Johnson-Laird, P. N., & Wason, P. C. (Eds.), *Thinking: Readings in Cognitive Science*. Cambridge University Press.

Condon, S., & Cech, C. (1995). Problems for reliable discourse coding schemes. In *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pp. 27–33.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21(2)*, 203–225.

Grosz, B., & Sidner, C. (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, *12(3)*, 175–204.

Hays, W. L. (1988). *Statistics* (Fourth edition). Holt, Rinehart, and Winston.

Heim, I. (1982). *The Semantics of Definite and Indefinite Noun Phrases.* Ph.D. thesis, University of Massachusetts at Amherst.

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. 34th Annual Meeting of the Association for Computational Linguistics*, pp. 286–293.

Hobbs, J. (1978). Resolving pronoun references. *Lingua*, *44*, 311–338.

Hwang, C., & Schubert, L. (1992). Tense trees as the "fine structure" of discourse. In *Proc. 30th Annual Meeting of the Association for Computational Linguistics*, pp. 232–240.

Isard, A., & Carletta, J. (1995). Replicability of transaction and action coding in the map task corpus. In *Working Notes of AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pp. 60–66.

Kameyama, M., Passonneau, R., & Poesio, M. (1993). Temporal centering. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 70–77.

Kamp, H., & Reyle, U. (1993). *From Discourse to Logic.* Kluwer Academic Publisher, Dordrecht, The Netherlands.

Lascarides, A., Asher, N., & Oberlander, J. (1992). Inferring discourse relations in context. In *Proc. 30th Annual Meeting of the Association for Computational Linguistics*, pp. 1–8.

Lavie, A., & Tomita, M. (1993). GLR* - an efficient noise skipping parsing algorithm for context free grammars. In *Proc. 3rd International Workshop on Parsing Technologies.*

Levin, L., Glickman, O., Qu, Y., Gates, D., Lavie, A., Rosé, C., Van Ess-Dykema, C., & Waibel, A. (1995). Using context in the machine translation of spoken language. In *Proc. Theoretical and Methodological Issues in Machine Translation (TMI-95).*

Litman, D., & Passonneau, R. (1995). Combining multiple knowledge sources for discourse segmentation. In *Proc. 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 130–143.

Mann, W., & Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, *8(3)*, 243–281.

Mitkov, R., & Stys, M. (1997). Robust reference resolution with limited knowledge: high precision genre-specific approach for English and Polish. In *Recent Advances in Natural Language Processing (RANLP-97)*, pp. 74–81. European Commission, DG XIII.

Moser, M., & Moore, J. (1995). Investigating cue selection and placement in tutorial discourses. In *Proc. 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 130–143.

Nakhimovsky, A. (1988). Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, *14(2)*, 29–43.

O'Hara, T., Wiebe, J., & Payne, K. (1997). Instructions for annotating temporal information in scheduling dialogs. Tech. rep. MCCS-97-308, Computing Research Laboratory, New Mexico State University.

Passonneau, R., & Litman, D. (1993). Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 148–155.

Poesio, M., Vieira, R., & Teufel, S. (1997). Resolving bridging references in unrestricted text. In *Proc. Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts.* Association for Computational Linguistics.

Qu, Y., Eugenio, B. D., Lavie, A., Levin, L., & Rosé, C. (1996). Minimizing cumulative error in discourse context. In *ECAI Workshop Proceedings on Dialogue Processing in Spoken Language Systems.*

Rosé, C., Eugenio, B. D., Levin, L., & Van Ess-Dykema, C. (1995). Discourse processing of dialogues with multiple threads. In *Proc. 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 31–38.

Shum, B., Levin, L., Coccaro, N., Carbonell, J., Horiguchi, K., Isotani, H., Lavie, A., Mayfield, L., Rosé, C., Van Ess-Dykema, C., & Waibel, A. (1994). Speech-language integration in a multi-lingual speech translation system. In *Proceedings of the AAAI Workshop on Integration of Natural Language and Speech Processing.*

Sidner, C. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse.* Ph.D. thesis, MIT.

Sidner, C. (1983). Focusing in the comprehension of definite anaphora. In Brady, M., & Berwick, R. C. (Eds.), *Computational Models of Discourse*, pp. 267–330. MIT Press, Cambridge, MA.

Siegel, S., & Castellan, Jr., N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (Second edition). McGraw-Hill, New York.

Song, F., & Cohen, R. (1991). Tense interpretation in the context of narrative. In *Proc. 9th National Conference on Artificial Intelligence (AAAI-91)*, pp. 131–136.

Villa, D. (1994). Effects of protocol on discourse internal and external illocutionary markers in Spanish dialogs. In *Linguistic Association of the Southwest Conference XXIII.*

Walker, L., & Moore, J. (1997). Empirical studies in discourse. *Computational Linguistics*, *23(1)*, 1–12.

Walker, M. (1996). Limited attention and discourse structure. *Computational Linguistics*, *22(2)*, 255–264.

Webber, B. L. (1983). So what can we talk about now?. In Brady, M., & Berwick, R. (Eds.), *Computational Models of Discourse.* MIT Press, Cambridge.

Webber, B. (1988). Tense as discourse anaphor. *Computational Linguistics*, *14(2)*, 61–73.

Wiebe, J., Farwell, D., Villa, D., Chen, J.-L., Sinclair, R., Sandgren, T., Stein, G., Zarazua, D., & O'Hara, T. (1996). Artwork: Discourse processing in machine translation of dialog. Tech. rep. MCCS-96-294, Computing Research Laboratory, New Mexico State University.