1980 TOC Search

10

Retrieving time information from natural-language texts

Lynette Hirschman

10.1 Introduction

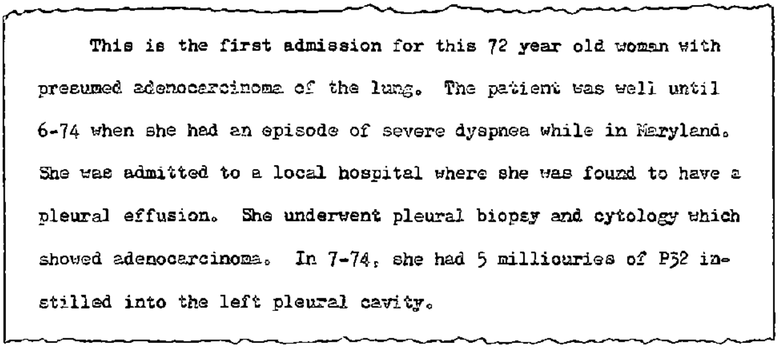
An understanding of time relations is central to processing information contained in a narrative. A typical narrative is concerned with the relative ordering or progression of events over time, and the information that one would like to retrieve is often of the type 'What happened after event x?' or 'Did event x precede event y?' Determination of causality also requires a knowledge of time relations, since an event *x* can cause an event *y* only if it precedes event *y* in time.

There has been considerable interest in the processing of time information among researchers in artificial intelligence. Several systems have been designed which compute time relations from a structured input of time specifications. Kahn and Gorry (1977) present a 'time specialist' program which accepts input about time relations in a LISP-like form; from this input, the 'specialist' computes time relations, checks for inconsistencies in the input and answers questions about time relations between the events given in the input. The 'specialist' provides several different ways of organising the time relations in order to facilitate retrieval.

Findler and Chen (1973) present a model for processing durational and time-point information which accepts imprecise and partial temporal specifications in symbolic form, for a set of events. From this information the program can check the consistency of the input and can retrieve relations such as possible causality, coexistence and relative duration for a pair of events.

Bruce (1972) presents a theoretical model for the time information expressed by tense. He illustrates this model by a small question–answering system, CHRONOS, which accepts interactive input in a 'stylized form of English' and answers questions about the relative time of the input events.

The work described in this chapter differs from the work referred to above in that it focuses specifically on the problem of extracting time information from coherent natural-language text. The analysis of narrative poses a more difficult problem than the analysis of dialogue with respect to implicit information. Natural language is an effective means of communication, because so much information can be packed into a very small space. One of the major problems in automatic language processing is how to 'unpackage' all this compressed



This is the first admission for this 72 year old woman with presumed adenocarcinoma of the lung. The patient was well until 6-74 when she had an episode of severe dyspnea while in Maryland. She was admitted to a local hospital where she was found to have a pleural effusion. She underwent pleural biopsy and cytology which showed adenocarcinoma. In 7-74, she had 5 millicuries of P32 instilled into the left pleural cavity.

Figure 10.1 First paragraph from a hospital record

information into an explicit, computable form. The program must restore all the pieces of information that the speaker/writer was able to omit because the information was 'redundant' —that is, reconstructable from context. Because the stretches of connected text produced by a writer are much longer in narrative than in dialogue, there is more 'context', which, in turn, allows of more complex types of omissions. In addition, in processing narrative there is no opportunity for a 'clarification' dialogue; the processing program must use whatever information is available, even when the information is incomplete or imprecise.

The program described in the following sections was developed in the course of ongoing research in natural-language processing at the Linguistic String Project of New York University. This research has been concerned with the creation of a database from natural-language input and the retrieval of information from such a database (Hirschman and Grishman, 1977; Sager, 1978). The work on narrative time was an outgrowth of an experiment on retrieval of information from narrative medical records (Hirschman *et al.*, to be published). In this experiment we converted a set of criteria for evaluating patient care into a set of retrieval routines and applied them to a database obtained by processing a set of hospital discharge summaries. Most of the retrieval requests in this application involved the computation of simple time relations, such as 'Was a blood culture done at admission?' A number of requests required computation of more complex time relations, such as 'Did the chest X-ray show improvement?' To answer this question for a given record, the retrieval routine had to locate both the first and the last chest X-ray associated with the current hospitalisation, then determine whether the later X-ray represented a 'better' state (according to a set of medical criteria) than the earlier X-ray.

A paragraph from a medical record (shown in *Figure 10.1*) illustrates the fact that most of the events in the narrative do not have an explicit time (for example, a date) associated with them. Nonetheless, sufficient time information is present in this narrative so that it is possible for a human reader to answer time-related questions about the events described. For example, a reader can answer the questions

(1) When was cancer first reported?
Answer: in June 1974.

(2) What signs or symptoms preceded or coincided with the first diagnosis of cancer?
Answer: pleural effusion and dyspnea.

Understanding how time is expressed in natural language is a necessary first step in processing time information. Clearly there are a number of ways of expressing time (other than by explicit dates) that permit a reader to determine time relations between the narrated events. Section 10.2 outlines these mechanisms, which include adverbial expressions (for example, *yesterday*), tense, multiple references to a single event, time-related conjunction and verb connectives (for example, *when, coincide with*), co-ordinate conjunction (for example, *and*) and change of state expressions (for example, *subsided* in *fever subsided*).

The next step is to design a program which both recognises time information as it occurs in the text and translates this information into a suitable representation. The time program described in this chapter is embedded in the larger natural-language processing system of the Linguistic String Project (Sager, 1978). This system receives as input the sentences of a text; it parses the sentences to determine the syntactic relations, performs information-preserving transformations to regularise the syntax and maps the resultant into a table-like structure whose columns correspond to the types of information in the material. *Figure 10.2* shows a schematic representation of the paragraph from *Figure 10.1* in its medical tabular form. This structured or 'formatted' information (the output of the natural-language processing system) is the input to the time program. Section 10.2 also explains how the time program utilises this structure to recognise the various types of time information and convert them into a suitable representation.

Much of the time information in medical records is not given explicitly; it is given only relative to other previously mentioned events and is often inexplicit or imprecise. The program for the retrieval of time relations between two events must use this relative and imprecise information. By making certain reasonable assumptions about the structure of narrative, it is possible to gather sufficient information about time relations between the events in the narrative to answer correctly a set of time-related questions (such as the two questions given above). The actual computation of time relations is discussed in Section 10.3. Section 10.4 gives a brief summary of the implementation and Section 10.5 outlines some planned extensions of the program and areas for future research.

10.2 Time expressions in natural language and their representation

This section identifies some of the ways in which time relations are expressed in natural language and discusses how time relations are represented in the time program described here. The natural-language processing divides the narrative into 'events', corresponding to simple assertions. The time program identifies time information as it occurs in the analysed text and associates with each event a time point or a time period. Each time point or time period is represented as a 'reference time' adjusted forward or backward by a certain

Code	Conn	Patient	Inst.	V-MD	V-TR	Med	V-PT	Body-part	Lab.	V-Show	Normalcy	Quant.	Sign-symp.	Diag.	Time	Tense
HIDSH 1.1.1!	BE															Present
HIDSH 1.1.1	REL-CL	(For this) 72 year old woman		(The 1^{st}) admission												
HIDSH 1.1.1		Woman					With	(Of the) Lung						(Presumed) Adenocarcinoma		
HIDSH 1.1.2	When	(The) Patient					Be				Well				Until 6/0/74	Past
HIDSH 1.1.2	While	She					Have					Severe	(An episode of) Dyspnea			Past
HIDSH 1.1.2*		She														
HIDSH 1.1.3	Where	She	(A local) Hospital	Admit (to)												Past
HIDSH 1.1.3	Embedded			Find												Past
HIDSH 1.1.3		She					Have	Pleural					(A) Effusion			
HIDSH 1.1.4	And	She		Undergo				Pleural	Biopsy							Past
HIDSH 1.1.4	REL-CL		Cytology†													
HIDSH 1.1.4			Cytology							Show				Adenocarcinoma		Past
HIDSH 1.1.5	REL-CL	She				(Of) P32	Have					5 milli curies			In 7/0/74	Past
HIDSH 1.1.5					Instill (into)	P32		(The left) Pleural cavity								

Note: Each connective (CONN) links the format line where it appears to the next format line(s).
Left and right modifiers appear in parentheses above and below the main entry in the column.

! This line corresponds to *this* in *This is the first admission*, but *this* cannot be placed in a column until the word it refers to (*admission*) is established.

* The second argument of *while* (namely *she be in Maryland*) did not contain medical information and was not formatted.

† There is no predicate on *cytology* in this line because the sequence *cytology which showed adenocarcinoma* was taken as a separate unit, rather than as a conjoined object of *undergo* (*she underwent cytology*?).

Figure 10.2 Tabular representation of medical information in a paragraph from a hospital discharge summary

Text	Relation	Reference point	Adjustment		
			Direction	Quantity	Time-unit
Ex. 1 on 7/3/74	At	7/3/74			
Ex. 2 Yesterday	At	Time of narration	−	1	Day
Ex. 3 Then	At	Last narrated topic	+		
Ex. 4 A week before admission	At	Admission	−	1	Week
Ex. 5 For the past week	From	Time of narration	−	1	Week
	To	Time of narration			

Figure 10.3 Time representation for adverbial time expressions

number of time units, as specified in the text. *Figure 10.3* illustrates this representation for five types of time expressions (taken from five different texts).

10.2.1 Adverbial time expressions

Adverbial expressions offer the richest time information. During the natural-language processing stage, they are separated into a special TIME column (*see* lines 4 and 13 in *Figure 10.2*). Adverbial expressions include dates (for example, *on 9/23/75*), prepositional phrases (for example, *after discharge*), adverbial time expressions (for example, *last week*, *this admission*) and simple adverbs (for example, *later*, *yesterday*). A date is itself a reference point: it anchors the event to a time scale external to the narrative. For example, in the sentence *she had P32 instilled on 7-3-74*, the date *7-3-74* gives the time of the event *she had P32 instilled* (*see* example 1, *Figure 10.3*).

Other adverbial expressions describe time relative to another event in the narrative — for example, *At admission, haematocrit was 24.* Here *admission* serves as the reference event. There are also composite time adverbials, such as *a week before admission* in the sentence *A week before admission, patient developed fever of 105 degrees.* This composite expression also gives the time of one event (*patient developed fever of 105 degrees*) in terms of the time of a reference event (*admission*) adjusted by a certain amount (*a week before admission = time of admission – 1 week; see* example 4, *Figure 10.3*).

The representation of time in terms of reference event plus adjustment is also adequate for other adverbial expressions, provided that the reference event is reconstructed appropriately. For example, certain adverbs contain an implicit reference to the time of narration — that is, the time when the sentence was uttered or written: *He is now dead from car accident.* Here *now* is taken to mean at the time of writing (*AT THE TIME OF NARRATION*). Similarly, the expression *yesterday* (*The patient developed cough yesterday*) also fits into this representation when broken down into its components *TIME OF NARRATION — 1 DAY* (*see* example 2, *Figure 10.3*). Other adverbs use as a reference event the time of the last narrated topic. For example, in the sentence

Patient continued to have stiff neck and then to show improvement in appetite, *then* provides the information that the time of the second event in the clause (*to show improvement in appetite*) is later than the event in the first clause (*patient continued to have stiff neck*). Therefore the first event is used as the reference event for the second event. In the chart (example 3, *Figure 10.3*) the reference event is given as *LAST NARRATED TOPIC* — that is, the event from the preceding main clause (*patient continued to have stiff neck*, not shown in the figure). For the word *then*, the direction of adjustment is '+' (later); the amount of adjustment is left blank because the word *then* does not specify the amount of time between these two events.

Durational expressions are handled by assigning the relations *FROM* and *TO* in the RELATION column (*see* example 5, *Figure 10.3*). These mark the beginning and end of the duration, respectively. The sentence *For the past week, the patient has had shortness of breath* contains the durational expression *for the past week*. The beginning of the duration is represented in relation to the TIME OF NARRATION, with ADJUSTMENT= *—1WEEK*. The word *past* provides both the negative DIRECTION and the reference event TIME OF NARRATION (*see* example 5, *Figure 10.3*).

The routines for treating adverbial time expressions cover some 35 time words, including prepositions (*at*, *on*, *for*, *of*), adjectives (*past*, *earlier*, *present*, *next*) and adverbs (*now*, *then*, *later*). These time words are divided into nine different classes, each one associated with a routine in the time program. The four major classes are:

(1) 'plus' expressions, mapping into '+' in DIRECTION (*after*, *later*, *subsequently*, etc.)
(2) 'minus' expressions, mapping into '–' in DIRECTION (*prior*, *previously*, *past*, *before*, etc.)
(3) 'zero' or 'same time' expressions, mapping into '0' in DIRECTION (*at*, *on*, *present*, *now*, etc.)
(4) durational expressions, mapping to the RELATIONs 'FROM' and 'TO' (*for*, *during*, *throughout*, *over*, etc.)

There are also smaller classes to handle prepositions which can be interpreted in more than one way, depending on what other time information is present (*in*, *of*); and there are separate classes for prepositions providing complex time information, such as *since* and *until*.

Certain of these time words or expressions also carry an implicit REFERENCE POINT equal to TIME OF NARRATION — for example, *now*, *yesterday*, *this morning*, *ago*, *past*. The other time words are taken to refer to the time of the reference event, if present, or to the time of the LAST NARRATED TOPIC otherwise (*see* example 3 with *then*, *Figure 10.3*).

10.2.2 Verb tense and aspect

Verb tense and aspect are another means of expressing time in relation to certain implicit time points. For example, the future tense relates the time of the narrated event to a time following the time of narration, as in *she will be followed in haematology clinic*. The present tense corresponds to time of narration and the past tense to a time earlier than the time of narration. The present perfect (present tense plus perfect aspect) indicates an event which

Time Mechanism	Text
Tense	1. High fever was first noted one day before admission. Fever has persisted.
Time connective	2. Patient was discharged when signs of pneumonia were gone.
Causal connective	3. Fever precipitated a convulsion.
Referential	4. A transfusion of 100 cc packed red blood cells was given. Patient lapsed into coma after completion of transfusion.
Change of state	5. Pain seemed to subside.

	Event #	Connective	Text	Relation	Ref. point	Adjustment		
						Dir.	Quant.	Unit
1. Tense	E1		High fever was first noted	At	Admission	–	1	Day
	E2		Fever has persisted	From To	E1 Time of narration			
2. Time conn.	E1	When	Patient was discharged	At	Last narrated topic	0/+		
	E2		Signs of pneumonia were gone	At	E1			
3. Causal conn.	E1	Precipitated	Fever	At	Last narrated topic	0/+		
	E2		A convulsion	At	E1	+		
4. Refer-ential	E1	Expand-ref	A transfusion of 100 cc packed red blood cells was given	At	Last narrated topic	0/+		
	E2		Patient lapsed into coma	At	E3	+		
	E3		Completion of transfusion	At	E1			
5. Change of state	E1	Change-state	Pain	At	E2	–		
	E2		Pain seemed to subside	At	Last narrated topic			

Figure 10.4 Representation of time for various natural-language constructions

started at or after the time of the LAST NARRATED TOPIC, and has continued up to the time of narration — for example, *has persisted* in the sequence *High fever was first noted 1 day before admission. Fever has persisted.* This requires the expression of a duration or time interval, indicated in example 1 of *Figure 10.4* by two time expressions, marked with RELATIONs FROM and TO. The processing of tense is facilitated by the fact that prior syntactic processing has extracted tense and aspectual information, and has placed it in a special TENSE column in the tabular representation (*see Figure 10.2*).

10.2.3 Connectives

Connections between events also provide relative time information about the events. For example, a number of subordinate conjunctions (*when*, *while*, *before*, *after*) provide specifically temporal connections between two clauses. In the sentence *Patient was discharged when signs of pneumonia were gone* the time program interprets the connective *when* as providing the information that the subordinate clause event and the main clause event occurred at the same time. The representation for this is shown in example 2 of *Figure 10.4*, where the time of the subordinate clause is expressed in terms of the time of the main clause*.

Connectives can appear in a number of syntactic forms other than subordinate conjunction. For example, the verb *precipitate* is a connective in the sentence *Fever precipitated a convulsion*. The time of the convulsion is taken to be later than the time of the (onset of) fever; *see* example 3 of *Figure 10.4*. Other connective verb phrases are *be associated with*, *cause*, *precede*, *follow*, *coincide with*, etc. There are also noun connectives — for example, *the source of*, *the result of*, and adjective phrases (*due to*).

The natural-language processing stage maps connectives into a special column CONN. The time program looks in the column CONN to find any connectives and then determines what, if any, time information the connective contributes.

There are a number of connectives which are not temporal. For example, a relative clause connective (e.g. *the woman who had adenocarcinoma*) is not explicitly temporal; neither is sentence embedding (CONN = EMBEDDED), where the subject or object is itself a (reduced) sentence, such as *She was found to have pleural effusion* = *() found that she have pleural effusion* (*see* lines 7–8, *Figure 10.2*, and E9–E10, *Figure 10.5*). When two events are connected by a connective with no specific temporal meaning, the time program assigns the two connected events the same time, provided that there is no information to the contrary: the reference point of the second event is taken as the event number of the first event with no adjustment (*see* E3, E10 and E13, *Figure 10.5*).

10.2.4 Special events with known time

In a given context certain special events may have a known time associated with them. For example, in medical records the date of birth of the patient is an event with a known time, because the birth date appears in the heading of the chart. Similarly, in hospital discharge summaries the words *admission* and *discharge* are associated with a known time, provided that they refer to the current admission. Therefore, given an expression such as *birth weight* or *admission haematocrit*, the time of the associated event (*weight*, *haematocrit*) can be equated with the appropriate date (provided that the statement refers to the current patient or current admission).

The events of admission and discharge are of particular importance in the hospital setting; they divide time into five distinct stages: (1) before admission; (2) at admission; (3) after admission and before discharge (during hospitalisation); (4) at discharge; and (5) after discharge. Many retrieval questions can be answered simply on the basis of this very general time information. To facilitate retrieval, each event is tagged with a 'general reference point' code (abbreviated GEN REF in *Figure 10.5*), to indicate to which of these periods or time points it belongs.

[*] However, if the subordinate clause had had explicit time information (for example, *she became ill when she went to Maryland in 6/74*) then the time of the main clause (*she became ill*) would have been expressed using the subordinate clause *when she went to Maryland in 6/74*) as a reference event. The subordinate clause itself would have used the date *6/74* as its reference event.

TEXT Date of admission: 9/6/74; Date of discharge: 10/9/74

This is the first admission for this 72 year old woman with presumed adenocarcinoma of the lung. The patient was well until 6/74 when she had an episode of severe dyspnea while in Maryland. She was admitted to a local hospital where she was found to have a pleural effusion. She underwent pleural biopsy and cytology which showed adenocarcinoma. In 7/74, she had 5 millicuries of P32 instilled into the left pleural cavity.

EV#	Connective	Regularized text (event)	Relation	Ref. point	Adjust.		Gen. ref.
E1		This	At	E2			0
E2	be present	the 1st admission for this 72 year old woman	At	Adm			0
E3	rel-clause	(woman) with presumed adenocarcinoma of the lung	At	E2			0
E4		The patient be past well from ( ) to 6/74	From To	Adm E5	--		−1 −1
E5	until	the patient be past [not] well (at) 6/74	At	6/74			−1
E6	when	she past have an episode of dyspnea	At	E5			−1
E7	while	[she be] in Maryland	At	E6			−1
E8		( ) admit past she to a local hospital	At	E5	0/+		−1
E9	where	( ) find past	At	E8			−1
E10	embedded	she have a pleural effusion	At	E9			−1
E11		She undergo past pleural biopsy	At	E8	0/+		−1
E12	and	cytology	At	E11	0/+		−1
E13	which	[cytology] show past adenocarcinoma	At	E12			−1
E14		In 7/74 she have past 5 millicuries of P32	At	7/74			−1
E15	rel-clause	( ) instill [P32] into the left pleural cavity	At	E14			−1

Gen. ref. = General reference code: −1 = before admission; 0 = at admission.

Figure 10.5 Time representation for a paragraph from a hospital discharge summary

10.2.5 Multiple references

The words *admission* (or *admit*), *discharge*, *hospitalisation* and their synonyms can be associated with dates —but only provided that they refer to the current admission. For example, in the first sentence of the paragraph shown in *Figure 10.5*, the word *admission* occurs:

> This is the first admission for this 72 year old woman with presumed adenocarcinoma of the lung.

With the aid of the present tense marker on *is*, the program establishes that *admission* here does refer to the current admission (since the present tense indicates time = *TIME OF NARRATION*, namely the time of the current hospitalisation). On the other hand, the verb *admit* occurs in sentence 3 of the same paragraph:

> She was admitted to a local hospital where she was found to have a pleural effusion.

In this sentence *admitted* does not refer to the current hospitalisation. The program detects this by noting that the institution (INST = *a local hospital*) occurs with the indefinite article *a* and therefore cannot refer to the hospital of the current admission, which has already been identified (in the header information).

This raises the question of identity of reference in determining the time relations in narrative. (The issue of identity of reference is, of course, also important to other aspects of narrative information.) Whenever there are two mentions of an event, it is important to determine whether these mentions are references to the same event (and therefore have the same time) or whether they refer to different events, as in the example above with *admission* and *admitted*. In the sequence below, a failure to determine identity of reference would produce a representation with two separate transfusion events, and no coherent time relations between them:

> A transfusion of 100 cc packed red blood cells was given. After completion of transfusion, patient lapsed into a coma.

A general solution to the problem of identity of reference would require that each event be checked as it is entered in the database, to determine whether it was an additional reference to a previously mentioned event. In normal English discourse the use of the definite article *the* is used to signal the fact that a given event has already been introduced. In the medical narrative such a search is made more difficult, since articles, both definite and indefinite, are routinely omitted. The program as currently implemented has not undertaken the complete referential analysis of each document. It has been sufficient in most cases to limit analysis of referential expressions to referentials occurring in explicit time expressions — that is, expressions such as *since admission*, *after the transfusion*, etc.

There are four types of search for a reference occurring in a time expression. The first is the search for the 'current' event of the appropriate type. The two-sentence sequence given above provides an example of this. The first sentence contains one event (E1: ( ) *give transfusion*...); the second sentence contains two events: the main event (E2: *patient lapsed into coma*) and the event *transfusion* contained in the time expression *after the transfusion* (*see* example 4, *Figure 10.4*). This event is given independent status and a separate event number (E3) in the representation. The time of E2: *patient lapsed into coma* can then be expressed with E3 being used as a reference event. Because *transfusion* in the second sentence occurs in a time expression, the program performs a referential search to determine whether *transfusion* refers to the most recently mentioned (current) transfusion. To do this, the program searches backwards looking for an occurrence of *transfusion* or a synonym; it finds one in the preceding sentence. These two occurrences of *transfusion* are taken to refer to the same event, since neither occurs with a time adjective (for example, *last*) and since there is no indefinite article *a* on the second occurrence. The time information that results from this identification of reference is recorded by using the event number of the first mention of transfusion in the first sentence (E1) as the reference point for the time of the second mention of *transfusion* in E3; this gives E3 the time of E1 (the two *transfusion* events have the same time).

The second type of search is for the 'previous' event of the appropriate type, as in the phrase *after the last hospitalisation*. This requires a search for the mention of the hospitalisation which occurred before the current hospitalisation. A complete search strategy for this would require:

(1) The identification of the 'current' hospitalisation.

(2) The retrieval of all hospital events entered.
(3) A comparison of all their times, to select the one with a time closest to but still less than the time of the 'current' hospitalisation.

Since this is a very costly procedure, a simpler strategy was implemented; this strategy locates the 'current' hospitalisation, then traces backwards through the events entered in the database to find the previous reference to hospitalisation. If this has a time less than the 'current' hospitalisation, it is accepted as the 'previous' hospitalisation.

Third, there is a search for the 'next' event, as in the sequence: *All his blood cultures drawn on admission grew out pneumococcus. Subsequent blood cultures were negative.* Again, the 'current' event of appropriate type is located in the first sentence, and the time of the 'next' event is then taken to be *after* that of the current event, owing to the presence of the word *subsequent.*

Finally, there is the search for the '*n*th' event, as in *after the first admission.* Here, in order to locate the desired event, it is necessary to go to the beginning of the section or paragraph and search forwards, counting events until the desired number *n* is found. (A more sophisticated search would require that all events of the appropriate type be found, and ordered with respect to time, before the *n*th event was selected.)

10.2.6 Change of state

Certain words express a complex time relationship by collapsing two events into a single assertion of a change of state. For example, the use of the preposition *until* in the sentence *she was well until 6/74* provides information about an earlier state, as well as the current (changed) state. The program expands the single assertion into two distinct events, as indicated below:

(E1) She was well during a time up to time of E2.
(E2) She was not well in 6/74.

where E1 and E2 are 'event numbers', used to distinguish events (*see* lines E4 and E5 of *Figure 10.5*). There are other 'change of state' words which require a similar expansion into two events to capture both the information content and the implicit time information. These words are 'change' words and 'termination' words, such as *become* or *subside.* For example, the program expands *pain seemed to subside* into two distinct events:

(E1) Pain existed before time of E2.
(E2) Pain seemed to subside (i.e. no pain at E2).

Note that the time of the original state (E1) is represented as occurring before the time of the changed state (E2). *See* example 5 of *Figure 10.4* for a more complete representation.

10.2.7 Narrative time progression

To augment the limited explicit time information available, the time program makes an assumption that time moves in a forward direction unless there is information to the contrary. Given the sequence:

(E1) She was admitted to the hospital.
(E2) She underwent a biopsy.

the program records that the time for undergoing the biopsy is later than or equal to the time of admission to the hospital. This assumption captures our intuition about how time works in a narrative. The relationship is represented by taking the first event as a reference point for the time of the second event, adjusted by the DIRECTION '0/+' (greater than or equal to). This is shown for events E8 and E11 in *Figure 10.5.*

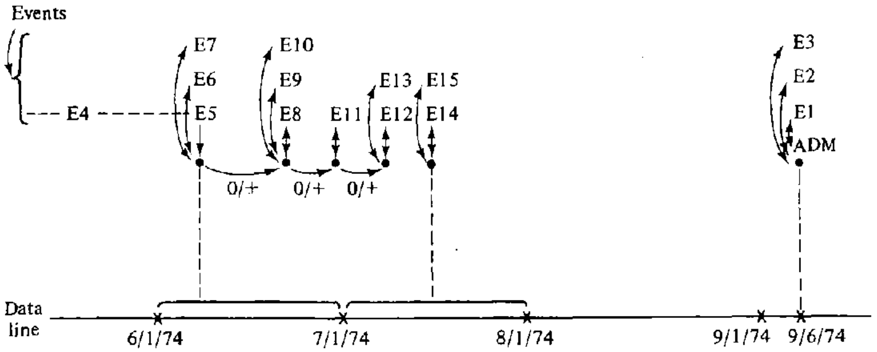
A co-ordinate conjunction can also indicate narrative time progression when it connects two assertions or verb phrases; for example, in the sentence *She developed seizures and was given phenobarbital and valium* the time of *given phenobarbital and valium* is later than the time of *she developed seizures.* However, when the co-ordinate conjunction connects two noun phrases (for example, *phenobarbital and valium*), then the time program assigns the same time to the events *given phenobarbital* and *given valium.* Conjoined noun phrases give rise to two distinct events (here *give phenobarbital* and *give valium*), because conjunction is regularised during the natural-language processing. This is done so that no row in the table contains a conjoined event — each column entry contains only a single piece of information.

The need for the specification of a direction 'greater than or equal to' for narrative time progression can be seen in the following example: *Blood gases were monitored throughout the next few days and gradually returned to normal.* Here the two conjoined events (*blood gases were monitored* and *[blood gases] returned to normal*) take place within the same time period, rather than at successive times. Further research is needed on how to distinguish these cases of concurrency from cases of actual time progression.

In addition to identifying the sources of time information in natural language, the time program must co-ordinate these sources, since it often occurs that a single event has several pieces of time information associated with it. As an example, the first sentence from the sample text of *Figure 10.1* (*This is the first admission for this 72 year old woman*) has three sources of time information. First, the tense of the verb *be* indicates that the time is the TIME OF NARRATION. Second, the special word *admission* refers to the current admission and therefore has associated with it the date of admission (9/6/74) as its reference point. The third source comes from the age expression *72 year old*, which, when added to the patient's birth date, gives another date as reference point.

The present implementation of the time program preferentially uses the most specific time information available. First it processes adverbial time expressions, multiple references and special events with known time. If none of these sources produces a date or a previously mentioned event as a reference point, then the remaining sources (connectives, tense and narrative time progression) are processed (in that order) to produce a reference point. Change of state expressions are expanded after the other time processing has been completed.

In some cases the program may combine information from several sources. In the first sentence of the sample paragraph of *Figure 10.1* the program resolves the reference of *admission* as referring to the current admission by checking the tense of the verb (present tense implies time of narration); this yields the date of admission as the reference point. The date of admission is used in the representation because it is more specific than the time calculated on the basis of the patient's age. At present, however, there is no consistency



Identification of events:

E4 be well	E8 Admit	E12 Cytology	E1 This
E5 be not well	E9 ( ) Find	E13 Show adenocarcinoma	E2 Admission
E6 have dyspnea	E10 Have effusion	E14 Have P32	E3 With adenocarcinoma
E7 be in Maryland	E11 Undergo biopsy	E15 Instill P32	

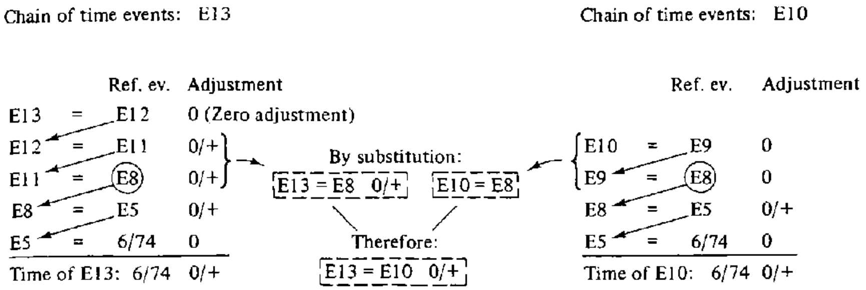
Figure 10.6 Schematic representation of time relations in the text of *Figure 10.1*

check between multiple sources of time information in a single sentence; the most specific information is used in the time representation, and the rest of the information is ignored. We plan to implement consistency checking in the next version of the program.

10.3 Computation of time relations

From the analysis of time relations, the time program obtains a representation of time for each narrated event, either in terms of a fixed external point (a date or a 'special event' with known time) or in terms of another event in the narrative. The output of the time program (events with their time representations) serves as the input to the retrieval program. The retrieval program has a special time comparison retrieval routine which compares the time information for two events and returns one of four values: greater than, less than, equal (including inclusion of one event within the other) or not comparable (that is, there is not sufficient information to calculate a relative ordering for the two events). The time comparison routine permits of the use of time relations in retrieval requests.

The time information obtained from the paragraph in *Figure 10.5* is shown schematically in *Figure 10.6*. The events E1, E2 and E3 are obtained from the first sentence (*This is the first admission for this 72 year old woman with presumed adenocarcinoma of the lung*); these events are associated with the time of admission, 9/6/74. The narrative then jumps back in time with the next sentence: *She was well until 6/74 when she had an episode of severe dyspnea while in Maryland.* This sentence contains E4: *she was well* (at a time before 6/74 up to E5). The narrative then proceeds in a forward direction for E5 (*not well*), and both E6 (*when she had an episode of severe dyspnea*) and E7 (*while in Maryland*) occur during this time. Although these events are anchored to the date 6/74, this expression actually represents a time 'point' within the month defined by



Time of E10 (effusion) is less than or equal to time of E13 (adenocarcinoma)

Figure 10.7 Computation of time for E13 (adenocarcinoma) and E10 (effusion) and their relative time

the date *6/74*. This is represented in *Figure 10.6* by the horizontal line running parallel to the month of 6/74. The dotted vertical line locates the time point at an unspecified place within the month.

The next sentence moves forward by narrative time progression: *She was admitted to a local hospital* (E8) *where she was found* (E9) *to have a pleural effusion* (E10). Events E8, E9 and E10 are found to occur at the same time, because of their connectives and because there is no information to the contrary. The next sentence also moves forward twice by narrative time progression, once for each conjunct: *She underwent pleural biopsy* (E11) *and cytology* (E12) *which showed adenocarcinoma* (E13). In this case the co-ordinate conjunction connects two separate sentential elements and is therefore treated as indicating narrative time progression (*see* sub-section 10.2.7, above). E13 has the time of E12, because the two events are connected by a non-temporal connective (a relative clause), and there is no specific time information to indicate otherwise. Finally, the last sentence is anchored to a date (7/74): *In 7/74, she had 5 millicuries of P32* (E14) *instilled into the left pleural cavity* (E15). Because of a peculiar parse obtained during the natural-language processing stage, this sentence is treated as two events (*she had P32* and *P32 was instilled*), although, in fact, it should be treated as a single event.

For purposes of computation, the time information can be seen in terms of 'time chains' which link together sets of events. The time chain for E13 (*cytology showed adenocarcinoma*) is shown on the left-hand side of *Figure 10.7*: each event in the chain is given in terms of the time of another event, adjusted by the appropriate amount.

In order to answer a question such as *When was cancer first detected*? the retrieval program performs the following steps. First, it scans the database (as represented in *Figure 10.2*) for unnegated entries under DIAG(nosis) of a synonym of *cancer*. It finds two such entries, one in the first sentence (line 3 of *Figure 10.2*, also shown as E3 in *Figure 10.5*), *woman with presumed adenocarcinoma of the lung*; and one in the fourth sentence *cytology showed adenocarcinoma* (line 12 of *Figure 10.2*, also shown as E13 in *Figure 10.5*). The time comparison routine is then invoked to determine whether E3 or E13 is earlier. In this case the calculation is very simple, because the time program assigned E3 a GENERAL REFERENCE (GEN REF) code of 0 (time = at admission), while E13 has a GEN REF = −1 (time = before admission). On the basis of this very simple check, E13 is found to be the earlier event, and therefore represents the first reported detection of cancer.

The time for E13 is calculated by tracing along the time chain created by the reference points until a fixed point (date or event with known time) is reached. As the routine traces back along this chain, it accumulates the sum of the successive adjustments of each event on the chain. In the case of E13 the chain consists of the events E13, E12, E11, E8 and, finally E5 = 6/74. The adjustments are three occurrences of the inequality 0/+ (greater than or equal to); the sum of three occurrences of 0/+ 'yields 0/+ as the total adjustment. Therefore, as *Figure 10.7* shows, the time for E13 is 6/74 0/+ (cancer was diagnosed first at a time later than or equal to 6/74)*.

The second retrieval question posed in Section 10.1 was *What signs or symptoms preceded or coincided with the diagnosis of cancer*? Again, a retrieval routine locates all unnegated entries in SIGN-SYMP(tom). Two candidate events are located, namely *dyspnea* (E6 in *Figure 10.5*, also shown in line 5, *Figure 10.2*) and *effusion* (E10 in *Figure 10.5*, also shown in line 9 of *Figure 10.2*). For each of these events, the time comparison routine compares its time with the time of E13 (the first diagnosis of cancer). In this case no easy comparison is possible, since all three of the events have a GEN REF = − 1. Therefore, the time comparison routine must look at the time chains for the events.

Using the method of calculating time described for E13, the time program finds that the time for E10 is also 6/74 0/+ (right-hand side of *Figure 10.7*). If the time program stopped at this point, E10 and E13 would be 'not comparable' with respect to each other, since they are both greater than or equal to the same date, 6/74. However, to circumvent the accumulated imprecision resulting from adding inequalities, the retrieval routine compares the two time chains for E10 and E13, to determine whether the chains share a common segment. As *Figure 10.7* shows, the chains merge at E8: the time for both events E10 and E13 is given in terms of their relation to E8. By substitution, E13 is greater than or equal to E8 and E10 is equal to E8. Therefore, by a second substitution, E13 is greater than or equal to E10, and the time comparison routine successfully determines that *effusion* (E10) precedes or coincides with the first diagnosis of cancer (E13).

For E6 (*dyspnea*) a similar calculation also gives the result that the time of E13 is greater than or equal to the time of E6, so that the retrieval program also finds that *dyspnea precedes or coincides with the first diagnosis of cancer*. By comparing the chains as well as the actual calculated dates, the time comparison routine utilises the available information, even when it is imprecise (such as the inequality derived from narrative time progression).

10.4 Implementation

The time program and the time comparison routine are implemented in an extension of University of Texas LISP 4.0 on a Control Data 6600. The

* In fact, this range should have an upper bound of 7/74. In its present implementation, however, the time program does not make use of this fact (derivable from narrative time progression), because each event is allowed only a single reference point. The next implementation of the program will allow multiple reference points for each event; this would permit the program to capture the upper bound for events E8 to E13.

program requires approximately 50 000 words of memory. The time program took approximately one minute to process the sample paragraph shown in *Figure 10.1.* Retrieval for the two sample questions took approximately 30 seconds.

10.5 Conclusion

The time program described here has been used in a larger context to retrieve information from hospital discharge summaries (Hirschman *et al.*, to be published). Specifically, it has been used to retrieve information about the patient's state at various times during the hospital stay (for example, whether the patient had a fever at admission and/or at discharge), as well as information about the performance of certain medical procedures and their results (for example, whether a chest X-ray was taken; whether the X-ray showed improvement during the patient's stay). A comparison of the machine-generated responses for three discharge summaries showed a 90 per cent agreement with the results obtained (for the same documents and the same set of questions) by a physician reviewer. An analysis of the discrepancies in the human-generated and machine-generated results showed that only one of the errors was due to incorrectly processed time information*.

Although the time program has been successfully tested in one application in the medical area, it remains to be more extensively tested both in the medical domain and in other domains. In addition to further testing, a number of extensions to the program are planned.

The present implementation performs computation of relative ordering of events as this information is needed during retrieval. To compare two arbitrary events, the program uses the transitivity of the inequality relations in the representation to obtain an ordering of the one event to the other. Originally this was done for efficiency, since the relations between most pairs of events are not of interest. However, if two events are compared more than once during the course of several different retrieval requests, their relative ordering must be recomputed each time. We are now developing an implementation which will compute the transitive closure of the partial ordering for all the events in the narrative and maintain this information in a compact form. This means that the relative ordering for any pair of events will be immediately available, with no further computation. The existence of this complete set of time relations will make it possible to implement a full treatment of referential expressions. With the complete set of time relations, each event entered in the database can be readily checked for identity of reference with all previously entered events. (At present, only referential expressions appearing in time adverbials undergo resolution of reference.) This information would also facilitate carrying out consistency checks; these are currently cumbersome to perform, because time relations must be recalculated each time they are used. Finally, it will also make it possible to

* The time preposition *for* was not correctly handled as indicating duration. As a result, in the sentence *no fever for 3 days before discharge*, *no fever* was assigned a time of DISCHARGE — 3 DAY, instead of a time duration which lasted until (and including) discharge. Therefore, the retrieval was not able to determine that the patient was afebrile at discharge.

represent more than one time relationship for a given event, so that the problem of capturing the 'upper bound' in narrative time progression can be resolved (*see* the discussion in Section 10.3).

A number of extensions are also needed for the representation of imprecision or 'fuzz'. Fuzz itself has several aspects, none of which have been handled in the present implementation of the time program. One type of fuzz results from the use of time 'points'. Of course, any time 'point' is an approximation with an implicit imprecision of measurement. For example, the time specification *9/6/74* specifies a 'point' within a period of 24 hours; the specification 9/74 specifies a 'point' with a period of one month, etc. This implicit uncertainty should be recorded along with the rest of the time information and preserved during any calculations.

Another type of fuzz arises in specifications such as *several* or *a few*. Fortunately, these expressions occur rarely in the medical material. Nonetheless, they must be handled if the time program is to extract time information from narrative in other domains. The 'fuzz' from narrative time progression is also a problematic area. More research is needed to determine when two events are concurrent and when they are sequential. Where there is progression of time, the unit of time progression seems to be defined by the context; for example, if the patient's state is being reported every several days, then the unit for narrative time progression appears to be the day (or a smaller unit), rather than a month or a year. Further study of the areas of resolution of multiple references and computation of imprecise time information should not only provide a more sophisticated and accurate time program, but also produce new insight into the organisation of discourse and the principles of discourse coherence.

Acknowledgements

The LISP programming for both the time program and the time retrieval routine was done bu Guy Story, who also contributed many valuable criticisms to the work described here.

This research was supported in part by the National Library of Medicine grant number LM02616, awarded by the National Institutes of Health, DHEW; and in part by the National Science Foundation under grant number IST-7920788 from the Division of Information Science and Technology.

References

BRUCE, B. C. (1972). 'A model for temporal references and its application in a question answering program', *Artificial Intelligence*, **3**, 1–25

FINDLER, N. and CHEN, D. (1973). 'On the problems of time, retrieval of temporal relations, causality and coexistence', *International Journal of Computers and Information Science*, **2**(3), 161–186

HIRSCHMAN, L. and GRISHMAN, R. (1977). 'Fact retrieval from natural language medical records', in SHIRES, D. B. and WOLF, H. (Eds), *Proceedings of the Second World Conference on Medical Informatics*

(*MEDINFO '77*) IFIP World Conference Series in Medical Informatics, Vol. 2, pp. 247–251, North-Holland, Amsterdam

HIRSCHMAN, L., STORY, G., MARSH, E., LYMAN, M. and SAGER, N. (to be published). 'An experiment in automated health care evaluation from narrative medical records', *Computers and Biomedical Research*

KAHN, K. and GORRY, G. A. (1977). 'Mechanizing temporal knowledge', *Artificial Intelligence*, **9**, 87–108

SAGER, N. (1978). 'Natural language information formatting: the automatic conversion of texts to a structured data base', *Advances in Computers*, **17**, 89–162