Part IV
Temporal Annotation

Introduction to Part IV

1 INTRODUCTION

The work presented in earlier sections of this book has, by and large, followed the methods of analytical linguistics, philosophy, and symbolic AI. That is, aspects of how temporal information is conveyed in language have been analyzed and modeled in formal (in varying degrees) or computational systems. The emphasis, however, has been on the analytical approach or model, and not on the data being analyzed. Native speaker intuitions both about acceptability and about the nature of the phenomena to be studied are taken as sufficient to ground the work in 'real' language. This methodology tends not to be explicitly defended; the presumption is that while theories or models remain unable to account for obvious usage, there is no need to go hunting in data for new challenges.

In the last decade, however, there has been a growing movement in both computational and noncomputational linguistics to place more emphasis on the study of real language as evidenced in collections of texts or **corpora**. Linguistic judgments about these texts are then recorded in the form of **annotations** associated with the texts, sometimes called **metadata**, which serve to label or flag phenomena of interest in the texts. The reasons for this shift in emphasis in the field are complex, and this is not the place either to defend this shift (see, for example McEnery and Wilson (1996) for strong advocacy for this move in linguistics) or to engage in speculation as to why it has come about. However, we can cite some of the clear benefits it brings:

- Corpus-guided research reveals both the variety of forms of expression in real samples of language and the distribution of these forms of expression. The former is important as it may direct attention to forms that the intuitions of armchair linguists have either overlooked or ruled out. The latter is important to help focus efforts of algorithm builders on the most frequently occurring cases, to aid in the construction of probabilistic models, and to guide psycholinguistic or cognitive psychological research.
- Annotation schemes, together with corpora annotated according to them, provide an objective data resource that can be shared, argued over, and refined by the (computational) linguistic community. Comparisons between human annotators annotating the same data according to the same scheme can be used to decide how well-defined and comprehensive the scheme is.
- Annotated corpora are resources that can be exploited by machine-learning algorithms to acquire annotation capability without the necessity of implementing an underlying analytical model.
- Annotated corpora provide an objective basis to evaluate competing algorithms.

This general movement towards corpus-based research has not been without effect on the study of temporal phenomena in text, and the papers in this part reflect this work. Unlike work in the previous sections, some of which dates back over fifty years, work on the annotation of temporal phenomena is quite recent, all of it less than ten years old, much of it more recent than that. As a consequence, no consensus about how temporal information in text should be annotated can yet be said to have emerged, though the collaboration of several research groups to produce the TimeML annotation scheme, described by PUSTEJOVSKY et al. (Chapter 27) is an ambitious attempt to provide a comprehensive approach which subsumes earlier efforts.

As an introduction to the papers in this part, the following provides a background to the issues these papers address. These issues include what to annotate and the status of automatic annotators, the process of and tools to support manual annotation, and the existing resources which are outcomes from annotation work.

The programme of defining appropriate standards for temporal annotation and then building algorithms that can carry out this annotation automatically is an exciting one and one which is only partially complete. The coming years will see to what extent it can be realized and what the impact of temporal annotation capability will be on applications of language technology.

2 ANNOTATING TEMPORAL REFERRING EXPRESSIONS

The most obvious temporal feature to annotate in texts, and the one which historically was addressed first, is **temporal referring expressions** (as found in temporal adverbials, for example); that is, expressions which refer to **times** (*July 1, 1867*), **durations** (*three months*), or **frequencies** (*weekly*). Being able to identify and distinguish these types of expression is crucial to being able to situate the events described in text either absolutely in terms of some conventional calendrical time frame or relatively with respect to other events.

The examples just given perhaps understate the complexity of the phenomena to be addressed. When devising an annotation scheme to capture temporal referring expressions one must deal with a variety of complications:

(1) *indexicals*: expressions like *now, yesterday*—and other contextually dependent expressions such as partially specified calendrical times (e.g. *Wednesday*—which Wednesday?) or relatives such as *next week, three weeks ago*, all of which depend for their interpretation on knowledge of a deictic centre;
(2) *relational expressions*: expressions which explicitly specify times in relation to other times (*two weeks after Christmas*) or to events (*5 seconds after the first explosion*); and
(3) *vagueness*: expressions referring to times whose boundaries are inherently vague (*spring, evening*) or which contain modifiers which blur the time reference (*several days ago, sometime after 7 p.m.*).

Work to devise annotation schemes for temporal referring expressions appears to have begun as part of the Named Entity (NE) tagging subtask within the DARPA Message Understanding Conference (MUC) series of evaluations, specifically in MUC-6 (MUC 1995). In this task participants' systems were to tag (by inserting SGML tags into running text) expressions which named persons, organizations, locations, dates, times, monetary amounts, and percentages. A key part of this exercise was that a set of texts

was manually tagged by human annotators to provide a 'gold standard' measure of correctness. Metrics, principally the **recall** and **precision** metrics adapted from information retrieval research, were used to compare system-supplied annotations (or **responses)** against human-supplied annotations (or **answer keys**). Recall, the proportion of the answer keys for which a correct response is supplied, is a measure of coverage or completeness of a system; precision, the proportion of responses which are correct, i.e. match the answer key, is a measure of correctness or soundness of a system.

In MUC-6 date and time (of day) expressions were labeled using a `TIMEX` tag. Only **absolute** time expressions were to be annotated, i.e. expressions which indicated a specific minute, hour, day, month, season, year, etc. **Relative** time expressions (e.g. *last July*) were excluded, though subexpressions within them (e.g. *July* in this example) were to be tagged. A set of thirty manually annotated newswire texts were used for a blind evaluation. The top scoring automated system scored .97 recall and .96 precision on the TIMEX tagging task.

In MUC-7 (MUC 1998) the principal change was to capture relative as well as absolute date and time expressions, though the two did not need to be distinguished in the tagging. Thus indexicals, such as *yesterday, last July*, were to be marked, as were so-called 'time-relative-to-event' phrases such as *the morning after the July 17 disaster*. For the final blind evaluation a set of 100 tagged texts was used and the highest scoring system scored .89/.99 recall/precision on the date tagging task and .81/.97 recall/precision on the time tagging task.

One of the principal limitations of the date and time NE task in both MUC-6 and MUC-7 is that while identifying temporal referring expressions in text is useful, what is really needed is the ability to interpret or evaluate or dereference these expressions to obtain the time they denote. Thus, according to the MUC-7 TIMEX tagging guidelines, an expression such as *yesterday* in an article datelined *June 12, 1998* would be tagged as a TIMEX of type `DATE`. However, what applications really need is the knowledge that in this context *yesterday* refers to *June 11, 1998*. This requirement is addressed by the TIMEX2 tagging guidelines, reviewed by Wilson et al. (Chapter 23). Interpretation is handled by adding the full calendrical time value for every temporal referring expression as an attribute of the tagged element, using an ISO standard time format as the attribute's value. Wilson et al. also describe an implemented tagger which annotates newswire text (in English and Spanish) with TIMEX2 tags with impressively high scores, achieving 96.2 f-measure (a combined measure of recall and precision) for tagging surface expressions and 83.2 f-measure in interpreting them.

The ability to evaluate a relational or indexical time expression, returning a calendrical time value, is clearly needed as part of the temporal interpretation process. However, there is utility in separating the evaluation process into two stages, first mapping the time expression into a semantic representation in the form of a functional expression, and second evaluating the functional expression. So, for example *last Thursday* might in the first stage be mapped into the expression `thursday(predecessor(week DCT))`, where DCT is the document-creation time of the article and in the second stage an absolute calendrical time is computed from this latter representation given the DCT. This separation of semantic interpretation from full evaluation has number of advantages. It fosters discussion of the correct semantic interpretation of complex temporal referring expressions, it permits separate evaluation of the two stages (an algorithm could be good at working out the semantics of *last* expressions, but bad at finding their anchors), it allows unevaluated semantic representations to be made available to other interpretation components which may require them rather than their values, and it

permits taggers to defer the evaluation of temporal functions until their values are actually required. Pustejovsky et al. propose an extension of the TIMEX2 standard to include temporal functional representations, and call the extended standard TIMEX3 (TIMEX3 includes a number of other refinements to the TIMEX2 standard, but this is the most significant).

Most of the work described above has been driven by the English-speaking research community, though as noted TIMEX2 has been applied to English and Spanish, and recently to Korean, French, Chinese, and Hindi. However, SCHILDER AND HABEL (Chapter 26) independently propose an approach for annotating German newswire texts which aims to capture the same sort of temporal referring expressions as the TIMEX2 and 3 standards. Their tagger outputs a semantic representation of relative time expressions which are evaluated in a subsequent stage, making its handling of these expressions similar to that proposed in TIMEX3.

3 ANNOTATING EVENTS AND STATES

To interpret a text temporally means not just identifying the times, durations, and frequencies mentioned in a text; it means positioning the events and states described in the text with respect to these times and to each other. However, before it is possible to discuss how to annotate relations between events, states, and times, agreement must be reached on how to annotate events and states themselves. To do this in turn requires making decisions about (a) what we are trying to annotate—just events? events and states? and what do we take the difference to be? (b) how events/states are realized in text; (c) what textual representative of the event/state will be annotated; (d) what attributes should be associated with annotated events/states.

3.1 What semantic types to annotate

To answer the first of these questions requires taking some position with respect to questions of event ontology raised in several places in the Introduction to Part I in this volume. At the most general level, temporal annotation can be taken as the task of correctly annotating the temporal position of all temporal entities in a text, i.e. of all *things that happen or are situated in time*. If, for purposes of the following discussion, we assume a top-level ontological class of eventualities or situations which is divided into events and states (cf. Introduction to Part I, 3.1.2), this would mean annotating all events and states.

Such a task is daunting, and since practical applications are primarily concerned with events, it might appear reasonable to start out with the more modest aim of annotating events, but excluding states. However, drawing a firm *conceptual* distinction between events and states is not straightforward, as the discussion in Part I has shown. One common distinguishing test is the so-called subinterval property (Dowty, Chapter 15): for any state p that holds over an interval t, p must hold for every subinterval of t. However, this is not a particularly easy test to apply and not one to expect annotators of texts to be able to carry out efficiently or effectively.

A second way to distinguish events and states, also discussed at greater length in Part I, is via *linguistic* tests. States tend to be expressed via constructions with the copula, or via certain verbs such as *have, know, believe*. This is a perhaps a more practical approach in

the context of producing realistic guidelines for annotation. If the point of making the distinction is to capture genuine semantic differences between events and states, however, then this approach depends on determining an accurate and complete set of linguistic correlates for states.

Most approaches to event annotation reported in this part, however, do not attempt to make a distinction between events and states. In general, the approach is to treat all verbs as expressing temporal entities suitable for tagging. This 'lumping' together assumes that the distinction is not important, or is too difficult, for purposes of annotation. While dismissing the problem in the short term, this ignores the fact that there are genuine semantic differences between events and states, and that these have consequences in terms of the inferences that can be drawn and the likely questions that can be asked concerning each. For example, states typically invite questions about when they began, ended, and how long they lasted; events invite questions about when they happened, but not so typically about their duration. Furthermore, the process of positioning states in time may differ from that of positioning events, so that an algorithm that attempts do this positioning automatically would need to know which it was dealing with.

The only work in this part which does propose to distinguish events and states and to annotate both is that of Pustejovsky et al. Note, however, that they treat states as a subtype of events—effectively identifying events with what we have here termed eventualities. In fact they go further than simply distinguishing events and states, and propose distinguishing seven types of events in their annotation scheme, two of which are stative and all of which are held to have distinctive temporal significance. Their distinguishing criteria, as presented, are primarily linguistic, though concerning states they do appeal to something like the subinterval property cited above. Further, they do not propose to annotate all states: they propose to annotate only those states which 'are directly related to a temporal expression . . . including those states that identifiably change over the course of a document'.

3.2 The linguistic realization of events and states

To date then, the work on temporal annotation of 'events' in text has not worried overly about the semantic distinction between events and states and has assumed that the 'things which are situated in time' which need to be annotated can be identified via a set of syntactic or lexical linguistic criteria.

Katz and Arioso (Chapter 24), for example, define their task in a (deliberately) restrictive way: 'The temporal interpretation of a sentence, for our purposes, can simply be taken to be the set of temporal relations that a speaker naturally takes to hold among the states and events described by the *verbs* of the sentence' (italics added). Thus, for example, event nominals such *destruction, election, war* are excluded, as are, presumably, stative adjectives such as *sunken*. However, their investigation is exclusively concerned with sentence-internal temporal relations and they are not aiming to position every event or state reference in time, or in relation to another event or state.

Filatova and Hovy (Chapter 25) take the locus of events to be syntactic clauses which contain a subject (one or more noun phrases) and predicate (verb phrase with one or more verbs), as output by a specific parser. Their concern is to time-stamp these clauses, that is, to associate a calendrical time reference with each clause. They too, ignore, event nominals and stative adjectives. However, again, they are not aiming at complete temporal interpretation, but at a more limited task.

Schilder and Habel have a broader target. They identify two types of event-denoting expressions: sentences and event-denoting nouns, especially nominalizations. The most inclusive treatment is that of Pustejovsky et al., who consider events expressed by tensed or untensed verbs, nominals, adjectives, predicative clauses, or prepositional clauses.

3.3 Textual representatives to annotate

Once a set of linguistic signals for events has been decided there is still the issue of deciding precisely what text spans will be annotated, i.e. what will count as the textual representative of the event. For the most part this follows straightforwardly from decisions made about the linguistic realizations of events and states. However, those decisions do not entirely specify the annotation.

Concerning events conveyed by clauses containing verbs, one could decide that the entire clause is the appropriate span to be annotated. This is the position taken by Filatova and Hovy. Or, one could decide to annotate just verb groups or just the heads of verb groups. This latter approach has been adopted by the other authors in this part, perhaps because it simplifies matters when dealing with embedded clauses or clauses with multiple verbs (Filatova and Hovy acknowledge problems with their approach for cases of co-ordinated verb phrases where the verbs have different tenses).

3.4 Attributes of events and states

As well as tagging a text span as event representative, some approaches chose to associate attributes with the event. In Schilder and Habel's approach, for example, each event has a `sem` attribute that holds a predicate-argument representation of the event. It also has a `temp` attribute whose value is triple consisting of a binary temporal relation, the time id of the event itself, and the id of a time related to the event time by the temporal relation. This attribute gets its value computed as part of the interpretation process.

These event attributes are effectively part of Schilder and Habel's implementation of a computational mechanism to assign times to events. Another sort of information that can be associated with events is descriptive linguistic information which may be of use during the interpretation process. So, for example, Filatova and Hovy make use of tense information associated with event clauses by their parser. Pustejovsky et al. associate tense, aspect, and subtype information with events. The event subtypes they propose are: occurrence (*crash, merge*), state (*on board, love*), reporting (*say, report*), i-action (*attempt, offer*), i-state (*believe, want*), aspectual (*begin, stop*), and perception (*see, hear*). These classes are distinguished because of the distinctive sorts of temporal inferences that may be drawn for events within them.

3.5 Automated event tagging

In the foregoing we have discussed *what* is to be annotated when annotating events or states. Now we briefly discuss the state of play with implemented systems that do event tagging. Of the papers in this part only three describe implemented systems that do event tagging: Filatova and Hovy, Schilder and Habel and Li et al. However, for none of these researchers is event tagging itself a goal—rather they are aiming to anchor

events in time and possibly also to relate events to each other temporally (Li et al.). Only Filatova and Hovy provide separate evaluation results for their system's ability to recognize events—in their case the ability to recognize clauses, since for them clauses are the textual representatives of events. They report figures of around 61 per cent recall and 56 per cent precision, errors being due in part to the parser they use and in part to their shallow algorithm for extracting clauses from the parse tree. As noted the others do not evaluate event recognition separately from temporal relation annotation.

4 ANNOTATING TEMPORAL RELATIONS

Given an approach to annotating temporal referring expressions and event/state denoting expressions, the next challenge for a programme of temporal annotation is to establish conventions for annotating the **relations** between times and events or between events and events (from now on we will use the term 'event' loosely to refer to events and possibly to states as well, making clear if necessary where remarks may only pertain to states or to nonstative eventualities).

4.1 Annotating relations between times and events

Time–event relational information may be conveyed in a variety of ways. The most explicit route is via a prepositional phrase in which a preposition signals a relation between a temporal referring expression (the complement of the phrase) and an event denoting expression (typically a verb or an event nominal modified by the phrase); for example, *John flew to Boston on Friday*. Sometimes the explicit prepositional marker is omitted and temporal referring expressions are used in adverbial (*Friday John flew to Boston*), nominal modifier (*John's Friday flight to Boston*) or elliptical/reduced relative clause (*John's flight, Friday at 5, will be crowded*) contexts. We refer to these cases as instances of **syntactically implicit** time–event relations.

However, in many cases the relational information may be implicit in a much less direct way, to be derived by the reader using world or lexical semantic knowledge, or narrative convention and discourse interpretation. In many of these cases relations between times and events are established indirectly by first establishing relations between events and then inferring relations between times and events. For example, consider the mini-narrative in (1):

(1) John arrived home at 9 p.m. He opened the door, dropped his briefcase, and poured himself a stiff whisky. Sipping his drink, he moved into the living room and collapsed in front of the TV.

The only explicit temporal relation here—between the event of arriving home and 9 p.m.—is asserted in the first sentence. The most likely interpretation of the ordering and absolute positioning of subsequent events is that they happened in the order they are recounted and within minutes of John's arrival home. This interpretation is made by assuming this narrative follows the general convention of recounting events in the order they occurred, and by drawing on script-like (Schank and Abelson 1977) world knowledge that tells us that on arrival home people don't drop their briefcases or pour themselves drinks before they open the door. Furthermore, since carrying around briefcases is cumbersome and at odds with what appears to be John's mood of

exhaustion/release from toil, suggested by what appears to be a late arrival home from work and by his need for relaxation, we assume John dropped the briefcase immediately after entering the house, and that the other events also occurred very soon thereafter. This interpretation is also boosted by the presumption, again based on narrative convention, that had other events intervened (e.g. half an hour spent walking the dog) we would have been told. We refer to the case's time–event relations such as those in this example (excluding the time–event relation in the first sentence) as *semantically implicit* time–event relations.

The question that arises here for a temporal annotation scheme is whether or not times ought to be associated with events when the relation is implicit, in either of the two senses just identified. Different positions are possible here and some of these are found in the work of authors in this part, reflecting fundamentally different conceptions of the ultimate goal of temporal annotation.

One position to take is that relations between time and events should be marked only in cases where explicitly signalled by prepositions or where they are syntactically implicit. This position is adopted by Schilder and Habel, who assume a default semantic relation of inclusion for all syntactically implicit relations. Time–event relations for events which do not occur in such syntactic contexts are simply not supplied.

Another possible position is to assign a calendrical time point or interval to *all* events in a text—so-called **time-stamping** of events. Filatova and Hovy pursue this line, developing a heuristic algorithm for news texts which assigns to each event (verb-bearing clause, as discussed in 3.2 above) a calendrical date, date range, or open-ended date interval (i.e. the interval before or after a given date). They use one set of rules which apply to cases of explicit time reference (e.g. temporal PPs), and another set that apply when no implicit information is available.

A further position to take is that time–event relations should only be marked in cases where they are explicitly signaled or are syntactically implicit (as with Schilder and Habel), but that event–event temporal relations (to be discussed later) should also be marked, so that calendrical time-points for some events can be recovered by inference from combinations of time–event and event–event relations (so, for example, if e_1 occurs at t and e_2 occurs after e_1 then we know e_2 occurs after t). The approaches of both Li et al. and Pustejovsky et al. admit event–event relations to be tagged as well as time–event relations and hence support this sort of indirect positioning of events in time.

4.2 Time-stamping events

Before discussing the annotation of event–event relations in detail, it is worth considering the time-stamping project in more detail. Time-stamping—by which we mean the assignment of a calendrical time reference (point or interval) to *every* event in running text—is an appealing aim. Motivating it is the intuition or wish, which is especially strong as concerns narrative texts such as newswires, that all events should be placeable on a time-line. This goal suggests that the target representation for a temporal annotator should be a mapping or anchoring of all events in a text on a calendrical time-line.

Despite its intuitive appeal, time-stamping all events has serious drawbacks which stem ultimately from the fact that natural language narratives underspecify event positions in time in a way that makes a time-line representation problematic. Put another way, narratives may only specify a partial ordering between events; a time-line

representation commits one to assigning a total ordering, information which simply may not be present in the text. This position is elaborated by Setzer and Gaizauskas (2002) who prefer a time–event graph, in which the nodes are times or events and the arcs are temporal relations, to a time-line as a target representation for temporal relation annotation. They present two arguments for this position which we repeat here in slightly modified form.

First, in many cases texts position events in time explicitly by relation to other events and any attempt to coerce these events onto a time-line must either lose information, invent information, or rely on a notion of an underspecified time-point constrained by temporal relations (i.e. introduce a representation of temporal relations by the back door). Consider this example:

(2) After the plane crashed, a search was begun. Later the coastguard reported finding debris.

and assume that an earlier sentence explicitly specifies the calendrical time of the plane crash. Attempting to map the information presented in this example onto a time-line we are faced with the situation depicted in Figure 1. While the crash event can be placed on the time-line the other two events cannot. Either time-points must be guessed, or an interval must be assigned. The first option is clearly not satisfactory. But if an interval is assigned the only possible interval, for both the searching and finding events, is the interval from the crash till the date of the article. However, if this is assigned to both events then the information about their ordering with respect to each other is lost.

A simpler representation, which does not attempt to be as specific but actually carries more information, is shown in Figure 2. This representation preserves the information that the searching event precedes the finding event, without forcing any early commitment to points on a time-line.

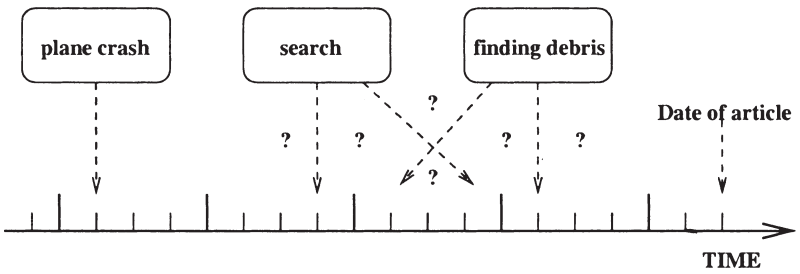

FIG. 1 A Time-line Representation.

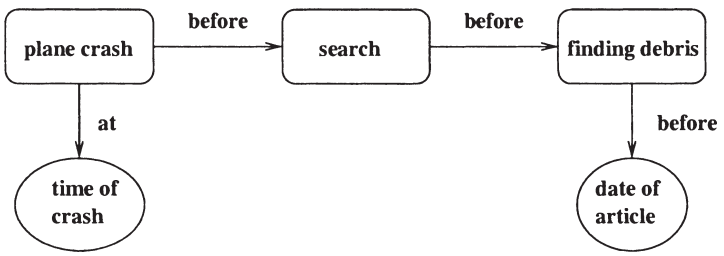

FIG. 2 A Time-Event Graph Representation.

The second argument for preferring a time-event graph representation that captures event–event temporal relations as well as time–event relations is that to position events on a time-line accurately requires the extraction of event–event relational information. In the example, the placing of the searching and finding events in the interval between the plane crash and the date of the article requires the recognition that these events occurred after the crash as signaled by the words 'after' and 'afterwards'. Without identifying the relations conveyed by these words the searching and finding events could only be positioned before the time of the article, and not following the plane crash. Thus, even if a time-stamp representation is viewed as the best target representation, achieving it requires the extraction of temporal relational information. In this case adopting a time–event graph as an intermediate representation is still a good idea, which begs the question of why it should not simply be taken as the final target representation.

4.3 Annotating relations between events and events

As with time–event relations, event–event temporal relations may be conveyed explicitly or implicitly. The primary mechanism for explicit relation is the temporal conjunction, typically used to relate the event expressed in a subordinated clause to one in a main clause; for example: *While chopping vegetables, John cut his finger* or *After the game John called Bob*. As with time–event relations, event–event temporal relations are frequently expressed implicitly, relying on world or lexical semantic knowledge, or narrative convention and discourse interpretation. Arguably all of the time–event relations in the mini-narrative (1) in Section 4.1 above (excluding the one in the first sentence) are established indirectly by first establishing event–event relations and then using plausible reasoning to position approximately the ordered events on a time-line. The event–event ordering is implicit in the narrative and established as described above.

Three of the papers in this part address the identification or annotation of event–event relation, though their concerns are different. Katz and Arioso are interested in the temporal relations between events, as signaled by verbs, within single sentences. Their primary concern is the study of how temporal information is conveyed within sentences such as *John kissed the girl he met at the party* where there are no explicit temporal relational markers. Is, for example, our knowledge that the kissing took place after the meeting dependent on lexical semantic knowledge of these two verbs? or on the recognition of the syntactic structure of matrix and subordinate clauses both with past tense verbs? To answer this question they propose adding to a large corpus of syntactically annotated sentences further annotations which capture temporal relational information. This resource could then be used for the induction of the sort of knowledge needed to resolve questions of temporal ordering in implicit contexts.

In their annotation scheme a human annotator adds labeled, directed edges between nodes in a graph which are the verbs in a syntactically annotated sentence. In addition to verb nodes, each sentence also has associated with it a node corresponding to its speech time (cf. Reichenbach, part I). The edges represent temporal relations and the edge labels and direction specify the relation (their set of relations contains just the two relations of precedence and inclusion, though their duals are also available by reversing the directionality of an edge). As noted above in the discussion of event annotation (Section 3.2), they do not consider event nominals.

While Katz and Arioso are concerned only with intrasentential temporal relations between verbs, the TimeML scheme proposed by Pustejovsky et al. aims to capture

event–event temporal relations as completely as possible and in a way that will facilitate the development of time, event, and temporal relational tagging systems for use in applications such as question answering and summarization. To that end they propose an approach to relational tagging that allows event–event relations be marked between any two event-denoting expressions (recall from Section 3.2 that for them events can be denoted by verbs, nominals, adjectives, predicative clauses, and event prepositional phrases). The approach relies on implementing a relational graph by using XML elements which consume no text but link, via pointers, XML elements surrounding event representatives and associate a relation type with the link. The set of relation types they employ are the thirteen proposed by Allen (Chapter 12). Note that these links, called TLINKS, can be asserted between any two event-denoting expressions (or between event and temporal referring expressions), regardless of whether or not they occur in the same sentence. This permits temporal relations that hold between events in different sentences to be recorded as in, for example, *John ate dinner. Later he phoned Bob*. However, it raises the question of whether temporal relations between *all* event–event pairs should be annotated in an extended text (an overwhelming task for a human annotator) or if not, which subset of them. This question is further addressed in Sections 5 and 6.2 below.

Li et al. are concerned to build a working temporal information extraction system for Chinese newswire texts. They distinguish single event, multiple event, and declared event statements which are statements reporting one event, two or more events, and events as asserted by a person or organization, respectively. In their model, single event statements are related to times (i.e. placed in a temporal relation to a calendrical time-point), while in multiple event statements the events are related to each other, using one of Allen's thirteen temporal relations. Thus, like Katz and Arioso, event–event relations are only marked within sentences. However, presumably event–event temporal relational information for events in separate sentences is available indirectly via the temporal relation of these single events to times on a time-line.

4.4 Subordinating and aspectual relations

If one considers verbs as event signals and examines sentences with multiple verbal elements with a view to labeling their temporal relations, several problem cases soon emerge. Consider, for example, *John might have kissed the girl he met at the party* or *John hoped to kiss the girl he met at the party (and did/did not)*. In neither case can we mark a temporal relation between *kiss* and *met*, because we do not know whether or not it occurred. These cases reveal that in contexts where verbs are modally subordinated, or occur as arguments in intensional constructions, they cannot straightforwardly be taken as denoting real events. However, there are some such contexts where the events the subordinated verbs denote are guaranteed to have occurred, such as *John forgot that he had already paid the bill* or *John knew Bill had gone*.

A further class of problem cases are those involving aspectual verbs, such as *start, keep*, which may signal the beginning, culmination, termination, or continuation of an activity, as in *John started chopping vegetables* or *Sue kept talking*. These verbs do not signal events distinct from the ones denoted by their verbal arguments, but rather draw attention to an *aspect* of these events. Attempting to assert a temporal relation between them, therefore, is problematic.

These cases demonstrate that proposing to annotate *temporal* relations between *all* verbs within a sentence is not sensible. There are two other possibilities. One is to ignore

them; the other is to annotate these verb–verb relations in some other way. Ignoring these contexts might have no impact on certain uses of temporal annotation, for example on Katz and Arioso's project of building an annotated corpus from which to induce the temporal profile of lexical items. For other applications, such as question answering or summarization, however, the ability to distinguish these contexts is certainly needed. Either to learn to ignore them, or to handle them appropriately, an annotation scheme for these contexts is desirable. This has been proposed in the TimeML specification, via the addition of two further sorts of relational links. Subordination links (SLINKS) are introduced to deal with cases of subordinate relations, and aspectual links (ALINKS) are introduced to deal with the cases of relations introduced by aspectual verbs.

4.5 Automated temporal relation tagging

The preceding discussion has focused on elucidating the targets of temporal relation tagging, rather than the implementation and performance of systems that actually attempt to do this automatically. As with automated event tagging discussed in Section 3.5 above we note that only three of the papers in this part describe implemented systems that actually do relation tagging: Filatova and Hovy, Schilder and Habel, and Li et al. As noted above, Filatova and Hovy are concerned with time-stamping all events (clauses). Their implementation does this using a rule-based system with two classes of rules: those for clauses with explicit date information and those for clauses without explicit date information. In the latter case tense information from the clause together with information about the most recently assigned time or the date of the article are used to assign a time-stamp to the clause. They report evaluation figures of 82 per cent correctly assigned time-stamps to clauses correctly identified. Schilder and Habel describe an implementation which extracts temporal relation information only from sentences containing time expressions and event expressions. They report figures of 84.49 per cent recall and precision for the system in extracting temporal relational information from these sentences. Finally, Li et al. describe an implemented system which extracts both time–event and event–event relational information from Chinese sentences. Their system relies on a complex set of rules which map from surface temporal indicators to temporal relations between times and events and events and events. They report figures of close to 93 per cent correct extraction of these relations. Given the differences in tasks and test corpora, none of these figures is directly comparable.

5 COMPARING ANNOTATIONS: SEMANTIC APPROACHES TO EVALUATION

One of the clear lessons to have emerged from the corpus annotation programme is the necessity, for any given annotation scheme, of some means to compare quantitatively two annotations of the same text created by separate annotators (human or machine). As noted in Section 1, above precise quantitative comparison permits validation of the annotation scheme (if two humans cannot agree then the scheme must be unclear), evaluation of competing algorithms, and hill-climbing approaches to machine learning of automated annotators.

Typically one wants to ensure that an annotator has annotated all the things it ought to and none that it ought not. These intuitions are captured in the metrics of precision and recall, mentioned above in 2. Defining these metrics concretely can only be done

with reference to a specific annotation scheme. For annotation schemes which specify a single correct annotation for any piece of text (i.e. are deterministic) this is relatively simple. For example, for texts annotated according to an annotation scheme for temporal referring expressions, there is typically a single way to tag the text correctly, tagging certain character sequences as times and perhaps associating additional attribute information, such as normalized time values, with these sequences. It may be desirable to decouple the evaluation of the accuracy with which text **extents** have been correctly tagged from the accuracy to which **types** or other attributes have been assigned to them. For example, one can separately assess the accuracy of a temporal referring expression tagger at marking the beginnings and endings of character sequences referring to times or dates, from its ability to distinguish which are times and which are dates, and this again from its ability to assign the correct normalized time. All this is quite straightforward, given the presumption of a single correct tagging.

What are much more problematic are annotation schemes which permit alternative equivalent taggings, particularly when it is not independent strings that are being tagged, but *relations* between multiple strings. However, this is precisely the situation which is likely to arise with annotating temporal relations *since the same information can be captured by distinct annotations*. Consider the abstract case, where two events A and B are said to occur at the same time and a third event C occurs later. If one annotator marks A and B as simultaneous and C after A and a second annotator also marks A and B as simultaneous, but instead marks C after B, then they do not differ in terms of their views on what happened when. Nor indeed would they differ from a third annotator who marked A and B as simultaneous and marked C after A *and* C after B.

This example is used by Setzer et al. (Chapter 29) to motivate their proposal for a semantic basis for comparing temporal annotations. They define the **temporal closure** of an annotation (of a specific text) to be the deductive closure, i.e. the set of all temporal consequences, that may be drawn from the annotation using a set of inference rules which capture essential properties of the temporal relations used in the annotation scheme. Two annotations are then said to be equivalent if their temporal closures are equivalent. So, for example, suppose the set of inference rules contains the plausible rule that says that for all times or events x, y, and z, if x and y are simultaneous and z is later than x then z is later than y. Using this rule, the temporal closures of each of the three annotations of the temporal relations between events A, B, and C introduced above are the same. Thus, we have means of abstracting away from the surface form an annotation to the temporal information it conveys, and based on this a way of comparing temporal annotations semantically. Setzer et al. go on to define measures of precision and recall in terms of the notion of temporal closure. In essence recall is the proportion of temporal closure of the key annotation which is found in the temporal closure of the response and precision is the proportion of the response annotation which is found in the key. They note that the problem of defining semantically based metrics for comparing temporal relation annotations is related to the problem of comparing coreference chain annotations, originally addressed in MUC-6, to which similar semantically based solutions have been proposed. A key difference is that whereas coreference is an equivalence relation which induces equivalence classes over sets of linked entities in the text, temporal relations are not, in general, equivalence relations. Thus, solutions based on the number of links which need to be added to transform the response (key) equivalence class into the key (response) equivalence class are not sufficient. Rather, the full implicational consequences of the annotations, i.e. the temporal closures, need to be compared instead.

Katz and Arioso propose a similar semantically based solution to the problem of comparing annotations, though with a more overtly model-theoretic character. They define the **model of an annotation** to be the assignment of pairs of entities in the domain of the annotation (the verbs, i.e. event realizations, in the annotation) to the two binary temporal relations in their annotation scheme—precedence and inclusion—that satisfy a set of axioms (for example, transitivity of precedence). An annotation is **satisfied** by a model if and only if all temporal relations in the annotation are satisfied by, i.e. are found in, the model. A distance measure may then be defined between annotations in terms of the numbers of models they share. More precisely, the distance is the sum of the number of models satisfying one annotation but not satisfying the other normalized by the number of models satisfying both.

Taken together, the measures already in use for evaluating temporal referring expressions (e.g. for TIMEX and TIMEX2) and the proposals discussed here for evaluating temporal relation annotations form the basis of a comprehensive framework for evaluating temporal annotations. This framework itself is part of an emerging publicly available set of resources for temporal annotation, to which we now turn.

6 ANNOTATION RESOURCES: STANDARDS, TOOLS, CORPORA

The preceding sections have discussed issues in the annotation of times, events, and temporal relations, and in the evaluation of these annotations. While a consensus in all these matters has not yet been achieved, sufficient progress has been made that various groups have made, or are in process of making, available annotation standards, tools to assist in the process of annotation and annotated corpora to promote research and reduce duplicated effort in this area. This section details some of these resources.

6.1 Annotation standards

Experience in creating annotated resources has shown that the creation and publication of an **annotation guidelines** or **specification** document is a key step towards creating an open standard. This document has as its ideal that two randomly selected individuals should, by following the annotation instructions in the document, produce identical annotations over any set of appropriate texts. Clearly this is an ideal, but it is a key aspect of an objective, scientific methodology (repeatability by independent observers) in this area. It also has the salutary effect of forcing would-be implementors of automated taggers to separate the 'what' from the 'how' in their work.

For some of the work reported in papers in this part it is not clear whether full temporal annotation specifications exist. The only temporal annotation proposals for which we are certain that annotation guidelines or specifications are publicly available are: TIMEX (MUC 1995, 1998), TIMEX2 (Ferro et al. 2001), TIMEX3 (available as part of the TimeML specification), and TimeML <http://www.timeml.org>. This is not to condemn work carried out without independent annotation specifications. In many cases the broad shape of what is proposed as a target annotation is clear from papers concerned with reporting algorithms as well as target representation. Furthermore, creating an annotation guidelines document suitable for creating large-scale annotated resources and for carrying out interannotator reliability experiments is a lot of work that

is not feasible and arguably not necessary for initial exploratory research, which characterizes much of the work reported here. It is to be hoped, however, that agreement can be reached on multilingual standards for temporal annotation, since this will benefit everyone engaged in this research area.

Apropos annotation standards, another topic of relevance is the language in which the annotations are to be expressed. A consensus appears to be emerging in the corpus linguistic community that linguistic annotations should be represented in some form of XML. Some annotated corpora, however, such as the TIMEX-annotated MUC corpora, were annotated prior to the advent of XML and are available with naive SGML-based annotation only (naive in the sense that no DTD is supplied, making the corpora unprocessable by an SGML parser). TimeML, including TIMEX3, has been specified in XML and an XML schema is available. Annotated corpora that are TimeML compliant should be possible to validate with an XML parser. The broader question of whether temporal annotations can or should become compliant with overarching schemes which have been proposed for linguistic annotation, such as the Text Encoding Initiative <http://www.tei-c.org> or the Corpus Encoding Standard <http://www.cs.vassar.edu/CES>, remains to be addressed.

6.2 Annotation tools and the process of annotation

As noted in the discussion of the TimeML approach to event–event annotation in Section 4.3 above, since any two events in a text may potentially be temporally related by an annotator, the issue arises as to whether an annotator should be directed to annotate *all* event pairs (probably infeasible) or only a subset, in which case it must be decided which subset. One solution that has been proposed to this problem (Setzer et al.) is to exploit further the notion of temporal closure to assist manual annotation. They propose a three-stage process for the manual annotation of all times, events, and temporal relations in a text. First, all times and events are marked, possibly with automated assistance as a preprocessing step (for example a tagger for temporal referring expressions could be run and its output corrected by a human). In the second stage a subset of the temporal relations is marked. These could be those which are explicitly expressed via temporal prepositions or conjunctions (if TimeML is adopted these signals are required to be annotated) and perhaps others which are particularly obvious (e.g. what were called syntactically implicit time–event relations in 4.1 above) or salient to the annotator. The key point here is that this set need not be defined precisely. In the final stage, the annotator takes part in an interactive process in which the temporal closure of the relations previously entered is computed and an as yet unrelated time–event or event–event pair from the set of all possible time–event/event–event pairs (as defined by the first stage) is selected and the annotator is promoted to supply the relation between them (possibly undefined). This process is repeated until a relation has been supplied or inferred for every time–event/event–event pair in the text. The advantage of the above procedure is that it significantly reduces the burden on the annotator. A pilot experiment reported by Setzer et al. revealed that almost 80 per cent of the relational pairs in the temporal closures of a small set of texts were supplied automatically by inference, with approximately 4 per cent being initially annotated and 16 per cent annotated under interactive prompting.

Regardless of whether an interactive annotation tool built on top of a closure procedure is available, the task of annotating text according to a scheme as rich as TimeML

is clearly demanding and requires sophisticated support. To this end a temporal annotation tool (TANGO) is under development; see <http://www.timeml.org>. TANGO allows a user to see the text being annotated in one window, a time–event graph with in a second window, and the underlying XML annotation in a third. Temporal, subordination, or aspectual links may be added or removed and a temporal closure algorithm run to add links by inference.

6.3 Temporally annotated corpora

As with annotation standards, the publication of annotated resources is a boon to the research community, enabling researchers to build on and contribute to prior work. Existing publicly available, though not necessarily free, temporally annotated corpora are the MUC-6 and MUC-7 TIMEX annotated corpora, which can be obtained from the Linguistic Data Consortium: <http://www.ldc.org>. The TIMEX2-tagged Enthusiast corpus and Korean corpus, as described in Wilson et al., are freely available to those who have licensed access to the underlying text (details at <http://timex2.mitre.org>. As part of the TERQAS workshop which stimulated the creation of the TimeML standard and the TANGO annotation tool a corpus of 300 newswire texts, called TimeBank (Pustejorslly et al. 2003), has been annotated with a subset of the full TimeML annotations. Further refinement of this corpus is still on-going, but it will in due course become available from the LDC.

7 CONCLUSION: ISSUES AND CHALLENGES

In the foregoing we have attempted to give an overview of the issues involved in the temporal annotation of text, as well as a snapshot of the current state of affairs concerning automated annotators and the resources that are publicly available to support research and development in this area. There remain many issues to be addressed and challenges to be overcome before either the scientific project of understanding how temporal information is conveyed in text or the engineering project of building reliable temporal annotators that can be embedded in language technology applications can be said to be complete. In concluding we raise some of these outstanding issues and challenges.

7.1 Multilinguality

Most of the work done to date on temporal annotation is for English, though Schilder and Habel's work for German and Li et al.'s work for Chinese are exceptions. Wilson et al. report that they have applied their TIMEX2 annotation standard to documents in Spanish, Korean, French, Hindi, and Chinese as well as to English. Katz and Arioso claim they have applied their temporal relation tagging scheme to sentences 'from a variety of languages', but give no indication of just which languages or how many. Clearly this work just scratches the surface of the world's languages. Applying ambitious annotation schemes such as TimeML to multiple languages will bring many benefits: the adequacy of annotation schemes will be further tested, rich insights will be gained into the range and distribution of mechanisms employed across the world's languages to convey temporal information, more researchers will be drawn into work in

this area, and many of the resulting resources, tools, and applications will be sharable across languages.

7.2 Annotation tools and resources

The foregoing should have made clear the need for, and the utility of, annotated resources. To date there is still a serious shortage of these, particularly for more complex relational annotations, such as those found in TimeML. The creation of such resources is difficult and time-consuming, hence expensive. Further development of specialized annotation tools, such as TANGO, should help, as will the integration into such tools of robust preprocessing components that will relieve some of the annotator's burden. More studies need to be made of the levels of agreement obtainable by annotators, with a view to improving annotation specifications, or recognizing fundamental limitations in the process.

7.3 Building temporal relation taggers

While automated taggers for annotating temporal referring expressions have achieved good levels of accuracy, even for the difficult task of evaluating indexicals and contextually sensitive expressions, the programme of constructing taggers to tag temporal relational information is still in its infancy. Some progress has been made, as reported by some authors in this part, but there are as yet no automated taggers that can begin to approximate the rich annotation proposed in TimeML. Building such taggers is an exciting challenge. However, achieving this capability may be a long-term project, and researchers may be well advised to target simpler, intermediate goals. So, for example, while in Section 4.2, we presented arguments against time-stamping as a suitable target representation for capturing *all* temporal relational information in text, this does not mean that taggers that accurately time-stamp a subset of events in text would be of no utility. In the short term, the creation of taggers that focus solely on relating times and events where these are explicitly signaled or straightforwardly associated may be a sensible goal. Of course defining this goal sufficiently precisely as to allow quantitative evaluation is itself no small task.

7.4 Applications

As automated temporal annotation taggers with increasing functionality become available, applications will be able to take advantage of this. Obvious examples, cited repeatedly by authors in this part are question answering, information extraction, and summarization. Questions may explicitly request temporal information (*When did the French Revolution begin*?) or be time-sensitive in less direct ways (*Who was president of Enron in 2000*/*when its share price was highest*?) Information extraction is concerned with extracting attributes of entities (e.g. titles of persons) or relations between entities (e.g. the `employee_of` relation). However, for most real applications attributes and relations will hold of entities in temporally bounded ways, and knowing these temporal bounds is critical for the extracted information to be of value. Summarization of multiple documents which overlap in their description of events and need to be reduced to a single nonredundant chronological narrative is a requirement in numerous areas,

ranging from assembling background information from news reports to condensing clinical records and courtroom proceedings. As with many areas of language technology, the challenge will be to deploy imperfect technology in settings where it can be of genuine value and where the lessons learned in its deployment can be fed back to fuel a positive development cycle.

REFERENCES

Ferro, L., Mani, I., Sundheim, B., and Wilson, G. (2001). 'TIDES temporal annotation guidelines, version 1.0.2'. Technical Report MTR 01W0000041, MITRE Corporation.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MUC (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. California: Morgan Kaufmann.

MUC (1998). *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. California: Morgan Kaufmann. Available at <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/>.

Pustejorslly, J., Hantts, P., Saurí, R., See, A., Gaizauollas, R., Setzer, A., Rader, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). 'The TIMEBANK corpus', in D. Archer, P. Rayson, A. Wilson, and T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, Lancaster, March 2003, 647–56.

Schank, R. and Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Lawrence Erlbourm Associates.

Setzer, A. and Gaizauskas, R. (2002). 'On the importance of annotating temporal event–event relations in text', in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002) Workshop on Annotation Standards for Temporal Information in Natural Language*, Las Palmas de Gran Canaria. European Language Resources Association, 52–60.