A Pilot Study on Annotating Temporal Relations in Text

Andrea Setzer and **Robert Gaizauskas**

Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
Sheffield S1 4DP, UK
{A.Setzer, R.Gaizauskas}@dcs.shef.ac.uk

Abstract

We describe a pilot study in which a scheme for annotating events and temporal relations between events in text is applied to a small corpus of newswire texts. High levels of agreement between human annotators are shown to be difficult to achieve and we investigate where discrepancies occur and how these might be addressed.

1 Introduction

Many natural language processing applications, such as information extraction, question answering, topic detection and tracking, would benefit significantly from the ability to accurately position reported events in time, either relatively with respect to other events or absolutely with respect to calendrical time.

To date relatively little work has been done on the extraction of temporal information from text. The Message Understanding Conferences (MUCs) addressed the problem in a limited way. So for example, the MUC-6 named entity subtask required the identification of absolute time expressions in text (MUC6, 1995), and the MUC-7 named entity subtask extended this requirement to include relative time expressions (MUC7, 1998), but neither of these tasks required placing events in time, or temporally relating events to each other. The MUC-5 and MUC-7 scenario tasks required participants to assign a calendrical time just to the specified scenario event types (joint venture announcements and rocket launchings, respectively), but this is a limited task and the scores were low, indicating its difficulty. More recently the TIDES Temporal Guidelines (Ferro et al., 2000) have been developed – a very thorough set of guidelines for annotating time expressions. These will clearly supercede the MUC named entity guidelines for time expressions, but again are not setting out to annotate events and temporal relations between events or between events and times.

In Setzer and Gaizauskas (2000a) we introduced an annotation scheme which does address these issues. While our guidelines also require the annotation of time expressions (and are much less detailed in this regard than the TIDES guidelines, which we are happy to adopt more or less completely), our guidelines go well beyond annotating time expressions and also describe how to annotate events and the temporal relations between events and other events or events and times. Thus our work should be seen as essentially complementary to the TIDES work, and in our view explores where annotation of temporal information in text should go next.

We have now carried out a pilot study in which a group of annotators has been supplied with our temporal annotation guidelines and asked to apply the annotation scheme to a trial corpus. In particular we were interested in answers to these questions:

- how unambiguous and comprehensive are our temporal guidelines?
- how much genuine disagreement is there about temporal relations in text?
- how burdensome is our annotation procedure? – i.e. is it feasible to think of

annotating a corpus of significant size at this level of detail?

This paper discusses the design of the pilot study, its outcome and the answers to these questions, insofar as we have been able to determine them.

2 Annotating Temporal Information in Text

In this section we briefly introduce the annotation scheme we propose. This scheme has been introduced in detail in Setzer and Gaizauskas (2000a; 2000b) to which the reader should turn for more information.

2.1 Conceptualising Time

Before we can propose an annotation scheme we must make clear the temporal entities and temporal relations we suppose exist. Our temporal ontology is a pragmatic one aimed at enabling us to identify events, determine their relative temporal order and, if possible, position them in calendrical time. It is not aimed at deriving at a philosophically 'true' description of temporal reality.

We provide an intuitive and restricted set of temporal entities and relations which enables us to achieve our goal to a certain extent. We presume the world contains the following primitive types: events, times, temporal relations and event identity.

Events Intuitively an event is something that happens, something that one can imagine putting on a time map. Events have to be anchorable in time, and they are usually conveyed by finite verbs and some nominalisations. Events can be ongoing or conceptually instantaneous, we do not distinguish between these. For example, the sentence *A small single-engine plane crashed into the Atlantic Ocean* conveys an event.

Times Like events, times can be viewed as having extent (intervals) or as being punctual (points). Both are treated as time objects. A time object must, however, be capable of being placed on a time line. Examples of time objects are *last Tuesday*, *April 4, 1998* and *March 1997* and also the referents of more complex expressions like *17 seconds after the crash.*

Temporal Relations Events stand in certain temporal relations to other events and to times, as in the following two examples. *The plane crashed after the pilot and his crew ejected* and *A small single engine plane crashed into the Atlantic Ocean on Wednesday.*

The full set of temporal relations we suppose at present is before, after, includes, included and simultaneous. This is a minimal set defined after a number of newswire articles had been analysed; it can easily be expanded should it prove useful or necessary.

Event Identity In newswire articles, events are usually referred to more than once, with subsequent references giving more detailed information. Event identity proves a useful and comparatively easily annotated relation. In Setzer (2000a), we also included subeventness as an additional event relation to be annotated. This has been excluded in the current work because subeventness is very difficult to define and poses a significant additional burden on annotators.

2.2 Annotation

We give a brief overview of the annotation scheme and describe how to annotate the entities, including the main attributes that are associated with them. For more detailed information about the attributes we again refer to Setzer and Gaizauskas (2000a).

Events The head of the finite verb group expressing the event is annotated as a representative. Should the event be conveyed by a nominalisation or a non-finite verb group then the head of that group is chosen. The main attributes for event annotations are a unique ID, the event class, verb tense, one or more other events or times to which the event is related and the temporal relation which holds between the event and the related events or times.

Times The text span referring to the time object is marked up. Times also have a unique ID and following the approach taken in MUC, we distinguish between dates and times (i.e. time objects of extent greater than or less than one day).

Temporal Relations and Event Identity Temporal relations and event identity are included in the attributes of events. If the temporal relation is explicitly signalled in the text (e.g. by a temporal preposition) then the text span conveying this signal is annotated as a separate entity and given a unique ID. The following two examples show how the entities are annotated, including their main attributes.

All 75 people on board the Aeroflot Airbus
<event eid=4 class=occurrence
relatedToEvent=5
eventRelType=simultaneous
tense=past signal=7>
died </event>
<signal sid=7> when </signal> it
<event eid=5 class=occurrence
tense=past >
ploughed </event>
into a Siberian mountain.

A small single-engine plane
<event eid=9 class=occurrence
relatedToTime=5
timeRelType=included
tense=past signal=9>
crashed </event>
into the Atlantic Ocean about eight miles
off New Jersey
<signal sid=9> on </signal>
<timex tid=5> Wednesday </timex>.

2.3 Comparing Temporal Annotations

As with coreference relations, it is possible to annotate semantically identical temporal relations in syntactically different ways. Should, for example, event A happen before event B then both 'A before B' and 'B after A' are valid representations. Further should A happen before B and B before C, then A happens before C, but in the absence of any explicit relational expression in the text, an annotator may or may not explicitly annotate the relation. This problem is similar to that addressed in defining the coreference scoring scheme adopted in the Sixth Message Understanding Conference (Vilain et al., 1995), and the solution we propose is related.

We proceed as follows. Once times, events and temporal relations have been annotated, relations in the text are normalised and then a deductive closure over these relations is calculated, using what is assumed to be a set of complete temporal inference rules. The result is a fully elaborated temporal model of the text and any two such models can be compared to calculate recall and precision figures (in fact this procedure is carried out interactively with the annotator, as discussed in section 3.2 below, but this is not relevant to the formal discussion here).

The procedure for deductive closure is as follows. The IDs of the events and times annotated in a text form two sets, E and T, respectively. Each of our temporal relations is binary and thus can be viewed as a subset of $(E \cup T) \times (E \cup T)$. For each temporal relation certain formal properties pertain. For example *simultaneous* is an equivalence relation, while *before* and *includes* are transitive, but asymmetric and irreflexive. Given a partially specified model of the temporal relations in a text, the deductive closure of each relation can be computed to arrive at a total model.

Let us denote sets of pairs from $(E \cup T) \times (E \cup T)$ which constitute the denotations of simultaneous, before and includes S, B and I respectively. The set of inference rules we need to compute the deductive closure contains for example:
$\forall x, y, z \in (E \cup T) \times (E \cup T)$

- $(x, y) \in S \Rightarrow (y, x) \in S$
- $(x, y) \in S \wedge (y, z) \in S \Rightarrow (x, z) \in S$
- $(x, y) \in B \wedge (y, z) \in S \Rightarrow (x, z) \in B$

- $(x, y) \in I \wedge (y, z) \in S \Rightarrow (x, z) \in I$

We can now specify what precision and recall mean in this framework. Letting S_k and S_r denote the annotated *simultaneous* relations in the answer key and system response respectively and $S_k^{\vdash}$ and $S_r^{\vdash}$ their deductive closures, respectively (and similarly for B and I). The recall and precision for the simultaneous relation is given by:

$$R = \frac{|S_k^{\vdash} \cap S_r^{\vdash}|}{|S_k^{\vdash}|} \qquad P = \frac{|S_k^{\vdash} \cap S_r^{\vdash}|}{|S_r^{\vdash}|}$$

Recall and precision measures can be defined in a parallel fashion for the other relations. An overall recall and precision measure for all temporal relations can then be defined as follows:

$$R = \frac{|S_k^{\vdash} \cap S_r^{\vdash}| + |B_k^{\vdash} \cap B_r^{\vdash}| + |I_k^{\vdash} \cap I_r^{\vdash}|}{|S_k^{\vdash}| + |B_k^{\vdash} + |I_k^{\vdash}|}$$

$$P = \frac{|S_k^{\vdash} \cap S_r^{\vdash}| + |B_k^{\vdash} \cap B_r^{\vdash}| + |I_k^{\vdash} \cap I_r^{\vdash}|}{|S_r^{\vdash}| + |B_r^{\vdash} + |I_r^{\vdash}|}$$

3 The Setup of the Experiment

3.1 The Corpus

The trial corpus consists of 6 newswire articles taken from the New York Times, 1996. There were part of the MUC7 (MUC7, 1998) training data. Basic statistics about the corpus are presented in Table 1.

	sentences	words	events	times	signals
text1	26	448	41	11	10
text2	18	333	31	5	3
text3	13	269	19	3	7
text4	13	213	27	0	1
text5	10	211	16	4	3
text6	13	399	26	5	4
total	93	1873	160	28	28

Table 1: The Corpus

3.2 The Process of Annotation

The recommended procedure is to annotate a text in the following two stages. All phases of the annotation are carried out using a GUI-based annotation tool specifically designed for this purpose. Stage I:

1. Annotate events, times and signals, but no temporal relations between them.
2. Annotate explicit temporal relations between events and events and events and times.
3. Annotate obvious implicit temporal relations, for example where times are clearly related to events, but this is not explicitly signalled in the text.

Once this stage is complete, the temporal deductive closure is calculated and additional implicit temporal relations are solicited interactively from the annotator. Stage II:
Until there are no temporally unrelated events in the temporal model of the text do:

1. Draw all inferences that can be drawn from the temporal relations given in the current temporal model of the text according to the temporal inference rules and add them to the temporal model.
2. Identify an unrelated event-event or time-event pair in the temporal model and prompt the user for the temporal relation (which may be "unknown").

After these two stages the result should be a temporal model of the text that is as complete as possible.

3.3 The Annotation Procedure

Each text was annotated by either two or three annotators, in addition to the first author who produced what in the following is taken to be the 'gold standard' or 'key' annnotation. To produce the gold standard, an initial annotation was created for each text, compared with the annotations provided by the annotators and then revised if necessary. The result is the 'key' annotation. On average, it took the annotators about an hour to annotate each text.

4 Experiments and Results

As mentioned in the introduction, the purpose of the pilot study was to ascertain how clearly defined our guidelines were, how much genuine disagreement there is about temporal relations, and how feasible this complex

annotation task is. To compare the annotators' output we have compiled several sets of figures, roughly corresponding to a measure of how much they agree after Stage I of the annotation process as described above (i.e. after the initial annotation of events, times, and temporal relations, but prior to any computation of deductive closure) and then after Stage II (i.e. after interactively extending the temporal model in conjunction with the deductive closure computations).

To measure stage I agreement we first calculated recall and precision figures for temporal entities (events, times, temporal signals) and then for all attributes associated with matched instances of these entities (similar to scoring matched slots in aligned template objects in the MUC IE tasks). To compute recall and precision for the entities alone, we attempted to match each entity in the response with an entity in the key. Two entities are judged to match if the same textspan was annotated with the same type (event, time or signal). The formulae for recall and precision are:

$$R = \frac{\text{number of matches}}{\text{number of key entities}}$$

$$P = \frac{\text{number of matches}}{\text{number of response entities}}$$

To obtain results for the attributes, we separately counted all attributes for each entity in the key and in the response (apart from entity IDs). For the entities in the response that matched entities in the key we counted the matching attributes.

$$R = \frac{\text{number of matching attributes}}{\text{number of key attributes}}$$

$$P = \frac{\text{number of matching attributes}}{\text{number of response attributes}}$$

To measure stage II agreement we used the recall and precision measure defined in section 2.3. We determined agreement on temporal relations after three different phases: after stage I, at the outset of stage II; after the initial deductive closure computation in stage II, but prior to any solicitation of additional temporal relations; after stage II was complete.

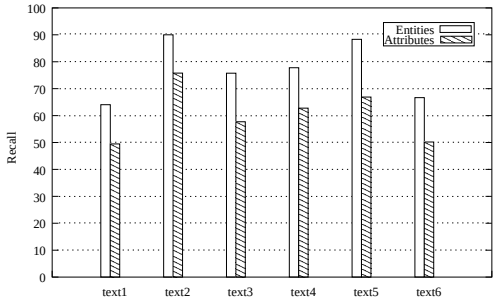

Figure 1: Recall after Stage I

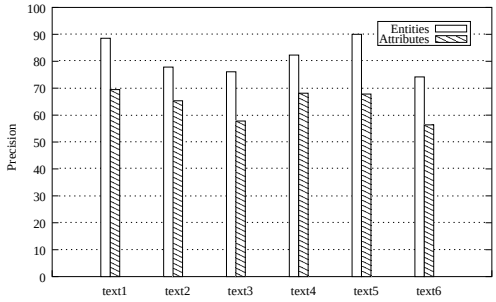

Figure 2: Precision after Stage I

4.1 Results

The results presented here were obtained by averaging the recall and precision values over all annotators for each text. Texts 1, 3, 5 and 6 were annotated by three annotators whereas texts 2 and 4 were annotated by two annotators. Further work is needed to report and analyse differences amongst the annotators and between individual annotators and the 'gold standard'.

The first two figures (1 and 2) show recall and precision figures for temporal entities and attributes (after stage I). The next two figures (3 and 4) show the recall and precision values after stage II (for each of the three phases). Table 2 shows the exact results from which Figures 1-4 were constructed (the averages here are calculated by weighting each text equally, not weighting texts based on how many entities/attributes/relations they contain).

4.2 Analysis

First, note that there is a strong correlation between the recall of temporal relations following stage II and the recall of temporal en-

	text1	text2	text3	text4	text5	text6	avg
Recall Entities	64.05	90.00	75.76	77.78	88.33	66.67	77.10
Precision Entities	88.53	77.84	76.10	82.31	90.00	74.18	81.49
Recall Attributes (all)	49.48	75.77	57.70	62.75	66.87	50.19	60.46
Precision Attributes (all)	69.49	65.31	57.80	68.12	67.82	56.38	64.15
Recall Phase1	12.12	29.16	33.33	20.00	20.00	19.30	22.32
Precision Phase1	22.79	43.75	42.93	17.05	20.00	41.25	31.30
Recall Phase2	20.07	50.97	44.45	33.44	32.04	23.05	34.00
Precision Phase2	47.17	75.40	43.44	14.78	5.80	43.86	39.00
Recall Phase3	26.69	50.97	50.68	33.44	49.26	29.35	40.07
Precision Phase3	86.80	72.31	70.91	54.02	58.67	63.61	67.72

Table 2: Recall and Precision values (as %)

	Times	Events
not anchored in time	71.43% (10 out of 14)	47.06% (32 out of 68)
wrong textspan	28.57% (4 out of 14)	14.70% (10 out of 68)
guidelines	0% (0 out of 14)	38.24% (26 out of 68)

Table 3: Mistakes for additional entities

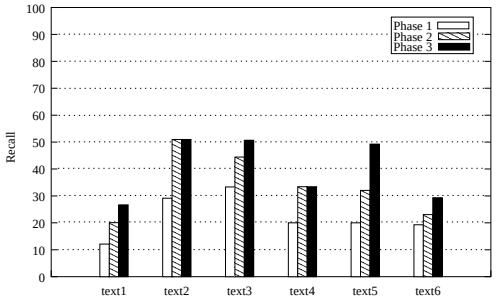

Figure 3: Recall after Stage II

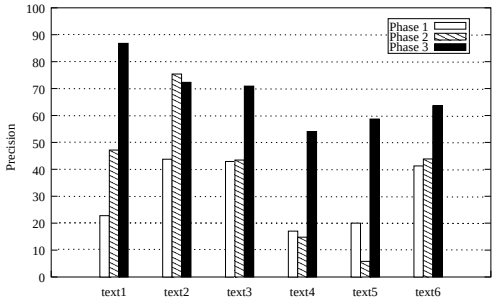

Figure 4: Precision after Stage II

tities after stage I. The number of event-event and event-time relations approaches $|(E \cup T) \times (E \cup T)|$, i.e. the number of entities squared. Thus, for example, if we imagine the number of entities in the key is 10 and the number of entities in the response is 6 (with a precision of 100%, i.e. all response entities are also in the key) then the potential number of relation facts in the temporal closure of the key is $10^2 = 100$ whereas the maximum number of facts in the response cannot go above $6^2 = 36$. This means that the recall of the response cannot be above 36%. So, although the values for the recall for the three phases of the temporal closure seem very low, they cannot be interpreted without looking at the values for the recall for the entities. And in fact, one can observe that those texts that have lower recall for the entities also have lower recall for the temporal closure.

When we look at the errors made for the entities only then we can broadly distinguish two groups. Entities that were missed out (and thus affect the recall) and entities that were annotated in addition to the entities of the key (affecting the precision). We can further subclassify additional entity mistakes as:

- not anchored in time The time or event is not anchorable in time but has been annotated despite this.
- wrong textspan It is clear what time or event was intended, but the textspan annotated for this time or event differs from the textspan in the key and is thus not recognised as the same entity by the scoring software (Note: such an error leads to a missing event/time and an additional event/time. We have left figures for both in the tables to indicate what proportion

	Times	Events
different opinion	35.29% (6 out of 17)	12.27% (13 out of 106)
wrong	41.18% (7 out of 17)	78.30% (83 out of 106)
wrong textspan	23.53% (4 out of 17)	9.43% (10 out of 106)

Table 4: Mistakes for missing entities

different opinion	missing	wrong	additional	guidelines
5.34%	26.70%	35.11%	13.43%	19.42%
33 out of 618	165 out of 618	217 out of 618	83 out of 618	120 out of 618

Table 5: Mistakes for attributes

of the overall errors these form).

- guidelines The guidelines have been applied incorrectly; for example, a hypothetical event has been annotated.

And, missing entity mistakes can be subclassified as:

- different opinion It is arguable, given the guidelines, whether or not an event or time should be annotated.
- wrong textspan As above.
- wrong An entity was not annotated for no apparent reason.

Tables 3 and 4 show the distribution of mistakes made, distinguishing between those made for times and those for events.

Examining the subgroups, it becomes evident that some of the mistakes could easily be rectified with more training of the annotators and an improved version of the guidelines. wrong textspan, guidelines and wrong fall under this category, which covers 68.06% of all errors. More difficult is the category where there is genuine ambiguity about the right results. Opinion and not anchored in time fall into this category, which covers the remaining 31.94% of all errors.

Looking at the attribute errors, we can distinguish the following groups:

- missing An attribute was not filled in.
- wrong An attribute is present but incorrect.
- different opinion It is debatable, for example, what the temporal relation is.
- additional Implicit temporal relations can be filled in, which results in attributes not present in the key. This is not actually a mistake, but the automatic recognition of mismatches cannot distinguish between genuine mistakes and additional information.
- guidelines The guidelines were not applied correctly. For example, for a reporting event the argument event should be entered into the argEvent slot, but was sometimes was entered into the relatedToEvent slot.

The distribution of the attribute errors is shown in table 5.

5 Discussion

Clearly recall and precision figures in comparing annotators' results with the answer keys are not as high as one would like. For stage I, the figures for entities alone are quite reasonable (77/81 recall/precision on average). The attribute figures are (60/64 recall/precision on average) are low, but these figures are artificially lowered by, e.g., the presence of correct implicit temporal relations in the responses but not in the keys. However, for stage II, the figures for temporal relations are low (40/68 recall/precision on average) and would need to be improved before a large scale annotation exercise can be undertaken. As noted above, recall on temporal relations is effectively bounded by the square of the recall on entities, so here we are effectively limited to 60% recall.

The causes of these problems and possible avenues for improvement include:

Imprecision/incompleteness of the guidelines The pilot study revealed misunderstandings on the part of the annotators as to the extent of text to be marked up, and confusions about, e.g. definitions of which events to be annotated. More explanation and further exemplification should help to avoid these problems.

Annotator Understanding of the Task While lack of clarity and comprehensiveness in the guidelines may have contributed to annotator mistakes, this is a difficult task and more extensive training of the annotators should also help to reduce errors.

Intrinsic difficulty of identifying which temporal relation holds Consider the sentence *All 75 people on board the Aeroflot Airbus died when it ploughed into a Siberian mountain in March 1994.* Is the relation between passengers dying and the plane crash one of causality and given that, did the passengers die, *after* the plane crash? Or is a plane crash an event that contains many subevents and is the death of the passengers part of it, which would imply that the death occurred *during* the plane crash? Or did they happen roughly at the same time, a relation covered by our temporal relation *simultaneous*? This example shows the type of genuine ambiguity encountered in our trial corpus. Aside from further exemplification and convention in the guidelines, a further possibility is that the set of temporal relations be broadened to facilitate decisions in difficult cases.

Annotator fatigue The phase during which the temporal closure is interactively computed is error prone due to the sheer number of questions asked. By intelligently analyzing existing chains of temporal relations in the temporal model and first asking those questions that might link existing chains and hence yield the largest number of inferences possible, the overall number of questions asked could be reduced. Allowing annotators to backtrack and revise their responses during the interactive cycle of completing the model, rather than having to do it correctly in one attempt, would also help in this regard.

Annotator carelessness Certain simple forms of annotator error could be reduced by improving the annotation tool with, for example, edit-checking, so that non-existing events IDs cannot be entered into the temporal relation attribute.

6 Conclusion

The task of annotating events, times and temporal relations in text is clearly not an easy one. Nevertheless we believe the pilot study reported here demonstrates the feasibility of carrying out this task. Further work needs to be done to increase the accuracy and ease with which annotators can perform the task. But clear indications of how this can be done have emerged from our study. We hope to produce a larger annotated corpus, but need to find the resources to do this.

Acknowledgements The authors would like to thank Kalina Bontcheva, Roberta Catizone, Patrick Herring, Diana Maynard, Horacio Saggion and Nick Webb and for help annotating texts.

References

L. Ferro, I. Mani, B. Sundheim, and G. Wilson. 2000. Tides temporal annotation guidelines. Tech. Rep. MTR 00W0000094, The MITRE Corporation, October. Draft-Version 1.0.

MUC6. 1995. *Proc. of the 6^{th} Message Understanding Conf. (MUC-6).* Morgan Kaufman.

MUC7. 1998. *Proc. of the 7^{th} Message Understanding Conf. (MUC-7).* Morgan Kaufman.

A. Setzer and R. Gaizauskas. 2000a. Annotating Events and Temporal Information in Newswire Texts . In *Proc. of the 2^{nd} Int. Conf. on Language Ressources and Evaluation (LREC2000).*

A. Setzer and R. Gaizauskas. 2000b. Building a Temporally Annotated Corpus for Information Extraction. In *Proc. of the Information Extraction Meets Corpus Linguistics Workshop at the 2^{nd} Int. Conf. on Language Ressources and Evaluation (LREC2000).*

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proc. of the 6^{th} Message Understanding Conf. (MUC-6).* Morgan Kaufmann.