Annotating and measuring temporal relations in texts

Philippe Muller
IRIT, Université Paul Sabatier
Toulouse, France
muller@irit.fr

Xavier Tannier
IRIT, Université Paul Sabatier
Toulouse, France
tannier@emse.fr

Abstract

This paper focuses on the automated processing of temporal information in written texts, more specifically on relations between events introduced by verbs in finite clauses. While this problem has been largely studied from a theoretical point of view, it has very rarely been applied to real texts, if ever, with quantified results. The methodology required is still to be defined, even though there have been proposals in the strictly human annotation case. We propose here both a procedure to achieve this task and a way of measuring the results. We have been testing the feasibility of this on newswire articles, with promising results.

1 Annotating temporal information

This paper focuses on the automated annotation of temporal information in texts, more specifically relations between events introduced by finite verbs. While the semantics of temporal markers and the temporal structure of discourse are well-developed subjects in formal linguistics (Steedman, 1997), investigation of quantifiable annotation of unrestricted texts is a somewhat recent topic. The issue has started to generate some interest in computational linguistics (Harper et al., 2001), as it is potentially an important component in information extraction or question-answer systems. A few tasks can be distinguished in that respect:

- detecting dates and temporal markers
- detecting event descriptions
- finding the date of events described
- figuring out the temporal relations between events in a text

The first task is not too difficult when looking for dates, e.g. using regular expressions (Wilson et al., 2001), but requires some syntactic analysis in a larger framework (Vazov, 2001; Shilder and Habel, 2001). The second one raises more difficult, ontological questions; what counts as an event is not uncontroversial (Setzer, 2001): attitude reports, such as beliefs, or reported speech have an unclear status in that respect. The third task adds another level of complexity: a lot of events described in a text do not have an explicit temporal stamp, and it is not always possible to determine one, even when taking context into account (Filatova and Hovy, 2001). This leads to an approach more suited to the level of underspecification found in texts: annotating relations between events in a symbolic way (e.g. that an event e_1 is before another one e_2). This is the path chosen by (Katz and Arosio, 2001; Setzer, 2001) with human annotators. This, in turn, raises new problems. First, what are the relations best suited to that task, among the many propositions (linguistic or logical) one can find for expressing temporal location ? Then, how can an annotation be evaluated, between annotators, or between a human annotator and an automated system ? Such annotations cannot be easy to determine automatically anyway, and must use some level of discourse modeling (cf. the work of (Grover et al., 1995)).

We want to show here the feasibility of such an effort, and we propose a way of evaluating the success or failure of the task. The next section will precise why evaluating this particular task is not a trivial question. Section 3 will explain the method used to extract temporal relations, using also a form of symbolic inference on available temporal information (section 4). Then section 5 discusses how we propose to evaluate the success of the task, before presenting our results (section 6).

2 Evaluating annotations

What we want to annotate is something close to the temporal model built by a human reader of a text; as such, it may involve some form of reasoning, based on various cues (lexical or discursive), and may be expressed in several ways. As was noticed by (Setzer, 2001), it is difficult to reach a good agreement between human annotators, as they can express relations between events in different, yet equivalent, ways. For instance, they can say that an event e_1

happens during another one e_2, and that e_2 happens before e_3, leaving implicit that e_1 too is before e_3, while another might list explicitly all relations. One option could be to ask for a relation between all pairs of events in a given text, but this would be demanding a lot from human subjects, since they would be asked for $n \times (n-1)/2$ judgments, most of which would be hard to make explicit. Another option, followed by (Setzer, 2001) (and in a very simplified way, by (Katz and Arosio, 2001)) is to use a few rules of inference (similar to the example seen in the previous paragraph), and to compare the closures (with respect to these rules) of the human annotations. Such rules are of the form "if r1 holds between x and y, and r2 holds between y and z, then r3 holds between x and z". Then one can measure the agreement between annotations with classical precision and recall on the set of triplets (event x,event y,relation). This is certainly an improvement, but (Setzer, 2001) points out that humans still forget available information, so that it is necessary to help them spell out completely the information they should have annotated. Setzer estimates that an hour is needed on average for a text with a number of 15 to 40 events.

Actually, this method has two shortcomings. First, the choice of temporal relations proposed to annotators, i.e. "before", "after", "during", and "simultaneously". The latter is all the more difficult to judge as it lacks a precise semantics, and is defined as "roughly at the same time" ((Setzer, 2001), p.81). The second problem is related to the inferential model considered, as it is only partial. Even though the exact mental processing of such information is still beyond reach, and thus any claim to cognitive plausibility is questionable, there are more precise frameworks for reasoning about temporal information. For instance the well-studied Allen's relations algebra (see Figure 2). Here, relations between two time intervals are derived from all the possibilities for the respective position of those intervals endpoints (before, after or same), yielding 13 relations. What this framework can also express are more general relations between events, such as disjunctive relations (relation between event 1 and event 2 is relation A or relation B), and reasoning on such knowledge. We think it is important at least to *relate* annotation relations to a clear temporal model, even if this model is not directly used.

Besides, we believe that measuring agreement on the basis of a more complete "event calculus" will be more precise, if we accept to infer disjunctive relation. Then we want to give a better score to the annotation "A or B" when A is true, than to an annotation where nothing is said. Section 5 gives more details about this problem.

We will now present our method to achieve the task of annotating automatically event relations. This has been tested on a small set of French newswire texts from the Agence France Press.

3 A method for annotating temporal relations

We will now present our method to achieve the task of annotating automatically event relations. This has been tested on a small set of French newswire texts from the Agence France Press. The starting point was raw text plus its broadcast date. We then applied the following steps:

- part of speech tagging with Treetagger (Schmid, 1994), with some post-processing to locate some lexicalised prepositional phrases;
- partial parsing with a cascade of regular expressions analyzers (cf. (Abney, 1996); we also used Abney's Cass software to apply the rules)[1]. This was done to extract dates, temporal adjuncts, various temporal markers, and to achieve a somewhat coarse clause-splitting (one finite verb in each clause) and to attach temporal adjuncts to the appropriate clause (this is of course a potentially large source of errors). Relative clauses are extracted and put at the end of their sentence of origin, in a way similar to (Filatova and Hovy, 2001). Table 1 gives an idea of the kind of temporal information defined and extracted at this step and for which potentially different temporal interpretations are given (for now, temporal focus is always the previously detected event; this is obviously an over-simplification).
- date computation to precise temporal locations of events associated with explicit, yet imprecise, temporal information, such as dates relative to the time of the text (e.g. *last Monday*).
- for each event associated to a temporal adjunct, a temporal relation is established (with a date when possible).
- a set of discourse rules is used to establish possible relations between two events appearing consecutively in the text, according to the tenses of the verbs introducing the events. These rules for French are similar to rules for English proposed in (Grover et al., 1995; Song and Cohen, 1991; Kameyama et al., 1993), but

[1] We have defined 89 rules, divided in 29 levels.

are expressed with Allen relations instead of a set of *ad hoc* relations (see Table 1 for a subset of the rules). These rules are only applied when no temporal marker indicates a specific relation between the two events.

- the last step consists in computing a fixed point on the graph of relations between events recognized in the text, and dates. We used a classical path-consistency algorithm (Allen, 1984). More explanation is given section 4.

Allen relations are illustrated Figure 2. In the following (and Table 1) they will be abbreviated with their first letters, adding an "i" for their inverse relations. So, for instance, "before" is "b" and "after" is "bi" (b(x,y)≡ bi(y,x)). Table 1 gives the disjunction of possible relations between an event e1 with tense X and a event e2 with tense Y following e1 in the text. This is considered as a first very simplified discourse model. It only tries to list plausible relations between two consecutive events, when there is no marker than could explicit that relation. For instance a simple past e_1 can be related with *e, b, m, s, d, f, o* to a following simple past event e_2 in such a context (roughly saying that e_1 is before or during e_2 or meets or overlaps it). This crude model is only intended as a basis, which will be refined once we have a larger set of annotated texts. This will be enriched later with a notion of temporal focus, following for instance (Kameyama et al., 1993; Song and Cohen, 1991), and a notion of temporal perspective necessary to capture more complex tense interactions.

The path consistency algorithm is detailed in the next section.

4 Inferring relations between events

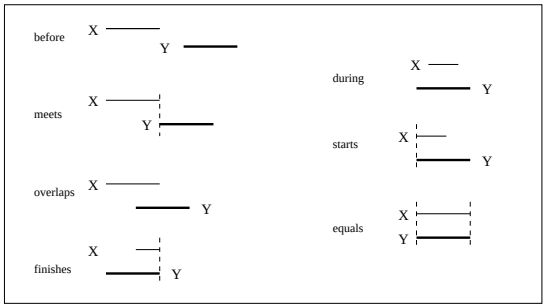

Figure 2: Allen Relations between two intervals X and Y (Time flies from left to right)

We have argued in favor of the use of Allen relations for defining annotating temporal relations, not only because they have a clear semantics, but also because a lot of work has been done on inference procedures over constraints expressed with these relations. We therefore believe that a good way of avoiding the pitfalls of choosing relations for human annotation and of defining inference patterns for these relations is to define them from Allen relations and use relational algebra computation to infer all possible relations between events of a text (that is saturate the constraint graph, see below), both from a human annotation and an annotation given by a system, and then to compare the two. In this perspective, any event is considered to correspond to a convex time interval.

The set of all relations between pairs of events is then seen as a graph of constraints, which can be completed with inference rules. The saturation of the graph of relations is not done with a few handcrafted rules of the form (relation between e1 and e2) + (relation between e2 and e3) gives (a simple relation between e1 and e3) (Setzer, 2001; Katz and Arosio, 2001) but with the use of the full algebra of Allen relation. This will reach a more complete description of temporal information, and also gives a way to detect inconsistencies in an annotation.

An algebra of relation can be defined on any set of relations that are mutually exclusive (two relations cannot hold at the same time between two entities) and exhaustive (at least one relation must hold between two given entities). The algebra starts from a set of base relations U= $\{r_1, r_2, ...\}$, and a general relation is a subset of U, interpreted as a disjunction of the relations it contains. From there we can define union and intersection of relations as classical set union and intersection of the base relations they consist of. Moreover, one can define a composition of relations as follows:

$$(r_1 \circ r_2)(x, z) \leftrightarrow \exists y \; r_1(x, y) \wedge r_2(y, z)$$

By computing beforehand the 13×13 compositions of base relations of U, we can compute the composition of any two general relations (because $r \cap r' = \emptyset$ when r, r' are basic and $r \neq r'$):

$$\{r_1, r_2, ...r_k\} \circ \{s_1, s_2, ...s_m\} = \bigcup_{i,j} (r_i \circ s_j)$$

Saturating the graph of temporal constraints means applying these rules to all compatible pairs of constraints in the graph and iterating until a fixpoint is reached. The following, so-called "path-consistency" algorithm (Allen, 1984) ensures this fixpoint is reached:

date(1/2) : non absolute date ("march 25th", "in June").

dateabs : absolute date "July 14th, 1789".

daterelST : date, relative to utterance time ("two years ago").

daterelTF : date, relative to temporal focus ("3 days later").

datespecabs : absolute date, with imprecise reference ("in the beginning of the 80s").

datespecrel : relative date, special forms (months, seasons).

dur : basic duration ("during 3 years").

dur2 : duration with two dates (*from February, 11 to October, 27*...).

durabs : absolute duration ("starting July 14").

durrelST : relative duration, w.r.t utterance time ("for a year").

durrelTF : relative duration, w.r.t temporal focus ("since").

tatom : temporal atom (*three days*, *four years*, ...).

Figure 1: Temporal elements extracted by *shallow parsing* (with examples translated from French)

e1/e2	imp	pp	pres	sp
imp	o, e, s, d, f, si, di, fi	bi, mi, oi	e, b	o, d, s, f, e, si, di, fi
pp	b, m, o, e, s, d, f	b, m, o, e, s, d, f, bi, mi	e, b	b, m, o
pres	U	U	b, m, o, si, di, fi, e	U
sp	b, m, o, e, s, d, f	e, s, d, f, bi, mi	e, b	e, b, m, s, d, f, o

Table 1: Some Discursive temporal constraints for the main relevant tenses, sp=simple past and perfect, imp=French imparfait, pp=past perfect, pres=present

Let $\begin{cases} A = \text{the set of all edges of the graph} \\ N = \text{the set of vertices of the graph} \\ U = \text{the disjunction of all 13 Allen relations} \\ R_{m,n} = \text{the current relation between nodes } m \text{ and } n \end{cases}$

1. $changed = 0$
2. for all pair of nodes $(i, j) \in N \times N$ and for all $k \in N$ such that $((i, k) \in A \wedge (k, j) \in A)$
   (a) $R_{1i,j} = (R_{i,k} \circ R_{k,j})$
   (b) if no edge (a relation $R_{2i,j}$) existed before between i and j, then $R_{2i,j} = U$
   (c) intersect: $R_{i,j} = R_{1i,j} \cap R_{2i,j}$
   (d) if $R_{i,j} = \emptyset$ (inconsistency detected) then : error
   (e) if $R_{i,j} = U$ (=no information) do nothing else update edge $changed = 1$
3. if $changed = 1$, then go back to 1.

It is to be noted that this algorithm is correct: if it detects an inconsistency then there is really one, but it is incomplete in general (it does not necessarily detect an inconsistent situation). There are sub-algebras for which it is also complete, but it remains to be seen if any of them can be enough for our purpose here.

5 Measuring success

In order to validate our method, we have compared the results given by the system with a "manual" annotation. It is not really realistic to ask humans (whether they are experts or not) for Allen relations between events. They are too numerous and some are too precise to be useful alone, and it is probably dangerous to ask for disjunctive information. But we still want to have annotation relations with a clear semantics, that we could link to Allen's algebra to infer and compare information about temporal situations. So we have chosen relations similar to that of (Bruce, 1972) (as in (Li et al., 2001)), who inspired Allen; these relations are equivalent to certain sets of Allen relations, as shown Table 2. We thought they were rather intuitive, seem to have an appropriate level of granularity, and since three of them are enough to describe situations (the other 3 being the converse relations), they are not to hard to use by naive annotators.

To abstract away from particulars of a given annotation for some text, and thus to be able to compare the underlying temporal model described by an annotation, we try to measure a similarity between annotations given by a system and human annotations, from the saturated graph of detected temporal relations in each case (the human graph is saturated after annotation relations have been translated as equivalent disjunctions of Allen relations). We do not want to limit the comparison to "simple" (base) relations, as in (Setzer, 2001), because it makes the evaluation very dependent on the choice of relations, and we also want to have a gradual measure of the imprecision of the system annotation. For instance, finding there is a "before or during" relation between two events is better than proposing "after" is the human put down "before", and it is less good

BEFORE	$\forall\, i\, \forall\, j\ (i \text{ before } j \ \Leftrightarrow\ ((i\ b\ j)\ \vee\ (i \text{ m } j)))$
AFTER	$\forall\, i\, \forall\, j\ (i \text{ after } j \ \Leftrightarrow\ ((i\ bi\ j)\ \vee\ (i \text{ mi } j)))$
OVERLAPS	$\forall\, i\, \forall\, j\ (i \text{ overlaps } j \ \Leftrightarrow\ ((i \text{ o } j)))$
IS_OVERLAPPED	$\forall\, i\, \forall\, j\ (i \text{ is\_overlapped } j \ \Leftrightarrow\ ((i \text{ oi } j)))$
INCLUDES	$\forall\, i\, \forall\, j\ (i \text{ includes } j \ \Leftrightarrow\ ((i \text{ di } j)\ \vee\ (i \text{ si } j)\ \vee\ (i \text{ fi } j)\ \vee\ (i \text{ e } j)))$
IS_INCLUDED	$\forall\, i\, \forall\, j\ (i \text{ is\_included } j \ \Leftrightarrow\ ((i \text{ d } j)\ \vee\ (i \text{ s } j)\ \vee\ (i \text{ f } j)\ \vee\ (i \text{ e } j)))$

Table 2: Relations proposed for annotation

than the correct answer "before".

Actually we are after two different notions. The first one is the consistency of the system's annotation with the human's: the information in the text is compatible with the system's annotation, i.e. the former implies the latter. The second notion is how precise the information given by the system is. A very disjunctive information is less precise than a simple one, for instance (a or b or c) is less precise than (a or b) if a correct answer is (a).

In order to measure these, we propose two elementary comparison functions between two sets of relations S and H, where S is the annotation proposed by the system and H is the annotation inferred from what was proposed by the human.

$$\text{finesse} = \frac{|S \cap H|}{|S|} \qquad \text{coherence} = \frac{|S \cap H|}{|H|}$$

The global finesse score of an annotation is the average of a measure on all edges that have information according to the human annotation (this excludes edges with the universal disjunction U) once the graph is saturated, while coherence is averaged on the set of edges that bear information according to the system annotation.

Finesse is intended to measure the quantity of information the system gets, while coherence gives an estimate of errors the system makes with respect to information in the text. Finesse and coherence thus are somewhat similar respectively to recall and precision, but we decided to use new terms to avoid confusion ("precision" being an ambiguous term when dealing with gradual measures, as it could mean how close the measure is to the maximum 1).

Obviously if S=H on all edges, all measures are equal to 1. If the system gives no information at all, S is a disjunction of all relations so $H \subseteq S$, $H \cap S = H$ and coherence=1, but then finesse is very low.

These measures can of course be used to estimate agreement between annotators.

6 Results

In order to see whether the measures we propose are meaningful, we have looked at how the measures behave on a text "randomly" annotated in the following way: we have selected at random pairs of events in a text, and for each pair we have picked a random annotation relation. Then we have saturated the graph of constraints and compared with the human annotation. Results are typically very low, as shown on a newswire message taken as example Table 3.

We have then made two series of measures: one on annotation relations (thus disjunctions of Allen relations are re-expressed as disjunctions of annotation relations that contains them), and one on equivalent Allen relations (which arguably reflects more the underlying computation, while deteriorating the measure of the actual task). In the first case, an Allen relation answer equals to *b or d or s* between two events is considered as "before or is_included" (using relations used by humans) and is compared to an annotation of the same form.

We then used finesse and coherence to estimate our annotation made according to the method described in the previous sections. We tried it on a still limited[2] set of 8 newswire texts (from AFP), for a total of 2300 words and 160 events, comparing to the English corpus of (Setzer, 2001), which has 6 texts for less than 2000 words and also about 160 events. Each one of these texts has between 10 and 40 events. The system finds them correctly with precision and recall around 97%. We made the comparison only on the correctly recognized events, in order to separate the problems. This course limits the influence of errors on coherence, but handicaps finesse as less information is available for inference.

The measures we used were then averaged on the number of texts. This departs from what could be considered a more standard practice, summing everything and dividing by the number of comparisons made. The reason behind this is we think comparing two graphs as comparing two temporal models of a text, not just finding a list of targets in a set of texts. It might be easier to accept this if one remembers that the number of possible relations between n events is $n(n-1)/2$. A text t1 with k more

[2]We are still annotating more texts manually to give more significance to the results.

	Finesse	Coherence
annotation relations	0.114	0.011
Allen relations	0.083	0.094

Table 3: Example of evaluation on a "random" annotation

events than a text t2 will thus have about k^2 times more importance in a global score, and we find confusing this non-linear relation between the size of a text and its weight in the evaluation process. Therefore, both finesse and coherence are generalized as global measure of a temporal model of a text. It could then be interesting to relate temporal information and other features of a given text (size being only one factor).

Results are shown Table 4. These results seem promising when considering the simplifications we have made on every step of the process. Caution is necessary though, given the limited number of texts we have experimented on, and the high variation we have observed between texts. At this stage we believe the quality of our results is not that important. Our main objective, above all, was to show the feasibility of a robust method to annotate temporal relations, and provide useful tools to evaluate the task, in order to improve each step separately later. Our focus was on the design of a good methodology.

If we try a first analysis of the results, sources of errors fall on various categories. First, a number of temporal adverbials were attached to the wrong event, or were misinterpreted. This should be fine-tuned with a better parser than what we used. Then, we have not tried to take into account the specific narrative style of newswire texts. In our set of texts, the present tense was for instance used in a lot of places, sometimes to refer to events in the past, sometimes to refer to events that were going to happen at the time the text was published. However, given the method we adopted, one could have expected better coherence results than finesse results. It means we have made decisions that were not cautious enough, for reasons we still have to analyze. One potential reason is that relations offered to humans are maybe too vague in the wrong places: a lot of information in a text can be asserted to be "strictly before" something else (based on dates for instance), while human annotators can only say that events are "before or meets" some other event; each time this is the case, coherence is only 0.5.

It is important to note that there are few points of comparison on this problem. To the best of our knowledge, only (Li et al., 2001) and (Mani and Wilson, 2000) mention having tried this kind of annotation, as a side job for their temporal expressions mark-up systems. The former considers only relations between events within a sentence, and the latter did not evaluate their method.

Finally, it is worth remembering that human annotation itself is a difficult task, with potentially a lot of disagreement between annotators. For now, our texts have been annotated by the two authors, with an *a posteriori* resolution of conflicts. We therefore have no measure of inter-annotator agreement which could serve as an upper bound of the performance of the system, although we are planning to do this at a later stage.

7 Conclusion

The aim of this study was to show the feasibility of annotating temporal relations in a text and to propose a methodology for the task. We thus define a way of evaluating the results, abstracting away from variations of human descriptions for similar temporal situations. Our preliminary results seem promising in this respect. Obviously, parts of the method need some polishing, and we need to extend the study to a larger data set. It remains to be seen how improving part of speech tagging, syntactic analysis and discourse modeling can influence the outcome of the task. Specifically, some work needs to be done to evaluate the detection of temporal adjuncts, a major source of information in the process. We could also try to mix our symbolic method with some empirical learning. Provided we can collect more annotated data, it would be easy to improve the discourse model by (at least local) optimization on the space of possible rules, starting with our own set. We hope that the measures of temporal information we have used will help in all these aspects, but we are also planning to further investigate their properties and that of other candidate measures not considered here.

References

Steven Abney, 1996. *Corpus-Based Methods in Language and Speech*, chapter Part-of-Speech Tagging and Partial Parsing. Kluwer Academic Publisher.

J. Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154.

B. Bruce. 1972. A model for temporal references

	Finesse	Standard Deviation	Coherence	SD
annotation relations	0.477499	0.286781	0.449899	0.175922
Allen relations	0.448222	0.289856	0.495755	0.204974

Table 4: Evaluation

and its application in a question answering program. *Artificial Intelligence*, 3(1-3):1–25.

Elena Filatova and Eduard Hovy. 2001. Assigning time-stamps to event-clauses. In Harper et al. (Harper et al., 2001).

Claire Grover, Janet Hitzeman, and Marc Moens. 1995. Algorithms for analysing the temporal structure of discourse. In *Sixth International Conference of the European Chapter of the Association for Computational Linguistics*. ACL.

Lisa Harper, Inderjeet Mani, and Beth Sundheim, editors. 2001. *ACL Workshop on Temporal and Spatial Information Processing*, 39^{th} Annual Meeting and 10^{th} Conference of the European Chapter. Association for Computational Linguistics.

M. Kameyama, R. Passonneau, and M. Poesio. 1993. Temporal centering. In *Proceedings of ACL 1993*, pages 70–77.

Graham Katz and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In Harper et al. (Harper et al., 2001), pages 104–111.

W. Li, K-F. Wong, and C. Yuan. 2001. A model for processing temporal reference in chinese. In Harper et al. (Harper et al., 2001).

I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of ACL 2000*.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Andrea Setzer. 2001. *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield, UK.

Franck Shilder and Christopher Habel. 2001. From temporal expressions to temporal information: Semantic tagging of news messages. In Harper et al. (Harper et al., 2001), pages 65–72.

F. Song and R. Cohen. 1991. Tense interpretation in the context of narrative. In *Proceedings of AAAI'91*, pages 131–136.

Mark Steedman. 1997. Temporality. In J. Van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*. Elsevier Science B.V.

Nikolai Vazov. 2001. A system for extraction of temporal expressions from french texts based on syntactic and semantic constraints. In Harper et al. (Harper et al., 2001).

George Wilson, Inderjeet Mani, Beth Sundheim, and Lisa Ferro. 2001. A multilingual approach to annotating and extracting temporal information. In Harper et al. (Harper et al., 2001).