Using Inference for Evaluating Models of Temporal Discourse

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

Axel Reymonet
IRIT, Université Paul Sabatier, Toulouse, France
reymonet@irit.fr

Abstract

This paper addresses the problem of building and evaluating models of the temporal interpretation of a discourse in Natural Language. The extraction of temporal information is a complicated task as it is not limited to finding pieces of information at specific places in a text. A lot of temporal data is made of relations between events, or relations between events and dates. Building such information is highly context-dependent, taking into account information more than a sentence at a time. Moreover it is not clear what the target representation should be: the way it is done by human beings is still a subject of study in itself. It seems to require some sort of reasoning, either purely temporal or involving complex world knowledge. This is the reason why evaluating this task is also problematic when trying to design a system for it. We present a method for enriching the detection of event-to-event relations with a basic reasoning model, that can be also used for helping to compare the extraction of temporal information by a system and by a human being. We have experimented with this method on a set of texts, comparing a very basic model of tense interpretation with a more complex model inspired by the Reichenbach's well-known theory of narrative discourse.

1. Introduction

This paper focuses on the extraction of temporal informations in texts, and on the issue of evaluating a system automating the task. While the semantics of temporal markers and the temporal structure of discourse are well-developed subjects in formal linguistics [20], the investigation of quantifiable systematic annotations of unrestricted texts is a somewhat recent topic. The issue has started to generate some interest in computational linguistics [7], as it is potentially an important component in information ex-

traction, automatic summarization or question-answer systems, generating some international effort towards a standard mark-up scheme for temporal information, TimeML (www.timeml.org). This effort has been only partially related to the vast literature on temporal representation and reasoning in Artificial Intelligence (AI) and Knowledge Representation (KR).

While detecting dates and temporal expressions (such as *after a few days, last year, at two o'clock,...*) is not a very difficult problem, relating such expressions with events introduced by verbs require some syntactic analysis [21, 18]. Stamping events with a precise date is even more difficult, as this kind of information is not always available or is highly contextual [5]. Finding events denoted by nominal phrases (e.g. *World War I, the destruction of Troy*) is not an easy task in general either, if they are not present in a typical prepositional phrase such as *after World War I*, and requires a specific lexicon.

Then there is an amount of information that is expressed only with relations between temporal entities (something happens before/during/after something else), and this level of vagueness raises new problems. First, the relations best suited to that task must be chosen among many propositions, (linguistically oriented or more concerned with knowledge representation issues) Then, the target representation must be compared and evaluated with respect to some standard. But the way it is done by human beings is still a subject of study in itself [15], and while it seems to require some sort of reasoning, either purely temporal or involving complex world knowledge, it is still unclear what representations humans have of the temporal ordering of events in a text they read.

The main proposal made for human annotation by [17], and imported in the TimeML recommendations, is to have a set of relations associated with a few rules of inference that are supposed to give a transitive closure of an annotation that can be the target of comparisons. This used a specific

model that was somewhat unaware of the large KR/AI tradition of temporal representation and reasoning. The use of a more well-studied inference model is advocated in [13], also as an arguably cleaner way of separating the problem of the intended representation of information from the process of handling and evaluating it. We use it here to compare a few strategies in extracting the temporal ordering of events in natural language texts.

The paper is organised as follows: we present the different linguistic levels at which some temporal information can be expressed, then we discuss the problem of comparing temporal annotations and the need for an inference model. We also present a few procedures to extract temporal relations between events and how they compare according to the methodology we proposed.

2. Structuring temporal information expressed in NL

The kind of temporal information that can be found in texts involves three classical levels: lexico-syntactic, semantic and pragmatic, and we deal with them separately.

At the lexico-syntactic level, we have specific temporal markers and patterns that group together expressions corresponding to similar semantic types of temporal adjuncts as follows (translated from their French counterparts):

- non absolute dates ("March 25th", "in june"),

- absolute dates "July 14th, 1789",

- dates, relative to utterance time ("two years ago"),

- dates, relative to some temporal focus ("3 days later"),

- absolute dates, with imprecise reference ("in the beginning of the 80s"),

- basic durations ("during 3 years"),

- durations with two dates (*from February, 11 to October, 27…*),

- absolute durations ("starting July 14"),

- relative durations, w.r.t utterance time ("for a year"),

- relative durations, w.r.t temporal focus ("since"),

- temporal atoms (*three days, four years*, …).

At the semantic level, each type of adjunct (date or duration) gets values for each instantiated attribute among: starting time, ending time, duration, type (absolute/relative). This uses only parts of the TimeML coding standard for dates and durations that are relevant for the following treatments.

Then, according to each type of temporal adjunct, we try to establish a link between an event and any temporal adjunct present in the same syntactic clause (the event is before, after, or during a date, or receive a duration that can be used later on).

At the pragmatic level, which is the level we want to investigate more precisely and the one in need of a methodology, we handle the semantic of verb tenses with respect to the temporal structure of a discourse (how tenses are chained, and what this means for relations between events for instance), how temporal references evolve through the interpretation and the role of the structure of discourse[1]. More details are given sections 4, 5.

3. Evaluating annotations

### 3.1. The problem of comparing temporal models of the same text

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 that may be expressed in several ways. As was noticed by [17], 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 [17] is to use a few rules of inference (similar to the example seen above), and to compare the closures (with respect to these rules) of the human annotations. Such rules are of the form "if r_1 holds between x and y, and r_2 holds between y and z, then r_3 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 [17] 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, and subjects get tired of it very quickly.

---

[1]This aspect has not been studied here as it raises many more theoretical questions, see [3], and might not be ready yet for the kind of evaluation we have in mind; but see a preliminary report in [10].

3.2. Separating the inference model from the annotation scheme

The previously introduced method has two shortcomings. First, the choice of temporal relations proposed to annotators, i.e. "before", "after", "during", and "simultaneously", is arbitrary and somewhat ill defined (the latter is defined as "roughly at the same time", [17], p.81). The second
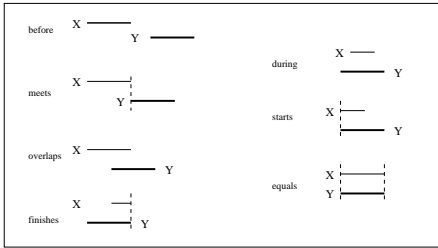
Figure 1. Allen Relations between two intervals X and Y (Time flows from left to right)

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 relation algebra (see Figure 1). Here, relations between two time intervals are derived from all the possibilities for the respective position of those interval endpoints (before, after or same), yielding 13 relations[2]. These relations are now the possible relations between events in the TimeML scheme. 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 disjunctions of atomic relations. 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 3.4 gives more details about this problem.

In order to validate the method, we have to compare the results given by the system with a "manual" annotation. It is not really realistic to ask humans (experts or not) for Allen relations between events. They are too numerous and some

---

[2]In the following (and Table 2) 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)).

are too precise to be useful alone, and it is probably dangerous to ask for disjunctive information (it remains to be seen what the TimeML conventions will yield in that respect, as they force the annotator to choose one and only one of Allen's relations). 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. Here we have chosen relations similar to that of [4] (as in [9]), who inspired Allen; these relations are equivalent to certain sets of Allen relations, as shown in Table 1. 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. This would have to be confirmed by an empirical study on a set of annotators.

Table 1 give their definition from Allen relations. Relations "includes" and "is_included" can be applied when two relations have the same temporal extent, for simplicity (since this is very rare), and it is also possible to indicate that two expressions refer to the *same* event (they are then treated as only one node).

3.3. The inference model

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 (i.e. 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) but with the use of the full algebra of Allen relations. This will reach a more complete description of temporal information, and also gives a way to detect inconsistencies in an annotation (which can be useful for a human annotator).

An algebra of relations 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 en-

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

Table 1. Relations proposed for annotation

tities). The algebra starts from a set of atomic 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 [2] ensures this fixpoint is reached:

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

1. $changed = False$

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; $changed = True$

3. if $changed$, then go back to 1.

This algorithm is proven to be correct: if it detects an inconsistency then there is really one, but incomplete in general (it does not necessarily detect an inconsistent situation).

Allen's original algorithm can be improved in various ways (using reference intervals to build clusters on which local consistency is enforced for instance), but this was not really necessary here since the graphs are rather small. There are sub-algebras for which this procedure is also complete, and it would be interesting to see if annotations are part of such a sub-algebras (see [16] for more details about temporal constraints and algorithms).

3.4. Comparing two temporal graphs

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 ones, from the saturated graphs of detected temporal relations. We do not want to limit the comparison to "simple" (atomic) relations, as in [17], because it makes the evaluation very dependent on the choice of relations, and we think it can give a misleading impression of how good the system performs (consider for instance the extreme case of a non-consistent annotation that could not be detected when looking only for basic relations).

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" if the human put down "before", and it is less good 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 can never contradict 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 use two elementary comparison functions between two sets of relations S and H (each set has as members the basic relations that constitutes the disjunction), 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 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.

Finesse is intended to measure the quantity of accurate 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.

4. Stages for the extraction of temporal relations

We will now present our method to achieve the task of annotating automatically event relations, before going through a small example. The starting point was raw text plus its broadcast date. We then applied the following steps prior to discourse interpretation:

- part of speech tagging, with some post-processing to locate some lexicalised prepositional phrases and mark specific lexical items (days, months,...);

- partial parsing with a cascade of regular expressions analysers (cf. [1]; we also used Abney's Cass software to apply the rules). 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 [5].

- 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.

At this point we have a sequence of events (recording also their types and tenses) with some relations between these events and dates (noted $t_1, t_2, ...$) recognized in the text:

$$input = \langle (e_1, \{(e_1 R_{1,j_1} t_{j_1}), ...\}), (e_2, \{...\}), ... \rangle$$

Then we consider the following steps:

1. *filtering* of events according to their lexical type (to exclude e.g. states, reports, or aspectual constructions and focus on "occurrences", in the TimeML terminology).

2. *chaining constraints*: 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 [6, 19, 8], but are expressed with Allen relations instead of a set of *ad hoc* relations. Let $tense(e_j)$ be the tense of the verbe denoting event i in the text. Then $V(tense(e_i), tense(e_{i-1}))$ denotes the possible relations between two successive events having the same tense as e_i and e_{i-1}, and it is taken from Table 2. For instance, if both event are simple past events, $V_{sp,sp} = \{e, b, m, s, d, f, o\}$

   These rules are only applied when no temporal marker (such as "when", "after that", "then",...) indicates a specific relation between the two events or when computing dates does not conflict with the result (leading to an inconsistent temporal graph).

3. *temporal perspective handling*: introducing state variables following Reichenbach [14] : E: the event reference point (the current narrative location, which is an event), S: speech time, R: the temporal perspective point (another event). Rules are used to determine how these variables change during the interpretation, according to various cues, the main one being a change of tense. For instance, a series of past tense verbs have as reference point the last event introduced (so E=R and $E < e_i$ for each new simple past event e_i, which then becomes R) while the use of the pluperfect *shifts the temporal perspective point* (if e_j has tense pluperfect, $e_j < R$, and for each subsequent pp event, $E < e_j < R$). The relations due to Reichenbach are summed up Table 3. Actually we observed that the ordering relation he assumed is too strong in practice, so the one we used is less constrained, we replace precedence with $\{b, m, s, f, d, o, e\}$ (before or included or overlap). The only tenses that depend on a temporal perspective are the pluperfect and the future perfect. Details and an example are given below.

The basic algorithm only applies the second step (see example below). The more elaborate one derived from

Reichenbach does some filtering and then handles variables as follows (mixing chaining constraints with temporal perspective): $i-1$ is the index of the previously handled event, i is the current event index. At the beginning, E=R=the first event.

1. get all possible relations
$relV_{i-1,i} = V(tense(e_i), tense(e_{i-1}))$
between tenses of events $i-1$ and i from Table 2

2. if a relation $(e_i relD_{i-1,i} e_{i-1})$ can be inferred from input (by date computation),

   (a) $rel = relD_{i-1,i} \cap relV_{i-1,i}$ (intersect the relations)

   (b) if $(rel \neq \emptyset)$, (relate current event i with E, R and S, according to its tense, following table 3, and to $i-1$, if it's not already E or R)

       i. add to graph: $(E\ rel_{E,i}\ e_i)$, $(R\ rel_{R,i}\ e_i)$, $(S\ rel_{S,i}\ e_i)$, $(e_{i-1}\ rel\ e_i)$

       ii. if (there is change of tense to a new one not dependant on R), then (reinitialise R and E).

       iii. else if $(rel \in \{\ b\ ,\ m\ ,\ o\ \})$, then $(E = e_i)$

       iv. else if $(rel \in \{\ bi\ ,\ mi\ ,\ oi\ \})$, $(R = E)$; $(E = e_i)$

   (c) else $(rel = \emptyset)$, record $(e_{i-1}\ relD_{i-1,i}\ e_i)$ (just keep the date constraint if any was found) and reinitialise R et E (we found an inconsistency so there must have been a narrative shift)

3. else $[\neg\ (\ \exists\ relD_{i-1,i}\ )]$, update S, R, E [3] as in 2)b) and the temporal graph.

5. Example of the processing pipeline

We will use the following text, extracted (and slightly modified for simplicity, since most sentences in the texts are very long) from a text of our corpus, dated Monday, June 2nd, 2003.

Les chances de percée au sommet d'Aqaba <u>ont diminué</u> (*lundi*). Le Premier Ministre israélien Ariel Sharon, son homologue palestinien Mahmoud Abbas et M. Bush se <u>sont réunis</u>, et <u>ont échoué</u> à trouver un accord sur le calendrier à suivre. La possibilité d'un cessez-le-feu <u>avait été évoquée</u> mais <u>avait été rejetée</u>.

Translation: *Chances of a breakthrough at the Aqaba summit <u>decreased</u> (Monday). PM A. Sharon, PM M. Abbas and P. G. Bush <u>met</u>, and <u>failed</u> to agree on a schedule. The possibility of a cease-fire had been <u>considered</u> but <u>had been rejected</u>.*

---

[3]As before, update follows table 3.

The preprocessing consists in tokenizing (separating words) lemmatizing and morphosyntactically tagging the text, yielding for the beginning of the second sentence (tags used are from the well-known Penn Treebank format, added with specific tags for verbs and time related words):

le/dta chance/nn de/of percée/nn à/a le/dta sommet/nn de/of Aqaba/nnp avoir/aux_pres diminuer/ver_ppas lundi/day...

Then the shallow parser can be applied. It is made of pattern matching rule of the form : rewrite a sequence of "dta nn" as "nx" (noun chunk), divided in stages where the output of each stage is the input of the next. The previous sentence would then be analysed as (brackets indicates the hierarchical structure; some simplifications in the structure are made for clarity, as there are many levels of noun phrases)

```
[c0
    [np2
        [np [nx [dta  le]
                [nn  chances]]]
        [of  de]
        [np [nx [nn  percée]]]
        [a  à]
        [np [nx
                [dta  le][nn  sommet]
                [name [of de][nnp  Aqaba]]]]]]]
    [vx  [ver_pc
            [aux_pres   avoir]
            [ver_ppas   diminuer]]]
    [cct  [daterelST   [day  lundi]]]
] ...
```

Here c0 indicates a basic clause, whith only one finite verb in it. The semantic interpretation of the sentence can thus be done, where any temporal adverbial (cct structures) is given a value if possible and is related to the event in the same clause, according to its type. Here "lundi" belongs to the category of date-relative-to-speech-time, and since it is used with a past-tensed verb it is computed as the monday before speech time, so is given the value (2003-5-26). Besides, as the adverbial phrase is a direct adjunct of the verb, it is assumed to be a simple localisation (so the event in included in the date). Every time another date is added to the interpretation, we compute any qualitative relation we can find with every date already introduced, using their computed value (including duration calculus is there is any). For the example text, the dates detected are (t0=date of the publication):

```
t3 : 2003-5-26   (lundi)
t0 : 2003-6-2
```

So we also have that $t_0 < t_3$ Here the system considered that a past tense used with "lundi" (monday) in the first sentence meant last monday before the publication of the text (while it is a (unsual) way of referring to the day of the publication in AFP news texts).

e1/e2	imp	pqp	pres	sp
imp	o, e, s, d, f, si, di, fi	bi, mi, oi	e, b	o, d, s, f, e, si, di, fi
pqp	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 2. Possible temporal relations between two successive events according to their tenses, for the main relevant tenses, sp=simple past and perfect, imp=French imparfait, pqp=pluperfect, pres=present; U stands for the universal relation (no information)

Tense of current event	relations between event i and E,R,S
Plus-que-parfait (pluperfect)	$E < i < R < S$
Passé simple (simple past)/Imparfait	$E = R < i < S$
Present	$i \subset S, E = R$
Simple Future	$S < E = R < i$
Futur antérieur (future perfect)	$S < E < i < R$

Table 3. Grammatical tense and Reichenbach perspective: relations between current event and E (last one), R (reference point) and S (speech time)

The last stage consists in determining the possible relations between successive events. We will first have a look at the basic algorithm. In the example, we have a sequence of clauses as follows (begin=1 stands for a fictitious event corresponding to the speech time), with number of event, tense of the verb (pp=past perfect, pqp=pluperfect), lemma of the verb.

```
1    begin
2    ver_pp     'diminuer'     (decrease)
4    ver_pp     'réunir'       (meet)
5    ver_pp     'échouer'      (fail)
8    ver_pqp    'évoquer'      (consider)
9    ver_pqp    'rejeter'      (reject)
```

The algorithm will then introduce a relation between speech time and each event according to its tense (past or future, so here $2 < 1, 4 < 1, 5 < 1,...$), add that e_1 is during t_0, and use table 2 to introduce relations between successive events. Event 4 follows event 2 so we look at the table for (pp,pp) which the same as (sp,sp) in current french, and find $2\{e, b, m, s, d, f, o\}4$. The same is found between 4 and 5. Then (pqp,pp) yields a relation between 5 and 8: $\{e, b\}$. Finally we have between 8 and 9, from (pqp,pqp): $\{b, m, o, e, s, d, f, bi, mi\}$. At each stage, the graph is saturated and if an inconsistency appears, we come back to the state before the last introduced event. Looking at the graph we see that nothing is inferred between 5 and 9, while the use of the pluperfect indicates that 9 have occcured before the sequence 2-5. This is the problem addressed in part by the method inspired by Reichenbach's work.

The interpretation following Reichenbach would consider two variables: E, the last event introduced, R the reference point and S, the speech time (=event 1). For the explanation, only the sequence of tenses is relevant since there is only one date and no temporal connector between clauses. This is how E and R evolve, and the relation between the current event and E, R and S, using Table 3 (for readability we left the symbol "<", but it should be seen as $\{b, m, o, e, s, d, f\}$):

+ event	E	R	relations	new E	new R
2(pp)			2<S	2	2
4(pp)	2	2	E<4<S	4	4
5(pp)	4	4	E<5<S	5	5
8(pqp)	-	5	8<R<S	8	5
9(pqp)	8	5	E<9<R	9	5

There are special rules for updating E and R when there is a change of tense depending whether the new tense needs a temporal perspective (plupefect, future perfect) (e.g. pp(5) to pqp(8), then R becomes the previous E, and E is reset to nothing), or not (all others tenses, then E and R are reset when there is a change).

6. Comparing theories of temporal interpretation

Our methodology has been used to compare a few strategies for the pragmatic level of discourse temporal interpretation. For each text we have 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). 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 limited set of newswire texts (from AFP), for a total of about 200 events. Each one of these texts has between 10 and 40 events. The human generated graphs have between 12 and 266 edges for an average of 115. The system average size was about 170 (for the last two methods combined, with more edges for the last one). We averaged this on the number of texts. Results are shown Table 4.

The first strategy is a "random" annotation made in the following way: to each pair of successive events in a text, we choose a random annotation relation. Then we saturate the graph of constraints and we compare with the human annotation. This strategy has poor results except for the coherence on Allen relations, which is very high maybe because random unrelated annotations do not produce a lot of coherent additional information and re-expressed as Allen relations they result in very disjunctive information (thus trivially coherent)[4].

The second one is a baseline using a similar strategy but assuming every event is in the order it has in the text (so instead of having a random relation between consecutive events, we always generate "before"). Again, coherence in Allen relations is surprisingly high, but the rest is very low.

The third one is the strategy based on simple tense chaining constraints. It was the first strategy we tried to check the feasability of this kind of study.

The last one is based on a filtered, localised Reichenbach model that performs better on coherence while damaging finesse a lot. This one takes less risk since it may "reset" the interpretation to record more local constraints and is perhaps more precise about the information it gives. Nonetheless, when it is converted back to the simplest annotation relations, there is not much difference with the more basic model, and it's even slightly worse. At first glance, it seems that the venerable ideas of Reichenbach are tricky to apply to real texts, where local coherence may not be ideal. We still have to analyse precisely what really happens here, but it means at least that we have to keep the measure on Allen relations to study more finely the methods used.

To the best of our knowledge, only [9] and [12] mention having tried this kind of annotation [5]. The former considers only relations between events in a same sentence, and the latter did not evaluate their method until they enrich it with a learning stage [11]. They indicate good results (about 75%) in both recall and precision on unambiguous (atomic) relations, but this is only on a partial ordering, and includes relations between dates and both events or dates (a much easier task, since dates are almost always explicitely, if not

unambiguously, related to an event). It seems to us that the measures we propose reflects more accurately the difficulty of the task.

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 two experts, 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. But we also did an experiment to see how human can agree on this task without too much training. We took 7 subjects to whom we explained the notion of relations between events and how to decide them and gave them a short newswire text to annotate (the text had 12 events, which were underlined in the text so only relations were to be added). The results were, as expected, not very good: on average on 42 comparisons[6], finesse = 0.51 and coherence is .49 for Allen relations, and finesse= .58 and coherence = .55 for annotation relations.

We are in the process of building a larger corpus of texts with precise annotation from experts, with an evaluation of the agreement between them[7].

Last, one could argue that trying to reach that level of precision in finding temporal relations in a text is irrealistic, and that humans don't do it and focus on the important temporal relations. This would have to be investigated but it appears difficult to implement something similar from a methodological point of view. Establishing a standard annotation for a text would still be an issue; moreover one runs the risk of getting only the obvious, explicit relations, while it has been shown that humans can agree on relations they had not seen by themselves, if they are asked a posteriori.

7. Conclusion

In this paper we have presented a method to handle temporal information in natural language texts, which tries to bridge the gap between natural language processing methods and knowledge representation techniques. Using well-studied inference processes is used for two tasks: improving the extraction of information, and helping the evaluation and comparison of the extracted information. Our pilot study shows the possibilities of combining techniques and leaves a lot of room for improvements and experimentation. We have chosen the simplest form of reasoning to stay within reasonable computational bounds but different models could be tried, as long as the target representation is linked but separated from the inference model. A lot of

---

[4]We have not investigated that unexpected effect yet.

[5]A lot of other proposals have not been evaluated beyond a set of fabricated examples, see among others [19, 8, 6].

[6]The measures are not symmetric because we consider only the events correctly found by one annotator against the other one, regarded as the correct annotation; however, note that precision(a,b)=coherence(b,a).

[7]Preliminary tests let us believe that measures around 0.7-0.8 should be reached, although not without effort.

Method	F (Allen)	C (Allen)	F (Ann. rel.)	C (Ann. rel.)
"Random" baseline	0.04	0.78	0.05	0.02
"Before" Baseline	0.03	0.84	0.03	0.015
Basic constraints	0.49	0.55	0.23	0.31
Modifi ed Reichenbach	0.195	0.75	0.215	0.225

Table 4. Comparing methods

tuning still has to be made to reach good results, but we also hope to help the task of the human in specifying temporal knowledge (as this is far from obvious from the agreement we observe between humans) and by bootstrapping the annotation either in a semi-automated way, or as a starting point for learning algorithm, as in [11].

References

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

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

[3] N. Asher and A. Lascarides. Temporal interpretation, discourse relations, and commonsense entailment. *Linguistics and Philosophy*, 16:437–493, 1993.

[4] B. Bruce. A model for temporal references and its application in a question answering program. *Artificial Intelligence*, 3(1-3):1–25, 1972.

[5] Elena Filatova and Eduard Hovy. Assigning timestamps to event-clauses. In Harper et al. [7].

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

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

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

[9] W. Li, K-F. Wong, and C. Yuan. A model for processing temporal reference in chinese. In Harper et al. [7].

[10] I. Mani and J. Pustejovsky. Temporal discourse models for narrative structure. In *ACL Workshop on Discourse Annotation*, Barcelona, Spain, 2004.

[11] I. Mani, B. Schiffman, and J. Zhang. Inferring temporal ordering of events in news. In *Proceedings of the Human Language Technology Conference (HLT-NAACL'03)*, 2003. (short paper).

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

[13] P. Muller and X. Tannier. Annotating and measuring temporal relations in texts. In *Proceedings of Coling 2004*, volume I, pages 50–56, Genève, 2004.

[14] H. Reichenbach. *Elements of Symbolic Logic*. McMillan, New York, 1947.

[15] W. Schaeken and P. N. Johnson-Laird. Strategies in temporal reasoning. *Thinking and Reasoning*, 6:193–219, 2000.

[16] Eddie Schwalb and Lluís Vila. Temporal constraints: A survey. *Constraints*, 3(2/3):129–149, 1998.

[17] Andrea Setzer. *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study*. PhD thesis, University of Sheffield, UK, September 2001.

[18] Franck Shilder and Christopher Habel. From temporal expressions to temporal information: Semantic tagging of news messages. In Harper et al. [7], pages 65–72.

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

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

[21] Nikolai Vazov. A system for extraction of temporal expressions from french texts based on syntactic and semantic constraints. In Harper et al. [7].