Temporally Anchoring and Ordering Events in News

Inderjeet Mani
Department of Linguistics
Georgetown University

Barry Schiffman
Department of Computer Science
Columbia University

Abstract

This paper describes a domain-independent approach to temporally anchoring and ordering events in news. The focus is on event-event and event-time linking. The approach involves mixed-initiative corpus annotation, with automatic preprocessing to identify clause structure, tense, aspect, and temporal adverbials. A controlled experiment reveals the capabilities of humans in ordering events in news. The paper then develops and evaluates two different approaches to machine learning of ordering information: learning from event-event ordering judgments, and learning from event-time anchoring judgments.

1 Introduction

The growing interest in practical NLP applications such as text summarization and question-answering places increasing demands on the processing of temporal information. In multi-document summarization of news articles, it is important to know the relative order of events so as to merge and present information from multiple news sources correctly. In question-answering (Voorhees 2002) (Pustejovsky et al. 2002), one would like to be able to ask when an event occurs, or what events occurred prior to a particular event. Such capabilities presuppose an ability to infer the temporal order of events in discourse.

A wealth of prior research by (Moens and Steedman 1988), (Passoneau 1988), (Webber 1988), (Hwang and Schubert 1992), (Kamp and Reyle 1993), (Lascarides and Asher 1993), (Allen 1995), (Hitzeman et al. 1995), (Kehler 2000) and others, has explored the different knowledge sources used in inferring the temporal ordering of events, including temporal adverbials, tense, aspect, rhetorical relations, pragmatic conventions, and background knowledge. For example, the narrative convention of events being described in the order in which they occur is followed in (1), but overridden by means of a rhetorical relation -- *Explanation* in (2)[1].

(1) Max *stood up*. John *greeted* him.

(2) Max *fell*. John *pushed* him.

The narrative convention can be viewed as a discourse relation, called the *Narration* relation in (Lascarides and Asher 1993). In addition to discourse relations, the ordering decisions humans carry out appear to involve a variety of knowledge sources, including tense (3a), aspect (3b), temporal adverbials (3c), and world knowledge (3d).

(3a) Max *entered* the room. He *had drunk* a lot of wine.

(3b) Max *entered* the room. Mary *was seated* behind the desk.

(3c) A drunken man *died* in the central Phillipines when he *put* a firecracker under his armpit.

[1] Examples 1 and 2 from (Lascarides and Asher 1993). Examples whose sources are not otherwise cited are taken from the North American News Corpus.

(3d) U. N. Secretary- General Boutros Boutros-Ghali Sunday *opened* a meeting of ....Boutros-Ghali *arrived* in Nairobi from South Africa, accompanied by …

As (Bell 1999) has pointed out, the narrative convention is not usually followed in the case of news stories; the temporal structure of news is dictated by perceived news value rather than chronology. Thus, the latest news is often presented first, with the possibility of multiple backward and forward movements through different time-frames. In addition, news often expresses multiple viewpoints, with commentaries, eyewitness recapitulations, etc., offered at different speech times.

Together, these observations suggest anecdotally that events in news stories may be very difficult to order by humans and machines. Assume, for the sake of argument, a naïve algorithm for ordering events. The first step would be to identify text units (sentences or clauses) with explicit time mentions, anchoring the event to the time value associated with the mention. The narrative convention along with tense and lexical aspect shifts might be used to temporally order events in successive text units. However, if the narrative convention isn't that strong, or if explicit time mentions are rare, the anchoring information from such an algorithm will be very sparse.

In this paper, we begin (Section 2) with a theoretical perspective on our work. In Section 3, we provide an outline of the relevance of temporally anchoring and ordering events to the TimeML annotation scheme. Then, we investigate by means of a human experiment how often the narrative convention is followed in a corpus of news examples. We also determine how rare explicit time mentions are. As a side-effect of the experiment, we identify the kinds of knowledge humans apparently bring to bear to help determine the order.

This initial investigation (Sections 4 and 5) allows us to characterize the relative difficulty of the temporal ordering problem, shedding light on which aspects are within reach of current computational methods. Based on the limited set of judgments from the experiment, we report on the use of shallow features involving time expressions, tense, aspect, etc. in a statistical classifier to temporally order pairs of clauses.

After this initial investigation, we address the sparseness of explicit times by inferring implicit times for each text unit. More precisely, for each finite clause, we identify the most likely 'reference' time (Reichenbach 1947) with respect to which the event in the clause can be anchored. Calculation of the reference time is done (Section 6) by means of a simple algorithm. The reference times are then corrected by a human. The human also anchors the event's time with respect to the reference time. Based on the training data, we explore statistical classifiers to compute the anchoring relation.

Finally, we construct partial orderings of events in a document using the explicit and implicit reference times (Section 7), and evaluate them.

2 Theoretical Perspective

2.1 Previous Work

Our work takes at its starting point a view of tense as anaphoric, i.e., referring to a time or set of times, a view which has its origin in the work of (Reichenbach 1947). Reichenbach makes a distinction between the point of speech (i.e., the time of the utterance), the point (i.e., time) of the event, and what he calls the point of reference (which we call the 'reference time'). Reichenbach argues that the past perfect tense introduces a reference time in between the speech time and the event time. Thus, in a sentence like "I had mailed the letter when John came and told me the news", the time of John's coming occurs after the event time of mailing the letter, and before the speech time, and coincides with the reference time. Reference times can also be explicit, as in the case where the above sentence is extended with the temporal adverbial "on Wednesday".

Type of Eventuality	Telic	Dynamic	Durative	Examples
State			+	know, have

Activity		+	+	march, paint
Accomplishment	+	+	+	destroy
Achievement	+	+		notice, win

Table 0: Analysis of Lexical Aspect

Work on lexical aspect derives from (Vendler 1967). A recent account by (Dorr and Olsen 1997) is based on analyses by Carlota Smith and others, and is shown in Table 0. While lexical aspect is useful, the aspectual class of a sentence can change due to compositional processes. For example, in "The regiment marched to Saigon", the activity has changed into an accomplishment.

A related issue that theoretical work has addressed is the influence of discourse-level information on the temporal ordering of events. The general intuition explaining the ordering difference between (1) and (3b) is that in the absence of a temporal adverbial or a discourse relation or a tense shift, a sentence with telic aspect (accomplishment or achievement) introduces a new reference time later than the current reference time (thereby instantiating narrative progression); non-telic sentences maintain the current reference time. In Discourse Representation Theory (Kamp and Reyle 1993), the ordering is addressed as part and parcel of the procedure for constructing a semantic interpretation of a sentence. The detailed algorithm actually confines itself to stative versus non-stative sentences. This flavor of approach has been criticized by (Dowty 1986) for its being based on lexical rather than compositional aspect; thus, you need the sentence meaning to decide whether a sentence is telic or not, but in order to construct the sentence meaning in DRT, you need to know whether you are dealing with a telic or non-telic eventuality, thus introducing a potential circularity.

The work of (Lascarides and Asher 1993) takes a non-anaphoric view of tense. Here tense orders the event just with respect to the speech time; the temporal ordering is derived entirely based on discourse relations. The approach thus views the temporal ordering as being based on 'pragmatics' rather than 'semantics'. The discourse relations are given a fairly detailed temporal semantics. For example, (Bras et al. 2001), developing work by Lascarides and Asher, define a *Narration* relation holding between constituents A and B if they don't have significant spatio-temporal gaps. In particular, *Narration* entails a temporal overlap between resulting state of main eventuality of A and the preceding state of B, in the absence of temporal adverbials. In addition to temporal succession, there is also a stronger sense where A and B tell the same story, i.e., are on the same 'topic' (however, the notion of what it means to be on the same topic isn't formalized). They go on to argue that the distinction between the weaker and stronger senses is found in the contrast between French 'puis' and 'un peu plus tard'.

This brief synopsis of some of the previous theoretical literature should be qualified with the observation that the theoretical accounts provided are fragmentary, and are not directly concerned with providing a complete specification of the interaction of tense, aspect, temporal adverbials, discourse relations, and world knowledge in inferring temporal ordering. This would be regarded as too ambitious and outside the scope of a theory of semantics or pragmatics. Previous computational approaches such as (Allen 1995) and (Hitzeman et al. 1995) have attempted to integrate these different influences based on an ad-hoc algorithm for inferring the temporal structure of discourse. While those approaches are promising, they are not derived from empirical information about the presence of these various features in a corpus, and nor are they evaluated.

2.2 Our Approach

The latter point forms the point of departure for our approach, which aims at learning rules to determine the influence of these various knowledge sources to order events. The idea here is to not prespecify ad-hoc orderings or combinations of knowledge sources; rather, let information from a corpus decide this.

Our approach adopts the view of tense as anaphoric, and thus keeps track of reference times. However, in order to be domain-independent, more shallow coverage is provided in comparison with the theoretical accounts above. Aspectual analysis based on Table 0 is carried out; however, compositional analysis of aspect isn't carried out, as this requires complete parses and substantial semantic lexicons. Discourse relations aren't modeled, as machine-derived ones tend to disagree with human ones (Marcu 2001). World knowledge isn't captured, as this requires domain-specific knowledge bases. Also, a full Reichenbachian tense analysis isn't used, though this could certainly be layered on. While these lacunae possibly impact the work, there are other positive aspects that we would like to emphasize. First and foremost, we empirically assess how well humans can order events; to the best of our knowledge, this has never been addressed. Second, the text units considered are clauses rather than the simple sentences discussed in the theoretical literature. Third, discourse-level information is introduced, at least to the extent that reference times, both explicit, as well as 'implicit', are tracked, along with tense and aspect shifts. Overall, our hypothesis is that semantic information at the word and phrase level, along with syntactic rather than semantic information at the sentence level, can be effectively used in a corpus-driven approach to temporal ordering.

3 Relevance to TimeML

TimeML is intended as a Metadata Standard for markup of events, their temporal anchoring, and how they are related to each other in news articles. TimeML 1.0 defines a TLINK tag that links events to other events and/or times. For example, given the sentence "John taught 5 minutes after the explosion", a TLINK tag relates an instance of the event of teaching to an instance of the explosion, with the relation type "AFTER", mediated by the textual signal "after". The annotation example here is:

<EVENT eid="e1" class="OCCURRENCE" tense="PAST" aspect="NONE">
taught
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1"/>
<TIMEX3 tid="t1" type="DURATION" value="PT5M">
5 minutes
</TIMEX3>
<SIGNAL sid="s1">
after
</SIGNAL>
the
<EVENT eid="e2" class="OCCURRENCE" tense="NONE" aspect="NONE">
explosion
</EVENT>
<MAKEINSTANCE eiid="ei2" eventID="e2"/>
<**TLINK** eventInstanceID="ei1" signalID="s1" relatedToEvent="ei2" relType="AFTER" magnitude="t1"/>.

Likewise, given the sentence "John taught in 1992", a TLINK tag will link the event instance of teaching to the time expression 1992, with the relation "IS_INCLUDED":

John
<EVENT eid="e1" class="OCCURRENCE" tense="PAST" aspect="NONE">
taught
</EVENT>
<MAKEINSTANCE eiid="ei1" eventID="e1"/>
<SIGNAL sid="s1">

in
</SIGNAL>
<TIMEX3 tid="t1" type="DATE" value="1992">
1992
</TIMEX3>
<**TLINK** eventInstanceID="ei1" signalID="s1" relatedToTime="t1" relType="IS_INCLUDED "/>

In this paper, a simplified form of TLINK is used. In the case where a TLINK links two events, we use a **TLINK_E** tag, expressed as a relation *links*(R, e_i, e_j), where e_i and e_j are the events corresponding to events i and j, and R is one of several temporal ordering relation. Thus, in "John taught 5 minutes after the explosion", a TLINK_E tag links an instance of the event of teaching to an instance of the explosion, with the relation type "AFT". Note that signals aren't represented directly in the tag. Also, the temporal ordering relations used are BEF, AFT, and AT, rather than the more extended set of 8 ordering relations used in TimeML. Thus, while TimeML represents temporal ordering and inclusion; we use only a coarser-grained temporal ordering.

In the case where a TLINK links an event and a time, a **TLINK_T** tag is used, expressed as the relation *anchors*(R, t(e_i), t_j), where t(e_i) is the time of event e_I and t_j is the time of a particular time expression. Here, given the sentence "John taught in 1992", a TLINK_T tag will link the time of the teaching event to the time expression 1992, with the relation "AT". It is somewhat different from the LINK case in that the event instance is not directly represented; rather, the time of the event instance is. This more coarse-grained representation was developed to simplify the annotation task, motivated in part by the experiment described in Section 3.

The representation of time expressions in this paper uses the **TIMEX2** scheme (Ferro et al. 2001). It represents three different kinds of time values: points in time (answering the question "when?"), durations (answering "how long?"), and frequencies (answering "how often?"). Points in time are calendar dates and times-of-day, or a combination of both, e.g., *Monday 3 pm*, *Monday next week, a Friday, early Tuesday morning, the weekend*. These are all represented with values (the tag attribute VAL) in the ISO format, which allows for representation of date of the month, month of the year, day of the week, week of the year, and time of day, e.g., *<TIMEX2 VAL="2000-11-29-T16:30">4:30 p.m. yesterday afternoon</TIMEX2>*. TimeML uses **TIMEX3**, an extension of TIMEX2 which uses temporal functions for relative times, e.g., "last Thursday" would be represented by "(thursday (predecessor (week DCT)))", rather than a particular value. The work in this paper uses an automatic tagger called TempEx (Mani and Wilson 2000), which currently supports only TIMEX2.

TIMEX3 also includes the extensions to TIMEX2 described in (Ferro et al. 2002). For example, given the sentence "The war went on for the past three weeks", the temporal expression "the past three weeks" would be represented in TIMEX3 (as in TIMEX2) with a VAL of "P3W " (i.e., a period of three weeks), but also (unlike TIMEX2) with the ANCHOR_VAL of "2000-W02" if the phrase were uttered during the second week of January 2002. In addition, the relative direction of the period with respect to the anchor is represented in TIMEX3 by ANCHOR_DIR, i.e., "BEFORE". Since TempEx doesn't support these extensions, we dispense with them as well. Put another way, we are able to infer TLINK_E and TLINK_T tags without requiring this additional level of annotation.

So far, we have been discussing the 'markup' level of TLINK_T and TLINK_E tags. In the rest of this paper, we will not discuss temporal information at the markup level, speaking instead at the more abstract level of *links* and *anchors* relations. It should be clear, however, that the result of the system's processing can be captured at the markup level in terms of TLINK_T, TLINK_E, TIMEX2, and various other tags discussed below.

4 Initial Experiment

4.1 Introduction

Rather than have subjects carry out the extremely tedious task of annotating the temporal order of events for entire news articles, we address the subproblem of ordering pairs of successively described events. We focus on pairs which exemplify two situations:

- **Past2Past.** Tense is maintained across the pair, and each event is described in simple past tense.
- **PastPerf2Past.** Tense shifts from past perfect to simple past.

Past2Past is the prototypical case, like (1), where the narrative convention would apply as a default. PastPerf2Past was chosen because here the clauses would be likely to span a discourse boundary, and where the decision was not likely to be trivial. In terms of a Reichenbachian tense representation (Reichenbach 1947), the reference time of the second described event could be at or near the event time of the first event, involving an *Elaboration* relation, as in (4), or the reference time could shift elsewhere, e.g., to the event time of the second event, e.g., to an explicit time as in (5).

(4) State government spokesman Roberto Alvarez said the five men were criminals involved in a robbery and *had attacked* the police. "Because it *took place* in Coyuca, there is the tendency to link this with politics, but this is merely a police matter," he told Reuters.

(5) But Chang and other Taiwan spokesmen pointedly refused to confirm local media reports that Lien was in Europe, much less to confirm that he *had flown* to France. Since a civil war *divided* them in 1949…

4.2 Experimental Design

The text units we focused on for determining event order were clauses rather than sentences, because in news texts, multi-clause sentences are common. Each subject was presented, on a web page, with a capsule of 3-4 sentences. The capsule contained the clauses in question, along with an additional sentence before and after as context. In each of the clauses, a Verb Group (VG), i.e., a verb preceded by modals and auxiliaries, and followed by a particle, was highlighted, each in a different color (shown italicized in examples in this paper)[2].

The experiment was conducted with 8 subjects, all graduate students in computational linguistics, who were otherwise naïve about the goals of the experiment. The experiment involved giving a subject a pair of clauses exemplifying the two tense sequences, and asking her to judge the order of events described in the clauses. To this end, 140 pairs of adjacent past tense clauses were selected at random without replacement from the North American News Corpus (class Past2Past above). Another 140 pairs were selected at random without replacement from this corpus where the first clause had past perfect tense and the second clause had past tense (class PastPerf2Past). From the two sets of pairs, 40 examples containing a roughly equal number of the two tense sequences were chosen at random without replacement for training the subjects, and another 40 examples (again including both tense sequences) were chosen at random without replacement for giving to three subjects for an inter-annotator study. The remaining 200 examples (including both sequences) were given to the remaining 5 subjects (40 distinct examples to each subject).

Each subject was asked to make a judgment of a relation between the first VG and the second, by selecting from one of six radio buttons: *Entirely Before*, *Entirely After*, *Upto* (occurs before, and also concurrently with), *Since* (occurs concurrently with, and also after), *Equal* (exactly simultaneous), and *Unclear* (can't clearly decide between the previous five).

The first five relations, *Entirely Before*, *Entirely After*, and *Equal* map to *< (before), > (after), and = (equal),* respectively, in Allen's interval logic (Allen 1984). *Upto* and *Since* are slightly more abstract than Allen's relations (Allen 1984): *Upto* maps to *o (overlaps)* (i.e., B starts after A and continues after A) *or fi (is finished by)* (i.e., B starts after A and ends when A ends), while *Since* maps to *oi (overlapped by)* *or si (is started by)* in his logic. We decided to exclude Allen's remaining *meets* and *during* relations, as we believed that having subjects make a 7-way choice would be too burdensome -- a belief that was later confirmed by the experimental results.

[2]The focus on verb groups means that nominalized events aren't considered.

4.3 Linguistic Processing

In order to automatically generate the examples for the experiment, several components were assembled: a sentence tokenizer and part-of-speech tagger, a time expression tagger, a clause tagger, and a variety of feature extractors.

The time expression tagger TempEx (Mani and Wilson 2000) tags and assigns values to temporal expressions, both "absolute" expressions like "June 1, 2001" and relative expressions like "Monday". It was cited in (Mani and Wilson 2000) as achieving a .83 F-measure against hand-annotated data. Inter-annotator reliability across 5 annotators (graduate students) on 193 TDT2-documents was .79F for extent and .86F for time values, with TempEx scoring .76F (extent) and .82F (value) on these documents.

COTAG1 (2): Clause 1 (2) is a complement clause {1, 0}
QUOTE: Presence of a quotation mark in either clause {1, 0}
RCTAG1 (2): Clause 1 (2) is a relative clause {1, 0}
STAG: Presence of a sentence boundary between clauses {1, 0}
STATIVE1 (2): Presence of a lexical stative verb in clause 1 (2) {1, 0}
TIMEPREP (2): Presence of a temporal preposition (like since, after, before, etc.) in clause 1 (2) {1, 0}
TIMECONJ: Presence of a temporal conjunction linking the two clauses {1, 0}
TIMEX1 (2): Presence of a time expression in clause 1 (2) {1, 0}
VERB1 (2): verb in clause 1 (2) {string}

Table 1: Linguistic Features computed for each clause

The clause tagger (CLAUSE-IT) identifies top-level clauses (C), top-level clauses with gapped subjects (GC), e.g., "<C>He returned the book</C> <GC>and went home</GC>", relative clauses (RC), and complement clauses (CO), which include all non-finite clauses. Here is an example of its output:

<S><C>The United States unleashed <RC>what appeared<CO>to be its fiercest daylight strike on Afghanistan on Monday but</CO></RC></C> <C>the administration faced concern from Saudi Arabia and Pakistan over the bombardment <CO>to force Taliban leaders</CO> <CO>to hand over Saudi militant Osama bin Laden</CO>. </C></S>

CLAUSE-IT uses two passes: In the first pass, specialized finite-state grammars implemented in CASS (Abney 1996) are used to identify NPs, PPs, and VPs, and links between verbs and their subjects. An initial set of clause boundaries is proposed based on the above. In the second pass, the proposed clause boundaries are confirmed or adjusted using verb subcategorization information. Here the Penn Treebank corpus is used to look up constituents to attach to a particular verb; for example, a PP can be attached to a VP containing an object NP if the verb has been followed in the PTB by a NP and a PP headed by the current preposition.

Finally, each clause is tagged automatically with the features shown in Table 2. These features test for the presence of time expressions, time adverbials, clause type, lexical aspect (stative or not), and the specific verbs used.

5 Experimental Results

5.1 Agreement on Event Ordering

The agreement between the three subjects who judged identical examples can be examined under a strict regimen, where all 3 subjects agree only if they make identical judgments; in this case, all 3 subjects agree 24/40 i.e., 60% of the time. In a more lenient measure, we discard the 7 examples which contain a *Unclear*; in this case, all 3 subjects agree 24/33 i.e., 72% of the time.

Of the 7 disagreement examples where *Unclear* was involved, one case was a part-of-speech tagging bug, one case seemed clear to us, and 5 examples involved cases where there wasn't enough context. Of the remaining cases of disagreement, there were only 4 instances (all involving class Past2Past) that involved a polar disagreement (*Entirely Before* vs. *Entirely After*), of which only one, in our view, was truly problematic, again because of lack of sufficient context to decide either way. The other disagreements involved *Entirely Before* versus *Equal*, e.g., (6), and *Entirely Before* versus *Upto*, e.g., (7). It appears that such fine-grained distinctions are hard for people to make.

(6) In an interview with Barbara Walters to be shown on ABC's "Friday night", Shapiro said he *tried* on the gloves and *realized* they would never fit Simpson's larger hands.

(7) They *had contested* the 1992 elections separately and *won* just six seats to 70 for MPRP.

If we move to less fine-grained categories by collapsing the categories *Entirely Before* and *Upto* (i.e., ignoring whether there is a gap between the two events), the agreement goes up considerably. The Kappa measure here is 0.5 under the fine-grained measure, and 0.61 under the collapsed measure. This means that provided we collapse the fine-grained categories, the subjects show enough agreement for us to trust the overall results, and even to use the data as a training set.

5.2 Inferred Event Ordering

The overall results are shown in Figure1, where we have collapsed *Entirely Before* and *Upto* into BEF, and *Entirely After* and *Since* into AFT. It is clear that the narrative convention holds in less than half the cases for Past2Past. A substantial number of events in case Past2Past are judged to be simultaneous. For PastPerf2Past, surprisingly, the percentage of times the first event is BEF the second is higher than in Past2Past. The percentage of AFT is the same for both. About 11-14% of the time, the subject couldn't make a decision, perhaps because it was ambiguous or we didn't provide enough context.

	Past2Past	**PastPerf2Past**
BEF	62	64
AFT	30	25
Equal	21	8
Unclear	18	12

Figure 1: Event Ordering for two tense sequences

In Figure 2, we show the reasons subjects gave for their decisions. Here, after each ordering decision, subjects could, if they wanted, select from one or more of five choices, or Not Applicable. It can be seen that while temporal expressions, surface order, and tense (all fairly easily computed by a program) often provide clues to the ordering, sentence meaning (hard to compute in the large) is involved at least as often as the other clues. Subjects chose more than one reason in nearly half the cases, and Surface Order was strongly correlated with the *Entirely Before* judgments.

	Past2Past	**PastPerf2Past**
Aspect	35	18
Order	14	27
Tempex	21	16

Tense	41	24
Meaning	40	41

Figure 2: Knowledge Sources Cited

5.3 A Classifier for Ordering Clauses

Using the subjects' judgments as training data, with feature vectors constructed from the features described in Table 1, we trained a clause ordering classifier using Ripper (Cohen 1995). The results for ten-fold cross-validation are shown in Figure 3 (with standard deviations shown in parentheses).

	Past2Past	**Past-Perf2Past**
MAJ (BEF)	51.1	64.8
Ripper	58.07 (±4.01)	70.38 (±3.89)

Figure 3: Accuracy of Clause Ordering Rules

Past2Past: The default (BEF) is the majority class (accuracy 51.1%). The only other rule the system learnt was the following rule for AFT, which had an average of 57.4% accuracy in a ten-fold cross-validation.

If the time-expression in the second clause has 'after', infer AFT.

PastPerf2Past: The default (BEF) is the majority class (64.8% accuracy). The system here learnt a number of rules for the other classes, with an average accuracy on ten-fold cross-validation of 70.38%. For example:

If the time-expression in the second clause has 'after', and the first verb is a reporting verb,
Or if the time expression in the second clause has 'since', or 'when',
Then infer AFT (72.73% accuracy).
If the time expression in the second clause has 'while', then infer Equal (91.91% accuracy).

It can be seen from this that the overall accuracy is not very high, perhaps due in part to the small number of examples. The temporal expression features, along with the presence of reporting verbs are the only ones used in the rules. Other features play no part in the rules.

6 More Dense Anchoring via Implicit Reference Times

Based on our initial experiment, we focus on coarse-grained temporal ordering. We note that the narrative convention isn't strong enough to handle implicit reference times. We also observe that the proportion of clauses with explicit time expressions is approximately 25%.

In addition to ordering events directly, relying for their temporal characterization on the sparse anchoring of events available from explicit reference times, it is also possible to order events indirectly, by introducing more instances of event anchoring. This idea leads us to *compute a reference time value for each clause*, either the time value of an explicit time mentioned in the clause, or the implicit reference time that is inferred from context.

To generate this **tval** (reference time) feature, the simple algorithm in Figure 4 was used. The system also anchors the event's time with respect to the tval (at, before, or after) when the tval is an explicit reference time. This feature is called **anchor-explicit.**

:

```
history_list := {doc_date}
for each clause c do
```

```
rtime = timex2(c)
if rtime then
        tval(c) = rtime
        unless type(c, rel_clause)
            push(rtime, history_list)
elsif reporting_verb(c)  then
        tval(c) = doc_date
elsif ∃j s.t. inside_quote(c, j) then
        tval(c) = tval(j)
else tval(c) = last (history_list)
```

Figure 4: Algorithm for Computing Reference Time (tval)

CTYPE: clause is a regular clause, complement clause, or relative clause {C, CO, RC}
CINDEX: clause relative index in main-clause {integer}
PARA: paragraph number {integer}
SENT: sentence number {integer}
SCONJ: subordinating conjunction (e.g., while, since, before) {symbol}
TPREP: preposition in a TIMEX2 PP {symbol}
TIMEX2: the extent of the TIMEX2 tag {string}
TMOD: temporal modifier not attached to a TIMEX2, (e.g., after [an altercation]) {symbol}
QUOTE: number of words in quotes {integer}
REPVERBC: reporting verb-p {boolean}
STATIVEC: stative verb-p {boolean}
ACCOMPC: accomplishment verb-p {boolean}
ASPECTSHIFT: shift in aspect from previous clause {symbol}
G-ASPECT: grammatical aspect {progessive, perfect,nil}
TENSE: tense of clause {past, present, future, nil}
TENSESHIFT: shift in tense from previous clause {symbol}
ANCHOR_EXPLICIT: {<, >, =, undef}

TVAL: reference time for clause, i.e., a time value {symbol}

Table 3: Linguistic Features for Anchors

A human unconnected with our project corrected the tval, based on a set of annotation guidelines, on a sample of 2069 clauses extracted at random from the North American News Corpus. She also anchored the event's time with respect to the tval (at, bef, aft, or undefined). This feature (not a machine feature) is called **anchors**.

The corrections showed that the algorithm in Figure 4 was right on tval for 1231 out of 2069, giving an accuracy of 59%. Tracking the sequence of corrected tvals revealed that the tval of the previous clause was kept 65.75% of the time, that it reverted to some other previous tval 22.99% of the time, and that it shifted to a new tval 11.26% of the times. Most of the errors in computed tvals had to do with the tval being assigned erroneously the document date rather than reverting to a non-immediately previous tval. Finally, the **anchor-explicit** relation is correct 83.8% of the time; however, just guessing "at" for the explicit anchor will get an accuracy of 90.2%.

We then used this training data to train a statistical classifier, C5.0 Rules (Quinlan 1997), to learn (1) **anchors** relation rules and (2) rules for tracking the **tval moves** (keep, revert, shift) across successive clauses. In order to do this, new linguistic features shown in Table 3 were computed for each clause[3]. Since the TIMEX2 and tval values form an open class, they were automatically grouped into classes based on the granularity of the time expression, namely, {time-of-day, day, week, month, year, or non-specific}.

The accuracy of anchors rules as well as tval change rules are shown in Figure 5. It can be seen that accuracy of machine learning here is significantly better than the majority class. The tval, tense, and tense shift play a useful role in anchoring, revealing that the tval is a useful abstraction. Here are some of the rules learnt (here t_e is the clause index, assumed to stand for the event time of the clause):

If no sconj and no tmod and no tprep and tval-class =day then anchors(AT, t_e,, tval) 80.4% accurate (156 examples).

If tense is present and no sconj and tval-class=month then anchors(AT, t_e,, tval) 77.8 (7).

If tense is present perfect and no sconj, then anchors(BEF, t_e,, tval) 83 (4).

If tense shift is present2past and no explicit time and no sconj, then anchors(AT, t_e,, tval) 90 (30)

	ANCHORS	**TVAL-MOVES**
MAJORITY	(AT) 76.9	(KEEP) 65.75
C5.0 Rules	80.2 (±1.8)	71.8 (±0.5)

Figure 5: Accuracy of Anchoring Rules

7 Partially Ordering Links

#C	**#W**	**#correct-anchor / #total-**	**Link Recall**	**Link Preci-sion**

[3] The statives and accomplishments were computed from Maryland's LCS lexicon, based on (Dorr and Olsen 1997) See www.umiacs.umd.edu/ ~bonnie/ LCS_ Database_Documentation.html.

		anchor		
40	525	15/18 (83.3%)	44/65 (67.7%)	53/63 (84.1%)
18	335	12/13 (92.3%)	59/59 (100%)	59/62 (95.2%)
27	509	17/22 (77.2%)	23/40 (57.5%)	23/58 (39.7%)
38	617	21/27 (77.8%)	94/172 (54.7%)	94/190 (49.5%)
22	296	11/12 (91.7%)	39/42 (92.9%)	39/49 (79.6%)
14	242	6/7 (85.7%)	6/6 (100%)	6/7 (85.7%)
35	447	28/31 (90.3%)	297/339 (87.6%)	289/335 (86.3%)
194	2971	110/130 (84.6%)	562/723 (77.7%)	563/764 (73.7%)

Table 4: Document-Level Accuracy of Learnt Rules

Based on the best machine-learned rules for the **anchors** relation, **anchors** tuples are generated for each document. The tvals in the document's anchor tuples are also partially ordered, yielding tuples consisting of ordered pairs of tvals. The two sets of tuples are then used to provide a partial ordering of events in the document, in the form of **links** tuples: *links*(R, e_i, e_j), where e_i and e_j are the events corresponding to clauses i and j, and R is in {at, bef, aft, or undefined}. One of the authors evaluated the partial ordering for accuracy, on seven documents. Note that the naïve algorithm for tval is only 59% correct. While improvements to the naïve algorithm are clearly possible based on the corrected tval, to adequately test the machine learnt rules we use the corrected tval. The results of this evaluation are shown in Table 4. #C is the number of clauses, #W the number of words. #Correct-anchor is the number of the anchors tuples correctly classified and #total is the total number of anchors tuples classified. Link Recall is the percentage of human generated links tuples (723 in all) that are correctly identified by machine learned rules. Link Precision is the percentage of the machine generated links tuples that are correct. Overall, our approach achieves 75.4% F-measure accuracy in partially ordering events.

8 Related Work

In this section, we discuss the most closely related work. (Barzilay et al. 2002) describe methods for deciding on the order in which to present sentences for multi-document summarization, without deciding when events occur. (Mani and Wilson 2000) used a baseline method of blindly propagating TempEx time values to events based on proximity. On a small sample of 8,505 words of text, they obtained 394 correct event times in a sample of 663 verb occurrences, giving an accuracy of 59.4%.

(Filatova and Hovy 2001) obtained 82% accuracy on 'timestamping' clauses for a single type of event/topic on a data set of 172 clauses. While fundamental differences between the three evaluation methods preclude a comparison, it should be noted that we achieve 84.6% accuracy in temporal anchoring (Table 4). Finally, (Pustejovsky et al. 2002) report 50% recall of time-event TLINKs in an early version of their system; by contrast, we achieve 84.6% accuracy on TLINK_T tags. However, the fact that we are using 'perfect' tvals in these latter results precludes a precise comparison with these other methods.

Our approach is also distinct in its use of human experimentation, machine learning and the variety of linguistically motivated features (including temporal adverbials) that are brought to bear. The availability of a suitably large TimeML-annotated corpus will make such comparisons much easier in the future.

9 Conclusion

We have described a robust, domain-independent, corpus-derived approach to temporally anchoring and ordering events in news. This approach is highly relevant to TimeML, as it allows the automatic generation of TLINK_E and TLINK-T tags, which in turn provide a basis for TLINK tagging in TimeML. Our research has identified the capabilities of humans in ordering events in news, as well evaluated corpus-based methods for event ordering. While time adverbials, tense, tense shifts, and implicit reference times played an important role, aspectual features were not of much use in the learnt rules, perhaps due to the system's ignorance of aspectual ambiguity and aspectual composition. The role of aspect, as well as the representation of discourse relations will be examined in future work.

A more serious problem arises from the skewed distribution dominated by AT in both anchor and anchor-explicit features. It appears that in news, an overwhelming majority of events occur at the reference times mentioned. Future work on learning will investigate other learning algorithms more suited to this skewed distribution, as well as learning from sequences of vectors, rather than simply using contextual features in individual vectors.

Finally, we explored two different approaches to machine learning of ordering information: learning from event-event ordering judgments, and learning from event-time anchoring judgments. In both these cases, judgments are expensive, even within a mixed-initiative annotation framework. Future research will include more use of unsupervised learning methods.

Acknowledgments

We would like to thank Jiangping Zhang (MITRE) for help with some of the machine learning experiments and the evaluation in Table 4, and Gemma Bel-Enguix (a visiting post-doc at Georgetown University) for the annotations used in the training data for Figure 5. In addition, we would like to thank a dozen students in Computational Linguistics at Georgetown for TIMEX2 annotation and for acting as experimental subjects in the experiment reported in Section 4.2. We are indebted to Janet Hitzeman (MITRE) for comments on an earlier draft. Finally, we are grateful to George Wilson (MITRE) for use of TempEx.

References

S. Abney. Partial Parsing via Finite-State Cascades. Proceedings of the ESSLLI '96 Robust Parsing Workshop. 1996.

J. F. Allen. Towards a General Theory of Action and Time. 1984. Artificial Intelligence, 23, 123-154.

J. F. Allen. Natural Language Understanding. Chapter 16.5: Discourse Structure, Tense, and Aspect, Addison-Wesley 1995, 517-533.

R. Barzilay, N. Elhadad, and K. R. McKeown. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. JAIR, 17, 35-55.

A. Bell. News Stories as Narratives. In A. Jaworski and N. Coupland, The Discourse Reader, Routledge, 1999, 236-251.

M. Bras, A. Le Draoulec, and L. Vieu. Temporal Information and Discourse Relations in Narratives : The Role of French Connectives « Puis », « Un peu plus tard ». Workshop on Temporal and Spatial Information Processing, ACL'2001, Toulouse, 49-56.

W. W. Cohen. Fast Effective Rule Induction. Proceedings of the 12th International Conference on Machine Learning, 1995.

D. Dowty. The Effects of Aspectual Class on the Temporal Structure of Discourse: Semantics or Pragmantics? Linguistics and Philosophy, 9, 1986, 37-61.

B. Dorr and M. B. Olsen. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. ACL'1997, 151-158.

L. Ferro, I. Mani, B. Sundheim and G. Wilson "TIDES Temporal Annotation Guidelines Draft - Version 1.02". MITRE Technical Report MTR MTR 01W000004. McLean, Virginia: The MITRE Corporation, 2001.

E. Filatova, and E. Hovy. Assigning Time-Stamps to Event-Clauses. Workshop on Temporal and Spatial Information Processing, ACL'2001, Toulouse, 88-95.

L. Ferro, R. Kozierok, L. Gerber, B. Sundheim, I. Mani, and G. Wilson. Annotation Temporal Information: From Theory to Practice. HLT'2002 (poster).

J. Hitzeman, M. Moens, and C. Grover. Algorithms for Analyzing the Temporal Structure of Discourse. In Proceedings of the European ACL, Utrecht, Netherlands, 1995, 253-260.

C.H. Hwang and L. K. Schubert. Tense Trees as the fine structure of discourse. Proceedings of the 30th Annual Meeting of the ACL, 1992, 232-240.

H. Kamp and U. Reyle. From Discourse to Logic, Part 2. Chapter 5. Tense and Aspect. 483-546. Kluwer, 1993.

A. Kehler. Resolving Temporal Relations Using Tense Meaning and Discourse Interpretation. In M. Faller, S. Kaufmann, and F. Pauly, eds., Formalizing the Dynamics of Information, CSLI Publications, 2000, 1-20.

A. Lascarides and N. Asher. Temporal Relations, Discourse Structure, and Commonsense Entailment. 1993. Linguistics and Philosophy 16, 437-494.

I. Mani and G. Wilson. Robust Temporal Processing of News. ACL'2000, 69-76.

D. Marcu. The Theory and Practice of Discourse Parsing and Summarization. MIT Press, 2001.

M. Moens and M. Steedman. Temporal Ontology and Temporal Reference. Computational Linguistics, 14, 2, 1988, 15-28.

R. J. Passonneau. A Computational Model of the Semantics of Tense and Aspect. Computational Linguistics, 14, 2, 1988, 44-60.

J. Pustejovsky, J. Wiebe, and M. Maybury. Multi-perspective and temporal question answering. Third International Conference on Language Resources and Evaluation (LREC 2002), Workshop on Question Answering: Strategy and Resources, Canary Islands, Spain.

J. Pustejovsky et al. TERQAS: Time and Event Recognition for Question-Answering Systems. Final Presentation, ARDA/NRRC Workshop, July 2002.

R. Quinlan. 1997. C5.0. www.rulequest.com.

H. Reichenbach. The tenses of verbs. In H. Reichenbach, Elements of Symbolic Logic. The Macmillan Company, New York, 1947, Section 51, 287-298.

A. Setzer and R. Gaizauskas. A Pilot Study on Annotating Temporal Relations in Text. Workshop on Temporal and Spatial Information Processing, ACL'2001, Toulouse, 73-80.

Z. Vendler. Linguistics in Philosophy, Cornell University Press, Ithaca, 1967.

E. Voorhees. Overview of the TREC-2001 Question Answering Track. NIST Special Publication 500-250: TREC 2001.

B. Webber. Tense as Discourse Anaphor. Computational Linguistics, 14, 2, 1988, 61-73.

J. M. Wiebe, T. P. O'Hara, T. Ohrstrom-Sandgren, K. J. McKeever. An Empirical Approach to Temporal Reference Resolution. JAIR, 9, 1998, 247-293.