Journal of Artificial Intelligence Research 27 (2006) 85–117 Submitted 01/06; published 09/06

Learning Sentence-internal Temporal Relations

Mirella Lapata MLAP@INF.ED.AC.UK
Alex Lascarides ALEX@INF.ED.AC.UK
School of Informatics,
University of Edinburgh,
2 Buccleuch Place,
Edinburgh, EH8 9LW,
Scotland, UK

Abstract

In this paper we propose a data intensive approach for inferring sentence-internal temporal relations. Temporal inference is relevant for practical NLP applications which either extract or synthesize temporal information (e.g., summarisation, question answering). Our method bypasses the need for manual coding by exploiting the presence of markers like *after*, which overtly signal a temporal relation. We first show that models trained on main and subordinate clauses connected with a temporal marker achieve good performance on a pseudo-disambiguation task simulating temporal inference (during testing the temporal marker is treated as unseen and the models must select the right marker from a set of possible candidates). Secondly, we assess whether the proposed approach holds promise for the semi-automatic creation of temporal annotations. Specifically, we use a model trained on noisy and approximate data (i.e., main and subordinate clauses) to predict intra-sentential relations present in TimeBank, a corpus annotated rich temporal information. Our experiments compare and contrast several probabilistic models differing in their feature space, linguistic assumptions and data requirements. We evaluate performance against gold standard corpora and also against human subjects.

1. Introduction

The computational treatment of temporal information has recently attracted much attention, in part because of its increasing importance for potential applications. In multidocument summarization, for example, information that is to be included in the summary must be extracted from various documents and synthesized into a meaningful text. Knowledge about the temporal order of events is important for determining what content should be communicated and for correctly merging and presenting information in the summary. Indeed, ignoring temporal relations in either the information extraction or the summary generation phase may result in a summary which is misleading with respect to the temporal information in the original documents. In question answering, one often seeks information about the temporal properties of events (e.g., *When did X resign?*) or how events relate to each other (e.g., *Did X resign before Y?*).

An important first step towards the automatic handling of temporal phenomena is the analysis and identification of time expressions. Such expressions include absolute date or time specifications (e.g., *October 19th, 2000*), descriptions of intervals (e.g., *thirty years*), indexical expressions (e.g., *last week*), etc. It is therefore not surprising that much previous work has focused on the recognition, interpretation, and normalization of time expressions[1] (Wilson, Mani, Sundheim, & Ferro, 2001; Schilder & Habel, 2001; Wiebe, O'Hara, Öhrström Sandgren, & McKeever, 1998). Reasoning with time, however, goes beyond temporal expressions; it also involves drawing inferences about the temporal relations among events and other temporal elements in discourse. An additional challenge to this task stems from the nature of temporal information itself, which is often implicit (i.e., not overtly verbalized) and must be inferred using both linguistic and non-linguistic knowledge.

Consider the examples in (1) taken from Katz and Arosio (2001). Native speakers can infer that John first met and then kissed the girl; that he left the party after kissing the girl and then walked home; and that the events of talking to her and asking her for her name temporally overlap (and occurred before he left the party).

(1) a. John kissed the girl he met at a party.

b. Leaving the party, John walked home.

c. He remembered talking to her and asking her for her name.

The temporal relations just described are part of the interpretation of this text, even though there are no overt markers, such as *after* or *while*, signaling them. They are inferable from a variety of cues, including the order of the clauses, their compositional semantics (e.g., information about tense and aspect), lexical semantics and world knowledge. In this paper we describe a data intensive approach that automatically captures information pertaining to the temporal relations among events like the ones illustrated in (1).

A standard approach to this task would be to acquire a model of temporal relations from a corpus annotated with temporal information. Although efforts are underway to develop treebanks marked with temporal relations (Katz & Arosio, 2001) and devise annotation schemes that are suitable for coding temporal relations (Saurí, Littman, Gaizauskas, Setzer, & Pustejovsky, 2004; Ferro, Mani, Sundheim, & Wilson, 2000; Setzer & Gaizauskas, 2001), the existing corpora are too small in size to be amenable to supervised machine learning techniques which normally require thousands of training examples. The TimeBank[2] corpus, for example, contains a set of 186 news report documents annotated with the TimeML mark-up language for temporal events and expressions (for details, see Sections 2 and 7). The corpus consists of 68.5K words in total. Contrast this with the Penn Treebank, a corpus which is often used in many NLP tasks and contains approximately 1M words (i.e., it is 16 times larger than TimeBank). The annotation of temporal information is not only time-consuming but also error prone. In particular, if there are n kinds of temporal relations, then the number of possible relations to annotate is a polynomial of factor n on the number of events in the text. Pustejovsky, Mani, Belanger, Boguraev, Knippen, Litman, Rumshisky, See, Symonen, van Guilder, van Guilder, and Verhagen (2003) found evidence that this annotation task is sufficiently complex that human annotators can realistically identify only a small number of the temporal relations in text, thus compromising recall.

In default of large volumes of data labeled with temporal information, we turn to unannotated texts which nevertheless contain expressions that overtly convey the information we want our models to learn. Although temporal relations are often underspecified, sometimes there are temporal markers, such as *before*, *after*, and *while*, which make relations among events explicit:

1. See also the Time Expression Recognition and Normalisation (TERN) evaluation exercise (`http://timex2.mitre.org/tern.html`).
2. Available from `http://www.cs.brandeis.edu/~jamesp/arda/time/timebank.html`

(2) a. Leonard Shane, 65 years old, held the post of president before William Shane, 37, was elected to it last year.

b. The results were announced after the market closed.

c. Investors in most markets sat out while awaiting the U.S. trade figures.

It is precisely this type of data that we will exploit for making predictions about the temporal relationships among events in text. We will assess the feasibility of such an approach by initially focusing on sentence-internal temporal relations. From a large corpus, we will obtain sentences like the ones shown in (2), where a main clause is connected to a subordinate clause with a temporal marker, and we will develop a probabilistic framework where the temporal relations will be inferred by gathering informative features from the two clauses. Our models will view the marker from each sentence in the training corpus as the label to be learned. In the test corpus the marker will be removed and the models' task will be to pick the most likely label—or equivalently marker.

We will also examine whether models trained on data containing main and subordinate clauses together with their temporal connectives can be used to infer relations among events when temporal information is underspecified and overt temporal markers are absent (as in each of the three sentences in (1)). For this purpose, we will resort to the TimeBank corpus. The latter contains detailed annotations of events and their temporal relations irrespectively of whether connectives are present or not. Using the TimeBank annotations solely as test data, we will assess whether the approach put forward generalizes to different structures and corpora. Our evaluation study will also highlight whether a model learned from unannotated examples could alleviate the data acquisition bottleneck involved in the creation of temporal annotations. For example, by automatically creating a high volume of annotations which could be subsequently corrected manually.

In attempting to infer temporal relations probabilistically, we consider several classes of models with varying degrees of faithfulness to linguistic theory. Our models differ along two dimensions: the employed feature space and the underlying independence assumptions. We compare and contrast models which utilize word-co-occurrences with models which exploit linguistically motivated features (such as verb classes, argument relations, and so on). Linguistic features typically allow our models to form generalizations over *classes* of words, thereby requiring less training data than word co-occurrence models. We also compare and contrast two kinds of models: one assumes that the properties of the two clauses are mutually independent; the other makes slightly more realistic assumptions about dependence. (Details of the models and features used are given in Sections 3 and 4). We furthermore explore the benefits of ensemble learning methods for the temporal interpretation task and show that improved performance can be achieved when different learners (modeling sufficiently distinct knowledge sources) are combined. Our machine learning experiments are complemented by a study in which we investigate human performance on the interpretation task thereby assessing its feasibility and providing a ceiling on model performance.

The next section gives an overview of previous work in the area of computing temporal information and discusses related work which utilizes overt markers as a means for avoiding manual labeling of training data. Section 3 describes our probabilistic models and Section 4 discusses our features and the motivation behind their selection. Our experiments are presented in Sections 5–7. Section 8 offers some discussion and concluding remarks.

2. Related Work

Traditionally, methods for inferring temporal relations among events in discourse have utilized a semantics and inference-based approach. This involves complex reasoning over a variety of rich information sources, including elaborate domain knowledge and detailed logical form representations (e.g., Dowty, 1986; Hwang & Schubert, 1992; Hobbs et al., 1993; Lascarides & Asher, 1993; Kamp & Reyle, 1993; Kehler, 2002). This approach, while theoretically elegant, is impractical except for applications in very narrow domains. This is for (at least) two reasons. First, grammars that produce detailed semantic representations inevitably lack linguistic coverage and are brittle in the face of natural data; similarly, the representations of domain knowledge can lack coverage. Secondly, the complex reasoning required with these rich information sources typically involves nonmonotonic inferences (e.g., Hobbs et al., 1993; Lascarides & Asher, 1993), which become intractable except for toy examples.

Allen (1995), Hitzeman, Moens, and Grover (1995), and Han and Lavie (2004) propose more computationally tractable approaches to infer temporal information from text, by hand-crafting algorithms which integrate shallow versions of the knowledge sources that are exploited in the above theoretical literature (e.g., Hobbs et al., 1993; Kamp & Reyle, 1993). While this type of symbolic approach is promising, and overcomes some of the impracticalities of utilizing full logical forms and complex reasoning over rich domain knowledge sources, it is not grounded in empirical evidence of the way the various linguistic features contribute to the temporal semantics of a discourse; nor are these algorithms evaluated against real data. Moreover, the approach is typically domain-dependent and robustness is compromised when porting to new domains or applications.

Acquiring a model of temporal relations via machine learning over a training corpus promises to provide systems which are precise, robust, and grounded in empirical evidence. A number of markup languages have recently emerged that can greatly facilitate annotation efforts in creating suitable corpora. A notable example is TimeML (Pustejovsky, Ingria, Sauri, Castano, Littman, Gaizauskas, & Setzer, 2004; see also the annotation scheme by Katz & Arosio, 2001), a metadata standard for expressing information about the temporal properties of events and temporal relations between them. The scheme can be used to annotate a variety of temporal expressions, including tensed verbs, adjectives and nominals that correspond to times, events or states. The type of temporal information that can be expressed on these various linguistic expressions includes the class of event, its tense, grammatical aspect, polarity (positive or negative), the time denoted (e.g., one can annotate *yesterday* as denoting the day before the document date), and temporal relations between pairs of eventualities and between events and times. TimeML's expressive capabilities are illustrated in the TimeBank corpus which contains temporal annotations of news report documents (for details, see Section 7).

Mani, Schiffman, and Zhang (2003) and Mani and Schiffman (2005) demonstrate that TimeML-compliant annotations are useful for learning a model of temporal relations in news text. They focus on the problem of ordering pairs of successively described events. A decision tree classifier is trained on a corpus of temporal relations provided by human subjects. Using features such as the position of the sentence within the paragraph (and the position of the paragraph in the text), discourse connectives, temporal prepositions and other temporal modifiers, tense features, aspect shifts and tense shifts, their best model achieves 75.4% accuracy in identifying the temporal order of events. Boguraev and Ando (2005) use semi-supervised learning for recognizing events and inferring temporal relations (between an event and a time expression). Their method exploits TimeML annotations from the TimeBank corpus and large amounts of unannotated data. They first build a classifier from the TimeML annotations using a variety of features based on syntactic analysis and the identification of temporal expressions. The original feature vectors are next augmented with unlabeled data sharing structural similarities with the training data. Their algorithm yields performances well above the baseline for both tasks.

Conceivably, existing corpus data annotated with *discourse structure*, such as the RST treebank (Carlson et al., 2001), might be reused to train a temporal relations classifier. For instance, for text spans connected with RESULT, it is implied by the semantics of this relation that the events in the first span temporally precede the second; thus, a classifier of rhetorical relations could indirectly contribute to a classifier of temporal relations. Corpus-based methods for computing discourse structure are beginning to emerge (e.g., Marcu, 1999; Soricut & Marcu, 2003; Baldridge & Lascarides, 2005). But there is currently no automatic mapping from discourse structures to their temporal consequences; so although there is potential for eventually using linguistic resources labeled with discourse structure to acquire a model of temporal relations, that potential cannot be presently realized.

Continuing on the topic of discourse relations, it is worth mentioning Marcu and Echihabi (2002) whose approach bypasses altogether the need for manual coding in a supervised learning setting. A key insight in their work is that rhetorical relations (e.g., EXPLANATION and CONTRAST) are sometimes signaled by a discourse connective (e.g., *because* for EXPLANATION and *but* for CONTRAST). They extract sentences containing such markers from a corpus, and then (automatically) identify the text spans connected by the marker, remove the marker and replace it with the rhetorical relation it signals. A Naive Bayes classifier is trained on this automatically labeled data. The model is designed to be maximally simple and employs solely word bigrams as features. Specifically, bigrams are constructed over the cartesian product of words occurring in the two text spans and it is assumed that word pairs are conditionally independent. Marcu and Echihabi demonstrate that such a knowledge-lean approach performs well, achieving an accuracy of 49.70% when distinguishing six relations (over a baseline of 16.67%). However, since the model relies exlusively on word-co-occurrences, an extremely large training corpus (in the order of 40 M sentences) is required to avoid sparse data (see Sporleder & Lascarides, 2005 for more detailed discussion on the tradeoff between training size and feature space for discourse-based models).

In a sense, when considering the complexity of various models used to infer temporal and discourse relations, Marcu and Echihabi's (2002) model lies at the simple extreme of the spectrum, whereas the semantics and inference-based approaches to discourse interpretation (e.g., Hobbs et al., 1993; Asher & Lascarides, 2003) lie at the other extreme, for these latter theories assume *no independence* among the properties of the spans, and they exploit linguistic and non-linguistic features to the full. In this paper, we aim to explore a number of probabilistic models which lie in between these two extremes, thereby giving us the opportunity to study the tradeoff between the complexity of the model on the one hand, and the amount of training data required on the other. We are particularly interested in assessing the performance of models on smaller training sets than those used by Marcu and Echihabi (2002); such models will be useful for classifiers that are trained on data sets where relatively rare temporal markers are exploited.

Our work differs from Mani et al. (2003) and Boguraev and Ando (2005) in that we do not exploit manual annotations in any way. Our aim is however similar, since we also infer temporal relations between pairs of events. We share with Marcu and Echihabi (2002) the use of data with overt markers as a proxy for hand coded relations. Apart from the fact that our interpretation task is different from theirs, our work departs from Marcu and Echihabi (2002) in three further important ways. First, we propose alternative models and explore the contribution of linguistic information to the inference task, investigating how this enables one to train on considerably smaller data sets. Secondly, the proposed models are used to infer relations between events in a more realistic setting, where temporal markers are naturally absent (i.e., the test data is not simulated by removing the markers in question). And finally, we evaluate the models against human subjects performing the same task, as well as against a gold standard corpus.

3. Problem Formulation

Given a main clause and a subordinate clause attached to it, our task is to infer the temporal marker linking the two clauses. $P(S_M, t_j, S_S)$ represents the probability that a marker t_j relates a main clause S_M and a subordinate clause S_S. We aim to identify which marker t_j in the set of possible markers T maximizes the joint probability $P(S_M, t_j, S_S)$:

$$
\begin{aligned}
t^* &= \operatorname*{argmax}_{t_j \in T} P(S_M, t_j, S_S) \\
t^* &= \operatorname*{argmax}_{t_j \in T} P(S_M) P(S_S|S_M) P(t_j|S_M, S_S)
\end{aligned} \tag{3}
$$

We ignore the terms $P(S_M)$ and $P(S_S|S_M)$ in (3) as they are constant. We use Bayes' Rule to calculate $P(t_j|S_M, S_S)$:

$$
\begin{aligned}
t^* &= \operatorname*{argmax}_{t_j \in T} P(t_j|S_M, S_S) \\
t^* &= \operatorname*{argmax}_{t_j \in T} P(t_j) P(S_M, S_S|t_j) \\
t^* &= \operatorname*{argmax}_{t_j \in T} P(t_j) P(a_{\langle M,1\rangle} \cdots a_{\langle S,n\rangle}|t_j)
\end{aligned} \tag{4}
$$

S_M and S_S are vectors of features $a_{\langle M,1\rangle} \cdots a_{\langle M,n\rangle}$ and $a_{\langle S,1\rangle} \cdots a_{\langle S,n\rangle}$ characteristic of the propositions occurring with the marker t_j (our features are described in detail in Section 4.2). Estimating the different $P(a_{\langle M,1\rangle} \cdots a_{\langle S,n\rangle}|t_j)$ terms will not be feasible unless we have a very large set of training data. We will therefore make the simplifying assumption that a temporal marker t_j can be determined by observing feature *pairs* representative of a main and a subordinate clause. We further assume that these feature pairs are conditionally independent given the temporal marker and are not arbitrary: rather than considering all pairs in the cartesian product of $a_{\langle M,1\rangle} \cdots a_{\langle S,n\rangle}$ (see Marcu & Echihabi, 2002), we restrict ourselves to feature pairs that belong to the same class i. Thus, the probability of observing the conjunction $a_{\langle M,1\rangle} \cdots a_{\langle S,n\rangle}$ given t_j is:

$$
t^* = \operatorname*{argmax}_{t_j \in T} P(t_j) \prod_{i=1}^{n} \left(P(a_{\langle M,i,\rangle}, a_{\langle S,i\rangle}|t_j) \right) \tag{5}
$$

For example, if we were assuming our feature space consisted solely of nouns and verbs, we would estimate $P(a_{\langle M,i,\rangle}, a_{\langle S,i\rangle}|t_j)$ by taking into account all noun-noun and verb-verb bigrams that are attested in S_S and S_M and co-occur with t_j.

The model in (4) can be further simplified by assuming that the likelihood of the subordinate clause S_S is conditionally independent of the main clause S_M (i.e., $P(S_S, S_M|t_j) \approx P(S_S|t_j)P(S_M|t_j)$).

The assumption is clearly a simplification but makes the estimation of the probabilities $P(S_M|t_j)$ and $P(S_S|t_j)$ more reliable in the face of sparse data:

$$t^* \approx \underset{t_j \in T}{\operatorname{argmax}} P(t_j)P(S_M|t_j)P(S_S|t_j) \tag{6}$$

S_M and S_S are again vectors of features $a_{\langle M,1\rangle} \cdots a_{\langle M,n\rangle}$ and $a_{\langle S,1\rangle} \cdots a_{\langle S,n\rangle}$ representing the clauses co-occurring with the marker t_j. Now individual features (instead of feature pairs) are assumed to be conditionally independent given the temporal marker, and therefore:

$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(t_j) \prod_{i=1}^{n} \Big(P(a_{\langle M,i\rangle}|t_j)P(a_{\langle S,i\rangle}|t_j) \Big) \tag{7}$$

Returning to our example feature space of nouns and verbs, $P(a_{\langle M,i\rangle}|t_j)$ and $P(a_{\langle S,i\rangle}|t_j)$ will be now estimated by considering how often verbs and nouns co-occur with t_j. These co-occurrences will be estimated separately for main and subordinate clauses.

Throughout this paper we will use the terms *conjunctive* for model (5) and *disjunctive* for model (7). We effectively treat the temporal interpretation problem as a disambiguation task. From a (confusion) set T of temporal markers, e.g., {*after, before, since*}, we select the one that maximizes (5) or (7) (see Section 4 for details on our confusion set and corpus). The conjunctive model explicitly captures dependencies between the main and subordinate clauses, whereas the disjunctive model is somewhat simplistic in that relationships between features across the two clauses are not represented directly. However, if two values of these features for the main and subordinate clauses co-occur frequently with a particular marker, then the conditional probability of these features on that marker will approximate the right biases.

The conjunctive model is more closely related to the kinds of symbolic rules for inferring temporal relations that are used in semantics and inference-based accounts (e.g., Hobbs et al., 1993). Many rules typically draw on the *relationships* between the verbs in both clauses, or the nouns in both clauses, and so on. Both the disjunctive and conjunctive models are different from Marcu and Echihabi's (2002) model in several respects. They utilize linguistic features rather than word bigrams. The conjunctive model's features are two-dimensional with each dimension belonging to the same feature class. The disjunctive model has the added difference that it assumes independence in the features attested in the two clauses.

4. Parameter Estimation

We can estimate the parameters for our models from a large corpus. In their simplest form, the features $a_{\langle M,i\rangle}$ and $a_{\langle S,i\rangle}$ can be the words making up main and subordinate clauses. In order to extract relevant features, we first identify clauses in a hypotactic relation, i.e., main clauses of which the subordinate clause is a constituent. In the training phase, we estimate the probabilities $P(a_{\langle M,i\rangle}|t_j)$ and $P(a_{\langle S,i\rangle}|t_j)$ for the disjunctive model by simply counting the occurrence of the features $a_{\langle M,i\rangle}$ and $a_{\langle S,i\rangle}$ with marker t_j (i.e., $f(a_{\langle M,i\rangle},t_j)$) and ($f(a_{\langle S,i\rangle},t_j)$). In essence, we assume for this model that the corpus is representative of the way various temporal markers are used in English. For the conjunctive model we estimate the co-occurrence frequencies $f(a_{\langle M,i\rangle},a_{\langle S,i\rangle},t_j)$. Features with zero counts are smoothed in both models; we adopt the m-estimate with uniform priors, with m equal to the size of the feature space (Cestnik, 1990).

```
(S1 (S (NP (DT The) (NN company))
     (VP (VBD said)
       (S (NP (NNS employees))
         (VP (MD will)
          (VP (VB lose)
           (NP (PRP their) (NNS jobs))
           (SBAR-TMP (IN after)
            (S (NP (DT the) (NN sale))
             (VP (AUX is) (VP (VBN completed)))
  ))))))))
```

Figure 1: Extraction of main and subordinate clause from parse tree

4.1 Data Extraction

In order to obtain training and testing data for the models described in the previous section, subordinate clauses (and their main clause counterparts) were extracted from the BLLIP corpus (30 M words). The latter is a Treebank-style, machine-parsed version of the Wall Street Journal (WSJ, years 1987–89) which was produced using Charniak's (2000) parser. Our study focused on the following (confusion) set of temporal markers: {*after, before, while, when, as, once, until, since*}. We initially compiled a list of all temporal markers discussed in Quirk, Greenbaum, Leech, and Svartvik (1985) and eliminated markers with frequency less than 10 per million in our corpus.

We extract main and subordinate clauses connected by temporal discourse markers, by first traversing the tree top-down until we identify the tree node bearing the subordinate clause label we are interested in and then extract the subtree it dominates. Assuming we want to extract *after* subordinate clauses, this would be the subtree dominated by SBAR-TMP in Figure 1 indicated by the arrow pointing down (see *after the sale is completed*). Having found the subordinate clause, we proceed to extract the main clause by traversing the tree upwards and identifying the S node immediately dominating the subordinate clause node (see the arrow pointing up in Figure 1, *employees will lose their jobs*). In cases where the subordinate clause is sentence initial, we first identify the SBAR-TMP node and extract the subtree dominated by it, and then traverse the tree downwards in order to extract the S-tree immediately dominating it.

For the experiments described here we focus solely on subordinate clauses immediately dominated by S, thus ignoring cases where nouns are related to clauses via a temporal marker (e.g., *John left after lunch*). Note that there can be more than one main clause that qualify as attachment sites for a subordinate clause. In Figure 1 the subordinate clause *after the sale is completed* can be attached either to *said* or *will loose*. There can be similar structural ambiguities for identifying the subordinate clause; for example see (8), where the conjunction *and* should lie within the scope of the subordinate *before*-clause (and indeed, the parser disambiguates the structural ambiguity correctly for this case):

(8) [Mr. Grambling made off with $250,000 of the bank's money [before Colonial caught on and denied him the remaining $100,000.]]

We are relying on the parser for providing relatively accurate resolutions of structural ambiguities, but unavoidably this will create some noise in the data. To estimate the extent of this noise, we manually inspected 30 randomly selected examples for each of our temporal discourse markers

TMark	Frequency	Distribution (%)
when	35,895	42.83
as	15,904	19.00
after	13,228	15.79
before	6,572	7.84
until	5,307	6.33
while	3,524	4.20
since	2,742	3.27
once	638	0.76
TOTAL	83,810	100.00

Table 1: Subordinate clauses extracted from BLLIP corpus

i.e., 240 examples in total. All the examples that we inspected were true positives of temporal discourse markers save one, where the parser assumed that *as* took a sentential complement whereas in reality it had an NP complement (i.e., *an anti-poverty worker*):

(9) [He first moved to West Virginia [as an anti-poverty worker, then decided to stay and start a political career, eventually serving two terms as governor.]]

In most cases the noise is due to the fact that the parser either overestimates or underestimates the extent of the text span for the two clauses. 98.3% of the main clauses and 99.6% of the subordinate clauses were accurately identified in our data set. Sentence (10) is an example where the parser incorrectly identifies the main clause: it predicts that the *after*-clause is attached to *to denationalise the country's water industry*. Note, however, that the subordinate clause (as *some managers resisted the move and workers threatened lawsuits*) is correctly identified.

(10) [Last July, the government postponed plans [to denationalise the country's water industry [after some managers resisted the move and workers threatened lawsuits.]]]

The size of the corpus we obtain with these extraction methods is detailed in Table 1. There are 83,810 instances overall (i.e., just 0.20% of the size of the corpus used by Marcu and Echihabi, 2002). Also note that the distribution of temporal markers ranges from 0.76% (for *once*) to 42.83% (for *when*).

Some discourse markers from our confusion set underspecify temporal semantic information. For example, *when* can entail temporal overlap (see (11a), from Kamp & Reyle, 1993), or temporal progression (see (11c), from Moens & Steedman, 1988). The same is true for *once*, *since*, and *as*:

(11) a. Mary left when Bill was preparing dinner. (*temporal overlap*)
b. When they built the bridge, they solved all their traffic problems. (*temporal progression*)

(12) a. Once John moved to London, he got a job with the council. (*temporal progression*)
b. Once John was living in London, he got a job with the council. (*temporal overlap*)

(13) a. John has worked for the council since he's been living in London. (*temporal overlap*)

b. John moved to London since he got a job with the council there. (*cause* and hence *temporal precedence*)

(14) a. Grand melodies poured out of him as he contemplated Caesar's conquest of Egypt. (*temporal overlap*)

b. I went to the bank as I ran out of cash. (*cause*, and hence *temporal precedence*)

This means that if the model chooses *when*, *once*, or *since* as the most likely marker between a main and subordinate clause, then the temporal relation between the events described is left underspecified. Of course the semantics of *when* or *once* limits the range of possible relations, but our model does not identify which specific relation is conveyed by these markers for a given example. Similarly, *while* is ambiguous between a temporal use in which it signals that the eventualities temporally overlap (see (15a)) and a contrastive use which does not convey any particular temporal relation (although such relations may be conveyed by other features in the sentence, such as tense, aspect and world knowledge; see (15b)).

(15) a. While the stock market was rising steadily, even companies stuffed with cash rushed to issue equity.

b. While on the point of history he was directly opposed to Liberal Theology, his appeal to a 'spirit' somehow detachable from the Jesus of history run very much along similar lines to the Liberal approach.

We inspected 30 randomly-selected examples for markers with underspecified readings (i.e., *when*, *once*, *since*, *while* and *as*). The marker *when* entails a temporal overlap interpretation 70% of the time and *as* entails temporal overlap 75% of the time, whereas *once* and *since* are more likely to entail temporal progression (74% and 80%, respectively). The markers *while* and *as* receive predominantly temporal interpretations in our corpus. Specifically, *while* has non-temporal uses in 13.3% of the instances in our sample and *as* in 25%. Once our interpretation model has been applied, we could use these biases to disambiguate, albeit coarsely, markers with underspecified meanings. Indeed, we demonstrate with Experiment 3 (see Section 7) that our model is useful for estimating *unambiguous* temporal relations, even when the original sentence had no temporal marker, ambiguous or otherwise.

4.2 Model Features

A number of knowledge sources are involved in inferring temporal ordering including tense, aspect, temporal adverbials, lexical semantic information, and world knowledge (Asher & Lascarides, 2003). By selecting features that represent these knowledge sources, notwithstanding indirectly and imperfectly, we aim to empirically assess their contribution to the temporal inference task. Below we introduce our features and provide motivation behind their selection.

Temporal Signature (T) It is well known that verbal tense and aspect impose constraints on the temporal order of events and also on the choice of temporal markers. These constraints are perhaps best illustrated in the system of Dorr and Gaasterland (1995) who examine how *inherent* (i.e., states and events) and *non-inherent* (i.e., progressive, perfective) aspectual features interact with the time stamps of the eventualities in order to generate clauses and the markers that relate them.

FINITE	=	{past, present}
NON-FINITE	=	{0, infinitive, ing-form, en-form}
MODALITY	=	{∅, future, ability, possibility, obligation}
ASPECT	=	{imperfective, perfective, progressive}
VOICE	=	{active, passive}
NEGATION	=	{affirmative, negative}

Table 2: Temporal signatures

Feature	$once_M$	$once_S$	$since_M$	$since_S$
FIN	0.69	0.72	0.75	0.79
PAST	0.28	0.34	0.35	0.71
ACT	0.87	0.51	0.85	0.81
MOD	0.22	0.02	0.07	0.05
NEG	0.97	0.98	0.95	0.97

Table 3: Relative frequency counts for temporal features in main (subscript M) and subordinate (subscript S) clauses

Although we cannot infer inherent aspectual features from verb surface form (for this we would need a dictionary of verbs and their aspectual classes together with a process that assigns aspectual classes in a given context), we can extract non-inherent features from our parse trees. We first identify verb complexes including modals and auxiliaries and then classify tensed and non-tensed expressions along the following dimensions: finiteness, non-finiteness, modality, aspect, voice, and polarity. The values of these features are shown in Table 2. The features finiteness and non-finiteness are mutually exclusive.

Verbal complexes were identified from the parse trees heuristically by devising a set of 30 patterns that search for sequences of auxiliaries and verbs. From the parser output verbs were classified as passive or active by building a set of 10 passive identifying patterns requiring both a passive auxiliary (some form of *be* and *get*) and a past participle.

To illustrate with an example, consider again the parse tree in Figure 1. We identify the verbal groups *will lose* and *is completed* from the main and subordinate clause respectively. The former is mapped to the features {present, 0, future, imperfective, active, affirmative}, whereas the latter is mapped to {present, 0, ∅, imperfective, passive, affirmative}, where 0 indicates the verb form is finite and ∅ indicates the absence of a modal. In Table 3 we show the relative frequencies in our corpus for finiteness (FIN), past tense (PAST), active voice (ACT), and negation (NEG) for main and subordinate clauses conjoined with the markers *once* and *since*. As can be seen there are differences in the distribution of counts between main and subordinate clauses for the same and different markers. For instance, the past tense is more frequent in *since* than *once* subordinate clauses and modal verbs are more often attested in *since* main clauses when compared with *once* main clauses. Also, *once* main clauses are more likely to be active, whereas *once* subordinate clauses can be either active or passive.

TMark	Verb_M	Verb_S	$\text{Supersense}_\text{M}$	$\text{Supersense}_\text{S}$	Levin_M	Levin_S
after	sell	leave	communication	communication	say	say
as	come	acquire	motion	motion	say	begin
before	say	announce	stative	stative	say	begin
once	become	complete	stative	stative	say	get
since	rise	expect	stative	change	say	begin
until	protect	pay	communication	possession	say	get
when	make	sell	stative	motion	characterize	get
while	wait	complete	communication	social	say	amuse

Table 4: Most frequent verbs and verb classes in main (subscript M) and subordinate clauses (subscript M)

Verb Identity (V) Investigations into the interpretation of narrative discourse have shown that specific lexical information plays an important role in determining temporal interpretation (e.g., Asher & Lascarides, 2003). For example, the fact that verbs like *push* can cause movement of their object and verbs like *fall* describe the movement of their subject can be used to interpret the discourse in (16) as the pushing causing the falling, thus making the linear order of the events mismatch their temporal order.

(16) Max fell. John pushed him.

We operationalize lexical relationships among verbs in our data by counting their occurrence in main and subordinate clauses from a lemmatized version of the BLLIP corpus. Verbs were extracted from the parse trees containing main and subordinate clauses. Consider again the tree in Figure 1. Here, we identify *lose* and *complete*, without preserving information about tense or passivisation which is explicitly represented in our temporal signatures. Table 4 lists the most frequent verbs attested in main (Verb_M) and subordinate (Verb_S) clauses conjoined with the temporal markers *after*, *as*, *before*, *once*, *since*, *until*, *when*, and *while* (TMark).

Verb Class ($\mathbf{V_W}$, $\mathbf{V_L}$) The verb identity feature does not capture meaning regularities concerning the types of verbs entering in temporal relations. For example, in Table 4 *sell* and *pay* are possession verbs, *say* and *announce* are communication verbs, and *come* and *rise* are motion verbs. Asher and Lascarides (2003) argue that many of the rules for inferring temporal relations should be specified in terms of the *semantic class* of the verbs, as opposed to the verb forms themselves, so as to maximize the linguistic generalizations captured by a model of temporal relations. For our purposes, there is an additional empirical motivation for utilizing verb classes as well as the verbs themselves: it reduces the risk of sparse data. Accordingly, we use two well-known semantic classifications for obtaining some degree of generalization over the extracted verb occurrences, namely WordNet (Fellbaum, 1998) and the verb classification proposed by Levin (1995).

Verbs in WordNet are classified in 15 broad semantic domains (e.g., verbs of change, verbs of cognition, etc.) often referred to as supersenses (Ciaramita & Johnson, 2003). We therefore mapped the verbs occurring in main and subordinate clauses to WordNet supersenses (feature V_W). Semantically ambiguous verbs will correspond to more than one semantic class. We resolve ambiguity

TMark	Noun_N	Noun_S	$\text{Supersense}_\text{M}$	$\text{Supersense}_\text{S}$	Adj_M	Adj_S
after	year	company	act	act	last	new
as	market	dollar	act	act	recent	previous
before	time	year	act	group	long	new
once	stock	place	act	act	more	new
since	company	month	act	act	first	last
until	president	year	act	act	new	next
when	act	act	year	year	last	last
while	group	act	chairman	plan	first	other

Table 5: Most frequent nouns, noun classes, and adjectives in main (subscript M) and subordinate clauses (subscript M)

heuristically by always defaulting to the verb's prime sense (as indicated in WordNet) and selecting its corresponding supersense. In cases where a verb is not listed in WordNet we default to its lemmatized form.

Levin (1995) focuses on the relation between verbs and their arguments and hypothesizes that verbs which behave similarly with respect to the expression and interpretation of their arguments share certain meaning components and can therefore be organized into semantically coherent classes (200 in total). Asher and Lascarides (2003) argue that these classes provide important information for identifying semantic relationships between clauses. Verbs in our data were mapped into their corresponding Levin classes (feature V_L); polysemous verbs were disambiguated by the method proposed by Lapata and Brew (2004).[3] Again, for verbs not included in Levin, the lemmatized verb form was used. Examples of the most frequent Levin classes in main and subordinate clauses as well as WordNet supersenses are given in Table 4.

Noun Identity (N) It is not only verbs, but also nouns that can provide important information about the semantic relation between two clauses; Asher and Lascarides (2003) discuss an example in which having the noun *meal* in one sentence and *salmon* in the other serves to trigger inferences that the events are in a part-whole relation (eating the salmon was part of the meal). An example from our corpus concerns the nouns *share* and *market*. The former is typically found in main clauses preceding the latter which is often in a subordinate clause. Table 5 shows the most frequently attested nouns (excluding proper names) in main (Noun_M) and subordinate (Noun_S) clauses for each temporal marker. Notice that time denoting nouns (e.g., *year*, *month*) are relatively frequent in this data set.

Nouns were extracted from a lemmatized version of the BLLIP corpus. In Figure 1 the nouns *employees*, *jobs* and *sales* are relevant for the Noun feature. In cases of noun compounds, only the compound head (i.e., rightmost noun) was taken into account. A small set of rules was used to identify organizations (e.g., *United Laboratories Inc.*), person names (e.g., *Jose Y. Campos*),

3. Lapata and Brew (2004) develop a simple probabilistic model which determines for a given polysemous verb and its frame its most likely meaning overall (i.e., across a corpus), without relying on the availability of a disambiguated corpus. Their model combines linguistic knowledge in the form of Levin (1995) classes and frame frequencies acquired from a parsed corpus.

and locations (e.g., *New England*) which were subsequently substituted by the general categories `person`, `organization`, and `location`.

Noun Class (N_W) As with verbs, Asher and Lascarides (2003) argue in favor of symbolic rules for inferring temporal relations that utilize the semantic classes of nouns wherever possible, so as to maximize the linguistic generalizations that are captured. For example, they argue that one can infer a causal relation in (17) on the basis that the noun *bruise* has a cause via some act-on predicate with some underspecified agent (other nouns in this class include *injury*, *sinking*, *construction*):

(17) John hit Susan. Her bruise is enormous.

Similarly, inferring that *salmon* is part of a meal in (18) rests on the fact that the noun *salmon*, in one sense at least, denotes an edible substance.

(18) John ate a wonderful meal. He devoured lots of salmon.

As in the case of verbs, nouns were also represented by supersenses from the WordNet taxonomy. Nouns in WordNet do not form a single hierarchy; instead they are partitioned according to a set of semantic primitives into 25 supersenses (e.g., nouns of cognition, events, plants, substances, etc.), which are treated as the unique beginners of separate hierarchies. The nouns extracted from the parser were mapped to WordNet classes. Ambiguity was handled in the same way as for verbs. Examples of the most frequent noun classes attested in main and subordinate clauses are illustrated in Table 5.

Adjective (A) Our motivation for including adjectives in the feature set is twofold. First, we hypothesize that temporal adjectives (e.g., *old*, *new*, *later*) will be frequent in subordinate clauses introduced by temporal markers such as *before*, *after*, and *until* and therefore may provide clues for relations signaled by these markers. Secondly, similarly to verbs and nouns, adjectives carry important lexical information that can be used for inferring the semantic relation that holds between two clauses. For example, antonyms can often provide clues about the temporal sequence of two events (see *incoming* and *outgoing* in (19)).

(19) The incoming president delivered his inaugural speech. The outgoing president resigned last week.

As with verbs and nouns, adjectives were extracted from the parser's output. The most frequent adjectives in main (Adj_M) and subordinate (Adj_S) clauses are given in Table 4.

Syntactic Signature (S) The syntactic differences in main and subordinate clauses are captured by the syntactic signature feature. The feature can be viewed as a measure of tree complexity, as it encodes for each main and subordinate clause the number of NPs, VPs, PPs, ADJPs, and ADVPs it contains. The feature can be easily read off from the parse tree. The syntactic signature for the main clause in Figure 1 is [NP:2 VP:2 ADJP:0 ADVP:0 PP:0] and for the subordinate clause [NP:1 VP:1 ADJP:0 ADVP:0 PP:0]. The most frequent syntactic signature for main clauses is [NP:2 VP:1 PP:0 ADJP:0 ADVP:0]; subordinate clauses typically contain an adverbial phrase [NP:2 VP:1 ADJP:0 ADVP:1 PP:0]. One motivating case for using this syntactic feature involves verbs describing propositional attitudes (e.g., *said*, *believe*, *realize*). Our set of temporal discourse markers will have varying distributions as to their relative semantic scope to these verbs. For example, one would expect *until* to take narrow semantic scope (i.e., the *until*-clause would typically attach to the verb in the sentential complement to the propositional attitude verb, rather than to the propositional attitude verb itself), while the situation might be different for *once*.

Argument Signature (R) This feature captures the argument structure profile of main and subordinate clauses. It applies only to verbs and encodes whether a verb has a direct or indirect object, and whether it is modified by a preposition or an adverbial. As the rules for inferring temporal relations in Hobbs et al. (1993) and Asher and Lascarides (2003) attest, the predicate argument structure of clauses is crucial to making the correct temporal inferences in many cases. To take a simple example, observe that inferring the causal relation in (16) crucially depends on the fact that the subject of *fall* denotes the same person as the direct object of *push*; without this, a relation other than a causal one would be inferred.

As with syntactic signature, this feature was read from the main and subordinate clause parse-trees. The parsed version of the BLLIP corpus contains information about subjects. NPs whose nearest ancestor was a VP were identified as objects. Modification relations were recovered from the parse trees by finding all PPs and ADVPs immediately dominated by a VP. In Figure 1 the argument signature of the main clause is [SUBJ OBJ] and for the subordinate it is [OBJ].

Position (P) This feature simply records the position of the two clauses in the parse tree, i.e., whether the subordinate clause precedes or follows the main clause. The majority of the main clauses in our data are sentence initial (80.8%). However, there are differences among individual markers. For example, *once* clauses are equally frequent in both positions. 30% of the *when* clauses are sentence initial whereas 90% of the *after* clauses are found in the second position. These statistics clearly show that the relative positions of the main vs. subordinate clauses are going to be relatively informative for the the interpretation task.

In the following sections we describe our experiments with the models introduced in Section 3. We first investigate their performance on temporal interpretation in the context of a pseudo-disambiguation task (Experiment 1). We also describe a study with humans (Experiment 2) which enables us to examine in more depth the models' behavior and the difficulty of the inference task. Finally, we evaluate the proposed approach in a more realistic setting, using sentences that do not contain explicit temporal markers (Experiment 3).

5. Experiment 1: Temporal Inference as Pseudo-disambiguation

Method Our models were trained on main and subordinate clauses extracted from the BLLIP corpus as detailed in Section 4. In the testing phase, all occurrences of the relevant temporal markers were removed and the models were used to select the marker which was originally attested in the corpus. This experimental setup is admittedly artificial, but important in revealing the difficulty of the task at hand. A model that performs deficiently at the pseudo-disambiguation task, has little hope of inferring temporal relations in a more natural setting where events are neither connected via temporal markers nor found in a main-subordinate relationship.

Recall that we obtained 83,810 main-subordinate pairs. These were randomly partitioned into training (80%), development (10%) and test data (10%). Eighty randomly selected pairs from the test data were reserved for the human study reported in Experiment 2. We performed parameter tuning on the development set; all our results are reported on the unseen test set, unless otherwise stated. We compare the performance of the conjunctive and disjunctive models, thereby assessing

Symbols	Meaning
$*$	significantly different from Majority Baseline
†	significantly different from Word-based Baseline
$	significantly different from Conjunctive Model
‡	significantly different from Disjunctive Model
&	significantly different from Conjunctive Ensemble
#	significantly different from Disjunctive Ensemble

Table 6: Meaning of diacritics indicating statistical significance (χ^2 tests, $p < 0.05$)

the effect of feature (in)dependence on the temporal interpretation task. Furthermore, we compare the performance of the two proposed models against a baseline disjunctive model that employs a word-based feature space (see (7) where $P(a_{\langle M,i\rangle} = w_{\langle M,i\rangle}|t_j))$ and $P(a_{\langle S,i\rangle} = w_{\langle S,i\rangle}|t_j))$. This model resembles Marcu and Echihabi's (2002)'s model in that it does not make use of the linguistically motivated features presented in the previous section; all that is needed for estimating its parameters is a corpus of main-subordinate clause pairs. We also report the performance of a majority baseline (i.e., always select *when*, the most frequent marker in our data set).

In order to assess the impact of our feature classes (see Section 4.2) on the interpretation task, the feature space was exhaustively evaluated on the development set. We have nine classes, which results in $\frac{9!}{(9-k)!}$ combinations where k is the arity of the combination (unary, binary, ternary, etc.). We measured the accuracy of all class combinations (1,023 in total) on the development set. From these, we selected the best performing ones for evaluating the models on the test set.

Results Our results are shown in Table 7. We report both accuracy and F-score. A set of diacritics is used to indicate significance (on accuracy) throughout this paper (see Table 6). The best performing disjunctive model on the test set (accuracy 62.6%) was observed with the combination of verbs (V) with syntactic signatures (S). The combination of verbs (V), verb classes (V_L, V_W), syntactic signatures (S) and clause position (P) yielded the highest accuracy (60.3%) for the conjunctive model. Both conjunctive and disjunctive models performed significantly better than the majority baseline and word-based model which also significantly outperformed the majority baseline. The disjunctive model (SV) significantly outperformed the conjunctive one ($\mathrm{V_W V_L PSV}$).

We attribute the conjunctive model's worse performance to data sparseness. There is clearly a trade-off between reflecting the true complexity of the task of inferring temporal relations and the amount of training data available. The size of our data set favors a simpler model over a more complex one. The difference in performance between the models relying on linguistically-motivated features and the word-based model also shows that linguistic abstractions are useful in overcoming sparse data.

We further analyzed the data requirements for our models by varying the amount of instances on which they are trained. Figure 2 shows learning curves for the best conjunctive and disjunctive models ($\mathrm{V_W V_L PSV}$ and SV). For comparison, we also examine how training data size affects the (disjunctive) word-based baseline model. As can be seen, the disjunctive model has an advantage over the conjunctive one; the difference is more pronounced with smaller amounts of training data. Very small performance gains are obtained with increased training data for the word baseline model. A considerably larger training set is required for this model to be competitive against the more lin-

Model	Accuracy	F-score
Majority Baseline	$42.6^{\dagger\$\ddagger\#\&}$	NA
Word-based Baseline	$48.2^{*\$\ddagger\#\&}$	44.7
Conjunctive ($V_W V_L$PSV)	$60.3^{*\dagger\ddagger\#\&}$	53.3
Disjunctive (SV)	$62.6^{*\dagger\$\#\&}$	62.3
Ensemble (Conjunctive)	$64.5^{*\dagger\$\ddagger\&}$	59.9
Ensemble (Disjunctive)	$70.6^{*\dagger\$\ddagger\#}$	69.1

Table 7: Summary of results for the temporal pseudo-disambiguation task; comparison of baseline models against conjunctive and disjunctive models and their ensembles (V: verbs, V_W: WordNet verb supersenses, V_L: Levin verb classes, P: clause position, S: syntactic signature)

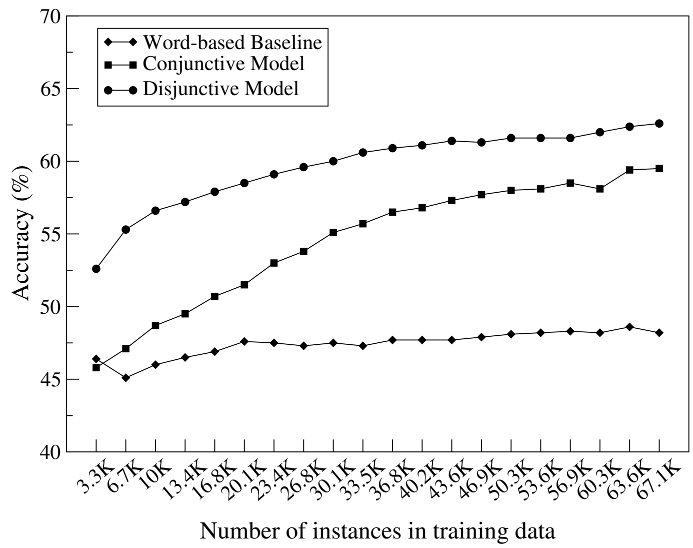

Figure 2: Learning curve for conjunctive, disjunctive, and word-based models.

guistically aware models. This result is in agreement with Marcu and Echihabi (2002) who employ a very large corpus (1 billion words, from which they extract 40 million training examples) for training their word-based model.

Further analysis of our models' output revealed that some feature combinations performed reasonably well on individual markers for both the disjunctive and conjunctive model, even though their overall accuracy did not match the best feature combinations for either model class. Some accuracies for these combinations are shown in Table 8. For example, NPRSTV was one of the best combinations for generating *after* under the disjunctive model, whereas SV was better for *before* (feature abbreviations are as introduced in Section 4.2). Given the complementarity of different models, an obvious question is whether these can be combined. An important finding in machine learning is that a set of classifiers whose individual decisions are combined in some way (an *ensemble*) can be more accurate than any of its component classifiers if the errors of the individual

	Disjunctive Model		Conjunctive Model	
TMark	Features	Accuracy	Features	Accuracy
after	NPRSTV	69.9	V_WPTV	79.6
as	ANN$_W$PSV	57.0	$V_W V_L$SV	57.0
before	SV	42.1	TV	11.3
once	PRS	40.7	V_WP	3.7
since	PRST	25.1	V_LV	1.0
when	V_LPS	85.5	V_LNV	86.5
while	PST	49.0	V_LPV	9.6
until	$V_L V_W$RT	69.4	$V_W V_L$PV	9.5

Table 8: Best feature combinations for individual markers (development set; V: verbs, V_W: WordNet verb supersenses, V_L: Levin verb classes, N: nouns, N_W: WordNet noun supersenses, P: clause position, R: argument signature, S: syntactic signature, T: tense signature)

classifiers are sufficiently uncorrelated (Dietterich, 1997). The next section reports on our ensemble learning experiments.

Ensemble Learning An ensemble of classifiers is a set of classifiers whose individual decisions are combined to classify new examples. This simple idea has been applied to a variety of classification problems ranging from optical character recognition to medical diagnosis and part-of-speech tagging (for overviews, see Dietterich, 1997; van Halteren, Zavrel, & Daelemans, 2001). Ensemble learners often yield superior results to individual learners provided that the component learners are accurate and diverse (Hansen & Salamon, 1990).

An ensemble is typically built in two steps: first multiple component learners are trained and next their predictions are combined. Multiple classifiers can be generated either by using subsamples of the training data (Breiman, 1996a; Freund & Shapire, 1996) or by manipulating the set of input features available to the component learners (Cherkauer, 1996). Weighted or unweighted voting is the method of choice for combining individual classifiers in an ensemble. A more sophisticated combination method is *stacking* where a learner is trained to predict the correct output class when given as input the outputs of the ensemble classifiers (Wolpert, 1992; Breiman, 1996b; van Halteren et al., 2001). In other words, a second-level learner is trained to select its output on the basis of the patterns of co-occurrence of the output of several component learners.

We generated multiple classifiers (for combination in the ensemble) by varying the number and type of features available to the conjunctive and disjunctive models discussed in the previous section. The outputs of these models were next combined using c5.0 (Quinlan, 1993), a decision-tree second level-learner. Decision trees are among the most widely used machine learning algorithms. They perform a general to specific search of a feature space, adding the most informative features to a tree structure as the search proceeds. The objective is to select a minimal set of features that efficiently partitions the feature space into classes of observations and assemble them into a tree (for details, see Quinlan, 1993). A classification for a test case is made by traversing the tree until either a leaf node is found or all further branches do not match the test case, and returning the most frequent class at the last node.

Conjunctive Ensemble						
APTV	$PSVV_WN_WV_L$	$NPVV_WV_L$	$PRTVV_WV_L$	PSV_WV_L	$PSVV_WV_L$	PVV_WV_L
SVV_WV_L	$NSVV_W$	$PSVV_W$ PVV_W	N_WPSVV_L	PSV_L	PVV_L	NPSV
NPV	NSV	PSV	PV	SV	TV	V
Disjunctive Ensemble						
AN_WNPSV	APSV	ASV	$PRSV_W$	PSV_N	SV_L	NPRSTV
PRS	PRST	PRSV	PSV	SV		

Table 9: Component models for ensemble learning (A: adjectives, V: verbs, V_W: WordNet verb supersenses, V_L: Levin verb classes, N: nouns, N_W: WordNet noun supersenses, P: clause position, R: argument signature, S: syntactic signature, T: tense signature)

Learning in this framework requires a primary training set for training the component learners; a secondary training set for training the second-level learner and a test set for assessing the stacked classifier. We trained the decision-tree learner on the development set using 10-fold cross-validation. We experimented with 133 different conjunctive models and 65 disjunctive models; the best results on the development set were obtained with the combination of 22 conjunctive models and 12 disjunctive models. The component models are presented in Table 9. The ensembles' performance on the test set is reported in Table 7.

As can be seen, both types of ensemble significantly outperform the word-based baseline, and the best performing individual models. Furthermore, the disjunctive ensemble significantly outperforms the conjunctive one. Table 10 details the performance of the two ensembles for each individual marker. Both ensembles have difficulty inferring the markers *since*, *once* and *while*; the difficulty is more pronounced in the conjunctive ensemble. We believe that the worse performance for predicting these relations is due to a combination of sparse data and ambiguity. First, observe that these three classes have fewest examples in our data set (see Table 1). Secondly, *once* is temporally ambiguous, conveying temporal progression and temporal overlap (see example (12)). The same ambiguity is observed with *since* (see example (13)). Finally, although the temporal sense of *while* always conveys temporal overlap, it has a non-temporal, contrastive sense too which potentially creates some noise in the training data, as discussed in Section 4.1. Another contributing factor to *while*'s poor performance is the lack of sufficient training data. Note that the extracted instances for this marker constitute only 4.2% of our data. In fact, the model often confuses the marker *since* with the semantically similar *while*. This could be explained by the fact that the majority of training examples for *since* had interpretations that imply temporal overlap, thereby matching the temporal relation implied by *while*, which in turn was also the majority interpretation in our training corpus (the non-temporal, contrastive sense accounting for only 13.3% of our training examples).

Let us now examine which classes of features have the most impact on the interpretation task by observing the component learners selected for our ensembles. As shown in Table 8, verbs either as lexical forms (V) or classes (V_W, V_L), the syntactic structure of the main and subordinate clauses (S) and their position (P) are the most important features for interpretation. Verb-based features are present in all component learners making up the conjunctive ensemble and in 10 (out of 12) learners for the disjunctive ensemble. The argument structure feature (R) seems to have some influence (it is present in five of the 12 component (disjunctive) models), however we suspect that there is some overlap with S. Nouns, adjectives and temporal signatures seem to have a small impact on

TMark	Disjunctive Ensemble Accuracy	Disjunctive Ensemble F-score	Conjunctive Ensemble Accuracy	Conjunctive Ensemble F-score
after	66.4	63.9	59.3	57.6
as	62.5	62.0	59.0	55.1
before	51.4	50.6	17.1	22.3
once	24.6	35.3	0.0	0.0
since	26.2	38.2	3.9	4.5
when	91.0	86.9	90.5	84.7
while	28.8	41.2	11.5	15.8
until	47.8	52.4	17.3	24.4
All	70.6	69.1	64.5	59.9

Table 10: Ensemble results on sentence interpretation for individual markers (test set)

the interpretation task, at least in the WSJ domain. Our results so far point to the importance of the lexicon for inferring temporal relations but also indicate that the syntactic complexity of the two clauses is an another key predictor. Asher and Lascarides' (2003) symbolic theory of discourse interpretation also emphasizes the importance of lexical information in inferring temporal relations, while Soricut and Marcu (2003) find that syntax trees are useful for inferring discourse relations, some of which have temporal consequences.

6. Experiment 2: Human Evaluation

Method We further assessed the temporal interpretation model by comparing its performance against human judges. Participants were asked to perform a multiple choice task. They were given a set of 40 main-subordinate pairs (five for each marker) randomly chosen from our test data. The marker linking the two clauses was removed and participants were asked to select the missing word from a set of eight temporal markers, thus mimicking the model's task. Examples of the materials our participants saw are given in Apendix A.

The study was conducted remotely over the Internet. Subjects first read a set of instructions that explained the task, and had to fill in a short questionnaire including basic demographic information. A random order of main-subordinate pairs and a random order of markers per pair was generated for each subject. The study was completed by 198 volunteers, all native speakers of English. Subjects were recruited via postings to local Email lists.

Results Our results are summarized in Table 11. We measured how well our subjects (Human) agree with the gold standard (Gold)—i.e., the corpus from which the experimental items were selected—and how well they agree with each other (Human-Human). We also show how well the disjunctive ensemble (Ensemble) agrees with the subjects (Ensemble-Human) and the gold standard (Ensemble-Gold). We measured agreement using the Kappa coefficient (Siegel & Castellan, 1988) but also report percentage agreement to facilitate comparison with our model. In all cases we compute pairwise agreements and report the mean.

	K	%
Human-Human	.410	45.0
Human-Gold	.421	46.9
Ensemble-Human	.390	44.3
Ensemble-Gold	.413	47.5

Table 11: Agreement figures for subjects and disjunctive ensemble (Human-Human: inter-subject agreement, Human-Gold: agreement between subjects and gold standard corpus, Ensemble-Human: agreement between ensemble and subjects, Ensemble-Gold: agreement between ensemble and gold standard corpus)

	after	as	before	once	since	until	when	while
after	**.55**	.06	.03	.10	.04	.01	.20	.01
as	.14	**.33**	.02	.02	.03	.03	.20	.23
before	.05	.05	**.52**	.08	.03	.15	.08	.04
once	.17	.06	.10	**.35**	.07	.03	.17	.05
since	.10	.09	.04	.04	**.63**	.03	.06	.01
until	.06	.03	.05	.10	.03	**.65**	.05	.03
when	.20	.07	.09	.09	.04	.03	**.45**	.03
while	.16	.05	.08	.03	.04	.02	.10	**.52**

Table 12: Confusion matrix based on percent agreement between subjects

As shown in Table 11 there is moderate agreement[4] among humans when selecting an appropriate temporal marker for a main and a subordinate clause. The ensemble's agreement with the gold standard approximates human performance on the interpretation task ($K = .413$ for Ensemble-Gold vs. $K = .421$ for Human-Gold). The agreement of the ensemble with the subjects is also close to the upper bound, i.e., inter-subject agreement (see Ensemble-Human and Human-Human in Table 11). Further analysis revealed that the majority of disagreements among our subjects arose for *as* and *once* clauses. *Once* was also problematic for the ensemble model (see Table 10). The inter-subject agreement was 33% for *as* clauses and 35% for *once* clauses. For the other markers, the subject agreement was around 55%. The highest agreement was observed for *since* and *until* (63% and 65% respectively). A confusion matrix summarizing the resulting inter-subject agreement for the interpretation task is shown in Table 12.

The moderate agreement is not entirely unexpected given that some of the markers are semantically similar and in some cases more than one marker are compatible with the temporal implicatures that arise from joining the two clauses. For example, *when* can be compatible with *after*, *as*, *before*, *once*, and *since*. Besides *when*, *as* can be compatible with *since*, and *while*. Consider for example the following sentence from our experimental materials: *More and more older women are divorcing when their husbands retire.* Although *when* is the right connective according to the corpus, *once*

4. Landis and Koch (1977) give the following five qualifications for different values of Kappa: .00–.20 is slight, .21–.40 is fair, .41–.60 is moderate, .61–.80 is substantial, whereas .81–1.00 is almost perfect.

or *after* are also valid choices. Indeed *after* is often chosen instead of *when* by our subjects (see Table 12). Also note that neither the model nor the subjects have access to the context surrounding the sentence whose marker must be inferred. In the sentence *A lot of them want to get out before they get kicked out* (again taken from our materials), knowing the referents of *them* and *they* is important in selecting the right relation. In some cases, substantial background knowledge is required to make a valid temporal inference. In the sentence *Are more certified deaths required before the FDA acts?* (see Appendix A), one must know what FDA stands for (i.e., Federal, Food, Drug, and Cosmetic Act). In a less strict evaluation setting where more than one connective are considered correct (on the basis of semantic compatibility), the inter-subject agreement is $K = .640$ (67.7%). Moreover, the ensemble's agreement with the subjects is $K = .609$ (67%).

We next evaluate the performance of the ensemble model on a more challenging task. Our test data so far has been somewhat artificially created by removing the temporal marker connecting a main and subordinate clause. Although this experimental setup allows to develop and evaluate temporal inference models relatively straightforwardly, it remains unsatisfactory. In most cases a temporal model would be required for interpreting events that are not only attested in main-subordinate clauses but in a variety of constructions (e.g., in parataxis or indirect speech) which may not contain temporal markers. We use the annotations in the TimeBank corpus for investigating whether our model, which is trained on automatically annotated data, performs well on a more realistic test set.

7. Experiment 3: Predicting TimeML Relations

Method As mentioned earlier the TimeBank corpus has been manually annotated with the TimeML coding scheme. In this scheme, verbs, adjectives, and nominals are annotated as EVENTs and are marked up with attributes such as the class of the event (e.g., state, reporting), its tense (e.g., present, past), aspect (e.g., perfective, progressive), and polarity (positive or negative). The TLINK tag is used to represent temporal relationships between events, or between an event and a time. These relationships can be inter- or intra-sentential. Table 13 illustrates the TLINK relationships with sentences taken from the TimeBank corpus. We focus solely on intra-sentential temporal relations between events; Table 13 does not include the IDENTITY relationship which is commonly attested inter-sententially.

Our intent here is to use the model presented in the previous sections to interpret the temporal relationships between events like those shown in Table 13 in the absence of overtly verbalized temporal information (e.g., temporal markers). However, one stumbling block to performing this kind of evaluation is that the corpus on which our model was trained uses different labels from those in Table 13 (e.g., (ambiguous) temporal markers like *when*). Fortunately, the temporal markers we considered and the TimeML relations are more or less semantically compatible, and so a mapping can be devised. First notice that some of the relations in Table 13 are redundant. For instance BEFORE is the inverse of AFTER, IS_INCLUDED is the inverse of INCLUDES, and so on. Furthermore, some semantic distinctions are too fine-grained for our model to identify accurately (e.g., BEFORE and IBEFORE (immediately before), SIMULTANEOUS and DURING). We therefore reduced the relations in Table 13 into a smaller set by collapsing BEFORE, IBEFORE, AFTER and IAFTER (immediately after) into one relationship. Analogously, we collapsed SIMULTANEOUSLY and DURING, INCLUDES and IS_INCLUDED, BEGINS and BEGUN_BY, and ENDS and ENDED_BY. The reduced relation set is also shown in Table 13 (within parentheses).

BEFORE	(BEFORE)	Pacific First Financial Corp. **said** shareholders **approved** its acquisition by Royal Trusstco Ltd. of Toronto for $27 a share, or $212 million.
IBEFORE	(BEFORE)	The first would be to launch the much-feared direct invasion of Saudi Arabia, hoping to **seize** some Saudi oil fields and **improve** his bargaining position.
AFTER	(BEFORE)	In Washington today the Federal Aviation Administration **released** air traffic control tapes from the night the TWA flight eight hundred **went** down.
IAFTER	(BEFORE)	In addition, Hewlett-Packard **acquired** a two-year option to **buy** an extra 10%, of which half may be sold directly to Hewlett-Packard by Octel.
INCLUDES	(INCLUDES)	Under the offer, shareholders will **receive** one right for each 105 common shares **owned**.
IS_INCLUDED	(INCLUDES)	The purchase price was **disclosed** in a preliminary prospectus **issued** in connection with MGM Grand's planned offering of six million common shares.
DURING	(INCLUDES)	According to Jordanian officials, a smaller line into Jordan **remained operating**.
ENDS	(ENDS)	The government may move to **seize** the money that Mr. Antar is **using** to pay legal fees.
ENDED_BY	(ENDS)	The Financial Times 100-share index **shed** 47.3 points to **close** at 2082.1, down 4.5% from the previous Friday.
BEGINS	(BEGINS)	DPC, an investor group led by New York-based Crescott Investment Associates, had itself **filed** a suit in state court in Los Angeles **seeking** to nullify the agreement.
BEGUN_BY	(BEGINS)	Saddam said he will **begin withdrawing** troops from Iranian territory on Friday and release Iranian prisoners of war.
SIMULTANEOUS		Nearly 200 Israeli soldiers have been **killed fighting** Hezbollah and other guerrillas guerrillas.

Table 13: TILINK relationships in TimeBank; the events participating in the relationship are marked with boldface; a more coarse-grained set of relationships is shown within parentheses.

We next defined a mapping between our temporal connectives and the reduced set of TimeML relations (see Table 14). Such a mapping cannot be one-to-one, since some of our connectives are compatible with more than one temporal relationship (see Section 4.1). For instance *when* can indicate an INCLUDES or BEFORE relationship. We also expect this mapping to be relatively noisy given that some temporal markers entail non-temporal relationships (e.g., *while*). Table 14 includes an additional relation, namely "no-temp-rel". We thus have the option of not assigning any temporal relation, thereby avoiding the pitfall of making a wrong prediction in cases where non-temporal

TMark	TimeMLRel	TrainInst	TestInst
after,before,once,when	BEFORE	31,643	877
as,when,while	INCLUDES	21,859	246
as,when,while	SIMULTANEOUS	22,165	360
since	BEGINS	2,810	19
until	ENDS	5,333	64
no-temp-rel	NO-TEMP-REL	22,523	967

Table 14: Mapping between temporal markers and coarse-grained set of TimeML relations; number of training and test instances per relation.

inferences are entailed by any two events. We next describe how training and test instances were generated for our experiments.

The disjunctive ensemble model from Experiment 1 was trained on the BLLIP corpus using the same features and component learners described in Sections 4.2 and 5. The training data consisted of our original 83,810 main-subordinate clause pairs labeled with the temporal relations from Table 14 (second column). To these we added 22,523 instances representative of the NO-TEMP-REL relation. Such instances were gathered by randomly concatenating main and subordinate clauses belonging to different documents (for a similar method, see Marcu & Echihabi, 2002). We hypothesize that the two clauses do not trigger temporal relations, since they are neither syntactically nor semantically related. Instances with connectives *since* and *once* were mapped to labels BEGINS and ENDS, respectively. In addition to BEGINS, *since* can signal BEFORE, INCLUDES, and SIMULTANEOUS temporal relations. However, in our experiments instances with *since* were used to exclusively learn the BEGINS relation. This is far from perfect, but we felt necessary since BEGINS is not represented by any other temporal marker. The training instances for *as* and *while* were equally split between the relationships INCLUDES and SIMULTANEOUS. Similarly, the data for *when* was equally split among BEFORE, INCLUDES, and SIMULTANEOUS. Instances with *after*, *before*, and *once* were exclusively used for learning the BEFORE relation. The number of training instances per relation (TrainInst) is given in Table 14.

As test data, we used sentences from the TimeBank corpus. We only tested the ensemble model on intra-sentential event-event relations. Furthermore, we excluded sentences with overt temporal connectives, as we did not want to positively influence the model's performance. The TimeBank corpus is not explicitly annotated with the NO-TEMP-REL relation. There are however sentences in the corpus whose events do not participate in any temporal relationship. We therefore hypothesized that these sentences were representative of NO-TEMP-REL. The total number of test instances (TestInst) used in this experiment is given in Table 14.

Results Our results are summarized in Table 15. We compare the performance of the disjunctive ensemble from Section 5 against a naive word-based model. Both these models were trained on main and subordinate clauses from the BLLIP corpus. We also report the accuracy of a majority baseline which defaults to the most frequent class in the BLLIP training data (i.e., BEFORE). Finally, we report the performance of a (disjunctive) ensemble model that has been trained and tested on the TimeBank corpus (see the column TestInst in Table 14) using leave-one-out crossvalidation. Comparison between the latter model and the BLLIP-trained ensemble will indicate whether unan-

Model	TrainCorpus	Accuracy	F-score
Majority Baseline	BLLIP	34.7	NA
Word-based Baseline	BLLIP	39.1*	21.1
Ensemble (Disjunctive)	BLLIP	53.0*†	45.8
Ensemble (Disjunctive)	TimeBank	42.7	40.5

Table 15: Results on predicting TimeML event-event relationships; comparison between word-based baseline and disjunctive ensemble models.

	BLLIP		TimeBank	
TimeMLRel	Accuracy	F-score	Accuracy	F-score
BEFORE	46.4	47.6	63.2	53.2
BEGINS	10.5	7.8	0.0	0.0
ENDS	14.1	3.7	4.7	7.7
INCLUDES	50.0	51.5	8.5	9.8
SIMULTANEOUS	46.7	47.8	6.7	8.9
NO-TEMP-REL	62.8	66.1	49.6	53.5
All	53.0	45.8	42.7	40.5

Table 16: Ensemble results on inferring individual temporal relations; comparison between ensemble model trained on BLLIP and TimeBank corpora.

notated data is indeed useful in reducing annotation effort and training requirements for temporal interpretation models.

As can be seen, the disjunctive model trained on the BLLIP corpus significantly outperforms the two baseline models. It also outperforms the ensemble model trained on TimeBank by a wide margin.[5] We find these results encouraging considering the approximations in our temporal interpretation model and the noise inherent in the BLLIP training data. Also note that, despite being linguistically informed, our feature space encodes very basic semantic and temporal distinctions. For example, aspectual information is not taken into account, and temporal expressions are not analyzed in detail. One would hope that more extensive feature engineering would result in improved results.

We further examined how performance varies for each class. Table 16 provides a comparison between the two ensemble models trained on BLLIP and the TimeBank corpus, respectively. Both models have difficulty with BEGINS and ENDS classes. This is not entirely surprising, since these classes are represented by a relatively small number of training instances (see Table 14). The two models yield comparable results for BEFORE, whereas the BLLIP-trained ensemble delivers better performance for INCLUDES, SIMULTANEOUS, and NO-TEMP-REL.

5. Unfortunately, we cannot use a χ^2 test to assess whether the differences between the two ensembles are statistically significant due to the leave-one-out crossvalidation methodology employed when training and testing on the TimeBank corpus. This was necessary given the small size of the event-event relation data extracted from TimeBank (2,533 instances in total, see Table 14).

We are not aware of any previous work that attempts to do a similar task. However, it is worth mentioning Boguraev and Ando (2005) who consider the interpretation of event-time temporal relations inter- and intra-sententially. They report accuracies ranging from 53.1% and 58.8% depending on the intervening distance between the events and the times in question (performance is better for events and times occurring close to each other). Interestingly, their interpretation model exploits unannotated corpora *in conjunction* with TimeML annotations to increase the amount of labeled data for training. Their method identifies unannotated instances that are distributionally similar to the manually annotated corpus. In contrast, we rely solely on unannotated data during training while exploiting instances explicitly marked with temporal information. An interesting future direction is the combination of such data with TimeML annotations as a basis for devising improved models (for details, see Section 8).

8. General Discussion

In this paper we proposed a data intensive approach to temporal inference. We introduced models that learn temporal relations from sentences where temporal information is made explicit via temporal markers and assessed their potential in inferring relations in cases where overt temporal markers are absent. Previous work has focused on the automatic tagging of temporal expressions (Wilson et al., 2001), on learning the ordering of events from manually annotated data (Mani et al., 2003), and inferring the temporal relations between events and time expressions from both annotated and unannotated data (Boguraev & Ando, 2005).

Our models bypass the need for manual annotation by training exclusively on instances of temporal relations that are made explicit by the presence of temporal markers. We compared and contrasted several models varying in their linguistic assumptions and employed feature space. We also explored the tradeoff between model complexity and data requirements. Our results indicate that less sophisticated models (e.g., the disjunctive model) tend to perform reasonably when utilizing expressive features and training data sets that are relatively modest in size. We experimented with a variety of linguistically motivated features ranging from verbs and their semantic classes to temporal signatures and argument structure. Many of these features were inspired by symbolic theories of temporal interpretation, which often exploit semantic representations (e.g., of the two clauses) as well as complex inferences over world knowledge (e.g., Hobbs et al., 1993; Lascarides & Asher, 1993; Kehler, 2002).

Our best model achieved an F-score of 69.1% on inferring temporal relations when trained and tested on the BLLIP corpus in the context of a pseudo-disambiguation task. This performance is a significant improvement over the baseline and compares favorably with human performance on the same task. Detailed exploration of the feature space further revealed that not only lexical but also syntactic information is important for temporal inference. This result is in agreement with Soricut and Marcu (2003) who find that syntax trees encode sufficient information to enable accurate derivation of discourse relations.

We also evaluated our model's performance on the more realistic task of predicting temporal relations when these are not explicitly signaled in text. To this end, we evaluated a BLLIP-trained model against TimeBank, a corpus that has been manually annotated with temporal relations according to the TimeML specifications. This experimental set-up was challenging from many perspectives. First, some of the temporal markers used in our study received multiple meanings. The ambiguity unavoidably introduced a certain amount of noise in estimating the parameters of our model and defining a mapping between markers and TimeML relations. Second, there is no guarantee that the relations signaled by temporal markers connecting main and subordinate clauses hold for events attested in other syntactic configurations such as non-temporal subordination or coordination. Given these approximations, our model performed reasonably, reaching an overall F-score of 45.8% on the temporal inference task and showing best performance for relations BEFORE, INCLUDES, SIMULTANEOUS and NO-TEMP-REL. These results show that it is possible to infer temporal information from corpora even if they are not semantically annotated in any way and hold promise for relieving the data acquisition bottleneck associated with creating temporal annotations.

An important future direction lies in modeling the temporal relations of events across sentences. In order to achieve full-scale temporal reasoning, the current model must be extended in a number of ways. These involve the incorporation of extra-sentential information to the modeling task as well as richer temporal information (e.g., tagged time expressions; see Mani et al., 2003). The current models perform the inference task independently of their surrounding context. Experiment 2 revealed this is a rather difficult task; even humans cannot easily make decisions regarding temporal relations out-of-context. In future work, we plan to take into account contextual (lexical and syntactic) as well as discourse-based features (e.g., coreference resolution). Many linguists have also observed that identifying the *discourse structure* of a text, conceptualized as a hierarchical structure of rhetorically connected segments, and identifying the temporal relations among its events are logically co-dependent tasks (e.g., Kamp & Reyle, 1993; Hobbs et al., 1993; Lascarides & Asher, 1993). For example, the fact that we interpret (1a) as forming a *narrative* with (1c) and (1c) as providing *background* information to (1b) yields the temporal relations among the events that we described in Section 1: namely, the temporal progression between kissing the girl and walking home, and the temporal overlap between remembering talking to her and walking home.

(1) a. John kissed the girl he met at a party.
 b. Leaving the party, John walked home.
 c. He remembered talking to her and asking her for her name.

This logical relationship between discourse structure and temporal structure suggests that the output of a *discourse parser* (e.g., Marcu, 1999; Soricut & Marcu, 2003; Baldridge & Lascarides, 2005) could be used as an informative source of features for inferring temporal relations across sentence boundaries. This would be analogous at the discourse level to the use we made here of a sentential parser as a source of features in our experiments for inferring sentence-internal temporal relations.

The approach presented in this paper can also be combined with the annotations present in the TimeML corpus in a semi-supervised setting similar to Boguraev and Ando (2005) to yield improved performance. Another interesting direction for future work would be to use the models proposed here in a bootstrapping approach. Initially, a model is learned from unannotated data and its output is manually edited following the "annotate automatically, correct manually" methodology used to provide high volume annotation in the Penn Treebank project. At each iteration the model is retrained on progressively more accurate and representative data. Another issue related to the nature of our training data concerns the temporal information entailed by some of our markers which can be ambiguous. This could be remedied either heuristically as discussed in Section 4.1 or by using models trained on unambiguous markers (e.g., *before*, *after*) to disambiguate instances with multiple readings. Another possibility is to apply a separate disambiguation procedure on the training data (i.e., prior to the learning of temporal inference models).

Finally, we would like to investigate the utility of these temporal inference models within the context of specific natural language processing applications. We thus intend to explore their potential in improving the performance of a multi-document summarisation system. For example, a temporal reasoning component could be useful not only for extracting temporally congruent events, but also for structuring the output summaries, i.e., by temporally ordering the extracted sentences. Although the models presented here target primarily interpretation tasks, they could also be adapted for generation tasks, e.g., for inferring if a temporal marker should be generated and where it should be placed.

Acknowledgments

This work was supported by EPSRC (Lapata, grant GR/T04540/01; Lascarides, grant GR/R40036/01). We are grateful to Regina Barzilay and Frank Keller for helpful comments and suggestions. Thanks to the anonymous referees whose feedback helped to substantially improve the present paper. A preliminary version of this work was published in the proceedings of NAACL 2004; we also thank the anonymous reviewers of that paper for their comments.

Appendix A. Experimental Materials for Human Evaluation

The following is the list of materials used in the human evaluation study reported in Experiment 2 (Section 6). The sentences were extracted from the BLLIP corpus following the procedure described in Section 4.1.

1	In addition, agencies weren't always efficient in getting the word to other agencies _____ the company was barred.	**when**
2	Mr. Reagan learned of the news _____ National Security Adviser Frank Carlucci called to tell him he'd seen it on television. _____	**when**
3	For instance, National Geographic caused an uproar _____ it used a computer to neatly move two Egyptian pyramids closer together in a photo.	**when**
4	Rowes Wharf looks its best _____ seen from the new Airport Water Shuttle speeding across Boston harbor.	**when**
5	More and more older women are divorcing _____ their husbands retire.	**when**
6	Together they prepared to head up a Fortune company _____ enjoying a tranquil country life.	**while**
7	_____ it has been estimated that 190,000 legal abortions to adolescents occurred, an unknown number of illegal and unreported abortions took place as well.	**while**
8	Mr. Rough, who is in his late 40s, allegedly leaked the information _____ he served as a New York Federal Reserve Bank director from January 1982 through December 1984.	**while**
9	The contest became an obsession for Fumio Hirai, a 30-year-old mechanical engineer, whose wife took to ignoring him _____ he and two other men tinkered for months with his dancing house plants.	**while**
10	He calls the whole experience "wonderful, enlightening, fulfilling" and is proud that MCI functioned so well _____ he was gone.	**while**
11	A lot of them want to get out _____ they get kicked out.	**before**
12	_____ prices started falling, the market was doing $1.5 billion a week in new issues, says the head of investment banking at a major Wall Street firm.	**before**
13	But _____ you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers.	**before**
14	The Organization of Petroleum Exporting Countries will travel a rocky road _____ its Persian Gulf members again rule world oil markets.	**before**
15	Are more certified deaths required _____ the FDA acts?	**before**
16	Currently, a large store can be built only _____ smaller merchants in the area approve it, a difficult and time consuming process.	**after**
17	The review began last week _____ Robert L. Starer was named president.	**after**
18	The lower rate came _____ the nation's central bank, the Bank of Canada, cut its weekly bank rate to 7.2% from 7.54%.	**after**
19	Black residents of Washington's low-income Anacostia section forced a three-month closing of a Chinese-owned restaurant _____ the owner threatened an elderly black woman customer with a pistol.	**after**
20	Laurie Massa's back hurt for months _____ a delivery truck slammed into her car in 1986.	**after**

Table 17: Materials for the temporal pseudo-disambiguation task; markers in bodlface indicate the gold standard completion; subjects were asked to select the missing word from the set of temporal markers {*after, before, while, when, as, once, until, since*}

21	Donald Lasater, 62, chairman and chief executive office, will assume the posts Mr. Farrell vacates _____ a successor is found.	**until**
22	The council said that the national assembly will be replaced with appointed legislators and that no new elections will be held _____ the U.S. lifts economic sanctions.	**until**
23	_____ those problems disappear, Mr. Melzer suggests working with the base, the raw material for all forms of the money supply.	**until**
24	A green-coffee importer said there is sufficient supply in Brazil _____ the harvest gets into full swing next month.	**until**
25	They will pump _____ the fire at hand is out.	**until**
26	_____ the gene is inserted in the human TIL cells, another safety check would be made.	**once**
27	_____ part of a bus system is subject to market discipline, the entire operation tends to respond.	**once**
28	In China by contrast, _____ joint ventures were legal, hundreds were created.	**once**
29	The company said the problem goes away _____ the car warms up.	**once**
30	_____ the Toronto merger is complete, the combined entity will have 352 lawyers.	**once**
31	The justices ruled that his admission could be used _____ he clearly had chosen speech over silence.	**since**
32	Milosevic's popularity has risen _____ he became party chief in Serbia, Yugoslavia's biggest republic, in 1986.	**since**
33	The government says it has already eliminated 600 million hours of paperwork a year _____ Congress passed the Paperwork Reduction Act in 1980.	**since**
34	It was the most serious rebellion in the Conservative ranks _____ Mr. Mulroney was elected four years ago.	**since**
35	There have been at least eight settlement attempts _____ a Texas court handed down its multi-billion dollar judgment two years ago.	**since**
36	Brud LeTourneau, a Seattle management consultant and Merit smoker, laughs at himself _____ he keeps trying to flick non-existent ashes into an ashtray.	**as**
37	Britain's airports were disrupted _____ a 24-hour strike by air traffic control assistants resulted in the cancellation of more thank 500 flights and lengthy delays for travelers.	**as**
38	Stocks plunged _____ investors ignored cuts in European interest rates and dollar and bond rallies.	**as**
39	At Boston's Logan Airport, a Delta plane landed on the wrong runway _____ another jet was taking off.	**as**
40	Polish strikers shut Gdansk's port _____ Warsaw rushed riot police to the city.	**as**

Table 17: (*continued*)

References

Allen, J. (1995). *Natural Language Understanding*. Benjamin Cummins.

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Baldridge, J., & Lascarides, A. (2005). Probabilistic head-driven parsing for discourse structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pp. 96–103, Ann Arbor, MI.

Boguraev, B., & Ando, R. K. (2005). TimeML-compliant text analysis for temporal reasoning. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 997–1003, Edingburgh, UK.

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, *2*(24), 123–140.

Breiman, L. (1996b). Stacked regressions. *Machine Learning*, *3*(24), 49–64.

Carlson, L., Marcu, D., & Okurowski, M. (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001*, Aalborg, Denmark.

Cestnik, B. (1990). Estimating probabilities: a crucial task in machine learning. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pp. 147–149, Stockholm, Sweden.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Assocation for Computational Linguistics*, pp. 132–139, Seattle, WA.

Cherkauer, K. J. (1996). Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. In *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, pp. 15–21, Portland, OR.

Ciaramita, M., & Johnson, M. (2003). Supersense tagging of unknown words in WordNet. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing*, pp. 168–175, Sapporo, Japan.

Dietterich, T. G. (1997). Machine learning research: Four current directions. *AI Magazine*, *18*(4), 97–136.

Dorr, B., & Gaasterland, T. (1995). Selecting tense aspect and connective words in language generation. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1299–1307, Montreal, Canada.

Dowty, D. (1986). The effects of aspectual class on the temporal sturcture of discourse: Semantics or pragmatics?. *Linguistics and Philosophy*, *9*(1), 37–61.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Ferro, L., Mani, I., Sundheim, B., & Wilson, G. (2000). TIDES temporal annotation guidelines. Tech. rep., The MITRE Corporation.

Freund, Y., & Shapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 148–156, Stanford, CA.

Han, B., & Lavie, A. (2004). A framework for resolution of time in natural language. *ACM Transactions on Asian Language Information Processing (TALIP)*, *3*(1), 11–32.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, *12*, 993–1001.

Hitzeman, J., Moens, M., & Grover, C. (1995). Algorithms for analyzing the temporal structure of discourse. In *Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics*, pp. 253–260, Dublin, Ireland.

Hobbs, J. R., Stickel, M., Appelt, D., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence*, *63*(1–2), 69–142.

Hwang, C., & Schubert, L. (1992). Tense trees as the finite structure of discourse. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 232–240, Newark, DE.

Kamp, H., & Reyle, U. (1993). *From Discourse to the Lexicon: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

Katz, G., & Arosio, F. (2001). The annotation of temporal information in natural language sentences. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, pp. 104–111, Toulouse, France.

Kehler, A. (2002). *Coherence, Reference and the Theory of Grammar*. CSLI Publications, Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lapata, M., & Brew, C. (2004). Verb class disambiguation using informative priors. *Computational Linguistics*, *30*(1), 45–73.

Lascarides, A., & Asher, N. (1993). Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, *16*(5), 437–493.

Levin, B. (1995). *English Verb Classes and Alternations*. Chicago University Press.

Mani, I., & Schiffman, B. (2005). Temporally anchoring and ordering events in news. In Pustejovsky, J., & Gaizauskas, R. (Eds.), *Time Event Recognition and Natural Language*. John Benjamins.

Mani, I., Schiffman, B., & Zhang, J. (2003). Inferring temporal ordering of events in news. In *Proceedings of the 1st Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 55–57, Edmonton, Canada.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 365–372, College Park, MD.

Marcu, D., & Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 368–375, Philadelphia, PA.

Moens, M., & Steedman, M. J. (1988). Temporal ontology and temporal reference. *Computational Linguistics*, *14*(2), 15–28.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., & Setzer, A. (2004). The specification of TimeML. In Mani, I., Pustejovsky, J., & Gaizauskas, R. (Eds.), *The Language of Time: A reader*, pp. 545–558. Oxford University Press.

Pustejovsky, J., Mani, I., Belanger, L., Boguraev, B., Knippen, B., Litman, J., Rumshisky, A., See, A., Symonen, S., van Guilder, J., van Guilder, L., & Verhagen, M. (2003). ARDA summer workshop on graphical annotation toolkit for TimeML. Tech. rep..

Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning*. Series in Machine Learning. Morgan Kaufman, San Mateo, CA.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, London.

Saurí, R., Littman, J., Gaizauskas, R., Setzer, A., & Pustejovsky, J. (2004). *TimeML Annotation Guidelines*. TERQAS Workshop. Version 1.1.

Schilder, F., & Habel, C. (2001). From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, pp. 65–72, Toulouse, France.

Setzer, A., & Gaizauskas, R. (2001). A pilot study on annotating temporal relations in text. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, pp. 73–80, Toulouse, France.

Siegel, S., & Castellan, N. J. (1988). *Non Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Soricut, R., & Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 1st Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 228–235, Edmonton, Canada.

Sporleder, C., & Lascarides, A. (2005). Exploiting linguistic cues to classify rhetorical relations. In *Proceedings of Recent Advances in Natural Language Processing*, pp. 532–539, Borovets, Bulgaria.

van Halteren, H., Zavrel, J., & Daelemans, W. (2001). Improving accuracy in word class tagging through combination of machine learning systems. *Computational Linguistics*, *27*(2), 199–230.

Wiebe, J. M., O'Hara, T. P., Öhrström Sandgren, T., & McKeever, K. J. (1998). An empirical approach to temporal reference resolution. *Journal of Artifical Intelligence Research*, *9*, 247–293.

Wilson, G., Mani, I., Sundheim, B., & Ferro, L. (2001). A multilingual approach to annotating and extracting temporal information. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*, pp. 81–87, Toulouse, France.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.