Learning Sentence-internal Temporal Relations

Mirella Lapata Alex Lascarides MLAP@INF.ED.AC.UK 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. Our approach bypasses the need for manual coding by exploiting the presence of temporal markers like *after*, which overtly signal a temporal relation. Our experiments concentrate on two tasks relevant for applications which either extract or synthesise temporal information (e.g., summarisation, question answering). Our first task focuses on interpretation: given a subordinate clause and main clause, identify the temporal relation between them. The second is a fusion task: given two clauses and a temporal relation between them, decide which one contained the temporal marker (i.e., identify the subordinate and main clause). We compare and contrast several probabilistic models differing in their feature space, linguistic assumptions and data requirements. We evaluate performance against a gold standard corpus and also against human subjects performing the same tasks. The best model achieves 69.1% F-score in inferring the temporal relation between two clauses and 93.4% F-score in distinguishing the main vs. the subordinate clause, assuming that the temporal relation is known.

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 summarisation, for example, information that is to be included in the summary must be extracted from various documents and synthesised into a meaningful text. Knowledge about the temporal order of events is important for determining what content should be communicated (*interpretation*) and for correctly merging and presenting information in the summary (*generation*). Indeed, ignoring temporal relations in either the information extraction phase or the summary generation phase potentially results 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 normalisation of time expressions¹ (Wilson, Mani, Sundheim, & Ferro,

^{1.} See also the Time Expression Recognition and Normalisation (TERN) evaluation exercise (http://timex2.mitre.org/tern.html).

2001; Schilder & Habel, 2001; Wiebe, O'Hara, Öhrström Sandgren, & McKeever, 1998). Reasoning with time, however, goes beyond temporal expressions; it involves interpretation of the order of events in discourse, analysis of their temporal relations, and generally the ability to draw inferences over time elements. An additional challenge to this task poses the nature of temporal information itself which is often implicit (i.e., not overtly verbalised) 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*, signalling 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), the semantic relationships among the words in the clauses, and real 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 TimeBank² corpus, for example, contains a set of 186 news report documents annotated with the TimeML mark-up language for temporal events and expressions (see Section 2 for details). The corpus consists of 68.5K words in total. Contrast this with the Penntreebank, 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 timeconsuming 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 et al. (2003) found evidence that this annotation task is sufficiently complex that human annotators can realistically identify only a small number of the temporal relations that hold in reality; i.e., recall is compromised.

In default of large volumes of data labelled 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:

- (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.

^{2.} Available from http://www.cs.brandeis.edu/~jamesp/arda/time/timebank.html

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. 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 learnt by gathering informative features from the two clauses.

In this paper we focus on two tasks, both of which are important for any NLP system requiring information extraction and text synthesis. The first task addresses the interpretation of temporal relations: given a main and a subordinate clause, identify the temporal marker which connected them. So for this task, our models view the marker from each sentence in the training corpus as the label to be learnt. In the test corpus the marker is removed and the models' task is to pick the most likely label—or equivalently marker. Our second task concerns the generation of temporal relations. Non-extractive summarisers that produce sentences by fusing together sentence fragments (e.g., Barzilay, 2003) must be able to determine whether to include an overt temporal marker in the generated text, where the marker should be placed, and what lexical item should be used. Rather than attempting all three tasks at once, we focus on determining the appropriate ordering among a temporal marker and two clauses. We infer probabilistically which of the two clauses is introduced by the marker, and effectively learn to distinguish between main and subordinate clauses. In this case the main vs. subordinate clause are treated as labels. The test corpus consists of sentences with overtly marked temporal markers, however information regarding their position is removed. By the very nature of these tasks, our models focus exclusively on sentence-internal temporal relations. It is hoped that they can be used to infer temporal relations among events in data where overt temporal markers are absent (e.g., as in (1)), although this is beyond the scope of this paper.

In attempting to infer temporal relations probabilistically, we consider different 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 utilise 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 generalisations 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 4.2). We furthermore explore the benefits of ensemble learning methods for the two tasks introduced above and show that improved performance can be achieved when different learners (modelling complementary knowledge sources) are combined. Our machine learning experiments are complemented by a study in which we investigate human performance on our two tasks, thereby assessing their 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 utilises overt markers as a means for avoiding manual labelling 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 utilised a semantics and inference-based approach. This involves complex reasoning over a variety of rich information sources, including representations of domain knowledge and detailed logical forms of the clauses (e.g., Dowty, 1986; Hwang & Schubert, 1992; Hobbs et al., 1993; Lascarides & Asher, 1993; Kamp & Reyle, 1993a; Kehler, 2002). This approach, while theoretically elegant, is impractical except for applications in very narrow domains for a number of 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 et al. (1995), and Han and Lavie (2004) propose more computationally tractable approaches to inferring 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, 1993a). While this type of symbolic approach is promising, and overcomes some of the impracticalities of utilising 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 in 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 (see Section 1).

Mani et al. (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 recognising events and inferring temporal relations (between two events or 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 unlabelled 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 example, 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 these discourse structures to their temporal consequences; so although there is potential for eventually using linguistic resources labelled with discourse structure to acquire a model of temporal relations, that potential cannot be presently realised.

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 signalled by an unambiguous discourse connective (e.g., because for EXPLANATION and but for CONTRAST). They extract sentences containing such unambiguous 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 labelled 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 and Lascarides (2005) for more detailed discussion).

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 discourse connectives 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, we 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 temporal relations. Apart from the fact that our interepretation 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, we apply the proposed models to a generation task, namely information fusion. And finally, we evaluate the models against human subjects performing the same task, as well as against a gold standard corpus. In the following section we present our models and formalise our interpretation and generation tasks.

3. Problem Formulation

Interpretation 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 maximises $P(S_M, t_j, S_S)$:

$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(S_M, t_j, S_S)$$

$$t^* = \underset{t_i \in T}{\operatorname{argmax}} P(S_M) P(S_S | S_M) P(t_j | S_M, S_S)$$

$$(3)$$

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

$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(t_j | S_M, S_S)$$

$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(t_j) P(S_M, S_S | t_j)$$

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

$$t_j \in T$$

$$(4)$$

 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 $a_{\langle M,1\rangle}\cdots a_{\langle M,n\rangle}$, 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:

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

For example, if we were assuming our feature space consisted solely of nouns and verbs, we will 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 and 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., $P(S_S, S_M|t_i) \approx P(S_S|t_i)P(S_M|t_i)$).

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

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

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

Fusion For the sentence fusion task, the identity of the two clauses is unknown, and our task is to infer which clause contains the marker. Conjunctive and disjunctive models can be expressed as follows:

$$p^* = \underset{p \in \{M,S\}}{\operatorname{argmax}} P(t) \prod_{i=1}^{n} \left(P(a_{\langle p,i,\rangle}, a_{\langle \overline{p},i\rangle} | t) \right)$$
(8)

$$p^* = \underset{p \in \{M,S\}}{\operatorname{argmax}} P(t) \prod_{i=1}^{n} \left(P(a_{\langle p,i \rangle} | t) P(a_{\langle \overline{p},i \rangle} | t) \right)$$
(9)

where p is generally speaking a sentence fragment to be realised as a main or subordinate clause $(\{\overline{p} = S | p = M\})$ or $\{\overline{p} = M | p = S\}$, and t is the temporal marker linking the two clauses. Features are generated similarly to the interpretation case by taking the co-occurrences of temporal markers and individual features (disjunctive model) or feature pairs (conjuctive model) into account.

```
(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. 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. Next, 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 (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). In the testing phase, all occurrences of the relevant temporal markers are removed for the interpretation task and the model must decide which member of the confusion set to choose. For the sentence fusion task, it is the textual order of the two clauses that is unknown and must be inferred.

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 identify 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

Marker	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

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. 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 (10), 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):

(10) [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 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):

(11) [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 (12) 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.

(12) [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 (13a), from Kamp & Reyle, 1993a), or temporal progression (see (13c), from Moens & Steedman, 1988). The same is true for *once* and *since*:

- (13) a. Mary left when Bill was preparing dinner.
 - b. When they built the bridge, they solved all their traffic problems.
- (14) a. Once John moved to London, he got a job with the council.
 - b. Once John was living was living in London, he got a job with the council.
- (15) a. John has worked for the council since he's been living in London.
 - b. John moved to London since he got a job with the council there.

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 to two, 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 (16a)) 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 real world knowledge; see (16b)). The maker *as* can also denote two relations, i.e., overlap (see 17a) or cause (see 17b).

- (16) 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.
- (17) a. Grand melodies poured out of him as he contemplated Caesar's conquest of Egypt.
 - b. I wen to the bank as I run out of cash.

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, whereas once and since are more likely to entail temporal progression (74% and 80%, respectively). The markers as and while 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 the interence procedure has taken place, we could use these biases to disambiguate, albeit coarsely, markers with underspecified meanings.

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,

FINITE = {past, present}

NON-FINITE = {0, infinitive, ing-form, en-form}

MODALITY = {0, future, ability, possibility, obligation}

ASPECT = {imperfective, perfective, progressive}

VOICE = {active, passive}

NEGATION = {affirmative, negative}

Table 2: Temporal signatures

Feature	$once_M$	onces	$since_{M}$	sinces
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

2003). By selecting features that represent, albeit indirectly and imperfectly, these knowledge sources, 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.

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, 0, imperfective, passive, affirmative}, where 0 indicates the verb form is finite

TMark	Verb _M	Verb _S	Supersense _M	Supersenses	Levin _M	Levins
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)

and 0 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.

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 and Lascarides 2003). For example, the fact that verbs like *push* can cause movement of the patient and verbs like *fall* describe the movement of their subject can be used to interpret the discourse in (18) as the pushing causing the falling, thus making the linear order of the events mismatch their temporal order.

(18) Max fell. John pushed him.

We operationalise lexical relationships among verbs in our data by counting their occurrence in main and subordinate clauses from a lemmatised 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 (V_W, 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 maximise the linguistic generalisations captured by a model of temporal relations. For our purposes, there is an

additional empirical motivation for utilising 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 generalisation 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 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 lemmatised form.

Levin (1995) focuses on the relation between verbs and their arguments and hypothesises that verbs which behave similarly with respect to the expression and interpretation of their arguments share certain meaning components and can therefore be organised 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 in Lapata and Brew (1999). Again, for verbs not included in Levin, the lemmatised verb form is 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 domain 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 quite frequent in this data set.

Nouns were extracted from a lemmatised version of the parser's output. 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 organisations (e.g., *United Laboratories Inc.*), person names (e.g., *Jose Y. Campos*), and locations (e.g., *New England*) which were subsequently substituted by the general categories person, organisation, and location.

Noun Class (N_W) As with verbs, Asher and Lascarides (2003) argue in favour of symbolic rules for inferring temporal relations that utilise the semantic classes of nouns wherever possible, so as to maximise the linguistic generalisations that are captured. For example, they argue that one can infer a causal relation in (19) 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*):

^{3.} Lapata and Brew (1999) 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.

TMark	Noun _N	Nouns	Supersense _M	Supersenses	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)

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

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 hypothesise 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 the marker interpretation task. 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 (20)).

(20) 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*, *realise*). 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 (18) 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 parsetrees. 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 the temporal interpretation and fusion tasks (Experiments 1 and 2) and then describe a study with humans (Experiment 3). The latter enables us to examine in more depth the models' performance and the difficulty of our inference tasks.

5. Experiment 1: Sentence Interpretation

Method Our models were trained on main and subordinate clauses extracted from the BLLIP corpus as detailed in Section 4. 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 3. 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 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

Symbols	Meaning
* *	(not) significantly different from Majority Baseline
† *	(not) significantly different from Word-based Baseline
\$ \$	(not) significantly different from Conjunctive Model
‡ <i>‡</i>	(not) significantly different from Disjunctive Model
# #	(not) significantly different Disjunctive Ensemble
& &	(not) significantly different Conjunctive Ensemble

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

Model	Accuracy	F-score
Majority Baseline	42.6†\$‡#&	NA
Word-based Baseline	48.2*\$‡#&	44.7
Conjunctive (V _W V _L PSV)	60.3*†‡#&	53.3
Disjunctive (SV)	62.6*†\$#&	62.3
Ensemble (Conjunctive)	64.5*†\$‡&	59.9
Ensemble (Disjunctive)	70.6*†\$‡#	69.1

Table 7: Summary of results for the sentence interpretation 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)

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 model on the test set (accuracy 62.6%) was observed with the combination of verbs (V) with syntactic signatures (S) for the disjunctive model (see Table 7). 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 (see Table 7). 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 ($V_W V_L PSV$).

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 favours a simpler model over a more complex one. The difference in performance between the models relying on linguistically-motivated

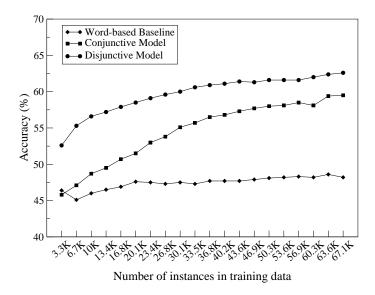


Figure 2: Learning curve for conjunctive, disjunctive, and word-based models; sentence interpretation

features and the word-based model also shows, in line with the findings in Sporleder and Lascarides (2005), that linguistic abstractions are useful in overcoming sparse data.

We further analysed 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 (SV and V_WV_LPSV). 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 linguistically aware models. This result is in agreement with Marcu and Echihabi (2002) who employ a very large corpus (1 billion words) for training their word-based model.

Further analysis of our models 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 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 in some way to classify new examples. This simple idea has been applied to a variety of classification problems ranging from optical character recognition to medical diagnosis

	Disjunctive Model		Conjunctive Model	
TMark	Features	Accuracy	Features	Accuracy
after	NPRSTV	69.9	$V_{W}PTV$	79.6
as	ANN_WPSV	57.0	$V_W V_L S V$	57.0
before	SV	42.1	TV	11.3
once	PRS	40.7	$V_{W}P$	3.7
since	PRST	25.1	$V_L V$	1.03
when	V_LPS	85.5	V_LNV	86.5
while	PST	49.0	$V_L PV$	9.6
until	$V_L V_W RT$	69.4	$V_W V_L P V$	9.5

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

and part-of-speech tagging (see Dietterich, 1997 and van Halteren, Zavrel, & Daelemans, 2001 for overviews). 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, i.e., first multiple component learners are trained and 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 (see Quinlan, 1993 for details). 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.

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 dis-

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	$N_W PSVV_L$	PSV_L	PVV_L	NPSV
NPV	NSV	PSV	PV	SV	TV	V
		Disjuncti	ve Ensemble			
AN _W NPSV	APSV	ASV	$PRSV_W$	PSV _N	SV_L	NPRSTV
PRS	PRST	PRSV	PSV	SV		

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

junctive 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 have fewest examples in our data set (see Table 1). Secondly, once is temporally ambiguous, conveying temporal progression and temporal overlap (see example (14)). The same ambiguity is observed with since (see example (15)). 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 can be explained by the fact that both markers convey similar relations: they both imply temporal overlap but also have contrastive usages (thereby entailing temporal progression).

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 the interpretation task, at least for the WSJ domain. Our results so far point to the importance of the lexicon for the marker interpretation task but also indicate that the syntactic complexity of the two clauses is crucial for inferring their semantic relation. Asher and Lascarides' (2003) symbolic theory of discourse interpretation also emphasises the importance of lexical information in inferring temporal relations.

	Disjunctive Ensemble		Conjunctive Ensemble	
TMark	Accuracy	F-score	Accuracy	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	17.06	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)

Model	Accuracy	F-score
Random Baseline	50.0†\$‡#&	NA
Word-based Baseline	64.0*\$‡#&	64.6
Conjunctive (NT)	68.3*†‡#&	67.2
Disjunctive (ARSV)	80.1*†\$#&	78.4
Ensemble (Conjunctive)	80.8*†\$‡&	89.4
Ensemble (Disjunctive)	97.3*†\$‡#	93.4

Table 11: Summary of results for the sentence fusion task; comparison of baseline models against conjunctive and disjunctive models and their ensembles (N:nouns, T:temporal signature, A:adjectives, S:syntactic signature, V:verbs, R:argument signature)

6. Experiment 2: Sentence Fusion

Method For the sentence fusion task we built models that used the feature space introduced in Section 4.2, with the exception of the position feature (P). Knowing the linear precedence of the two clauses is highly predictive of their type: 80.8% of the main clauses are sentence initial. However, this type of positional information is typically not known when fragments are synthesised into a meaningful sentence and was therefore not taken into account in our experiments. To find the best performing model, the feature space was exhaustively evaluated on the development set.

As in Experiment 1, we compared the performance of conjunctive and disjunctive models. These models were in turn evaluated against a word-based disjunctive model (where $P(a_{\langle p,i\rangle}=w_{\langle p,i\rangle}|t)$) and $P(a_{\langle \overline{p},i\rangle}=w_{\langle \overline{p},i\rangle}|t)$) and a simple baseline that decides which clause should be introduced by the temporal marker at random.

Results The best performing conjunctive and disjunctive models are presented in Table 11. The feature combination NT delivered the highest accuracy for the conjunctive model (68.3%), whereas ARSVV_W, was the best disjunctive model reaching an accuracy of 80.1%. Both models significantly outperformed the word-based model and the random guessing baseline. Similarly to the interpre-

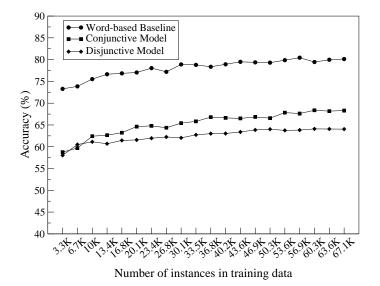


Figure 3: Learning curve for conjunctive, disjunctive, and word-based models; sentence fusion

	Conjunctive	Disjunctive Model		
TMark	Features	Accuracy	Features	Accuracy
after	NR	74.1	AVV_W	77.9
as	$NRSV_W$	54.4	AV	75.8
before	NRV_L	65.5	ANSTV	85.4
once	ANN_WSTVV_W	70.3	RT	100
since	NRV_LV_W	60.5	T	85.2
when	$NSTV_W$	53.8	RST	86.9
while	$ANSV_W$	61.9	SV_W	79.4
until	$ANRV_L$	65.5	TV	90.5

Table 12: Best feature combinations for individual markers (sentence fusion; development set; V: verbs, V_W: WordNet verb supersenses, V_L: Levin verb classes, N: nouns, N_W: WordNet noun supersenses, P: clause position, S: syntactic signature, R: argument signature)

tation task, the conjunctive model performs significantly worse than the disjunctive one. We also examined the amount of data required for achieving satisfactory performance. The learning curves are given in Figure 3. The disjunctive model achieves a good performance with approximately 3,000 training instances. Also note that the conjunctive model suffers from data sparseness (similarly to the word-based model). With increased amounts of training data, it manages to outperform the word-based model, without however matching the performance of the disjunctive model.

We next report on our experiments with ensemble models. Inspection of the performance of individual models on the development set revealed that they are complementary, i.e., they differ in their ability to perform the fusion task. Feature combinations with the highest accuracy (on the development set) for individual markers are shown in Table 12).

Conjunctive Ensemble						
AN_WNSTVV_L	ASV, N _W NS	N _W NST	NwST	N_WNT	NT	N _W NR
	I	Disjunctive 1	Ensemble			
$ANRSTVV_W$	AN _W NSTV	AN _W NV	AN _W RS	ANV	ARS	ARSTV
ARSV	ARV	AV	V_WHS	$V_{W}RT$	$V_{W}TV$	N_WRST
N_WS	N_WST	V_WT	$V_{\mathrm{W}}TV$	RT	STV	

Table 13: Component models for ensemble learning (sentence fusion; V: verbs, V_W: WordNet verb supersenses, V_L: Levin verb classes, N: nouns, N_W: WordNet noun supersenses, P: clause position, S: syntactic signature, R: argument signature)

	Conjunctive	Disjunctive
TMark	Accuracy	Accuracy
after	90.4	96.7
as	78.8	93.2
before	89.7	96.8
once	36.7	100
since	93.3	98.2
when	72.7	99.3
while	93.3	97.7
until	96.1	97.8

Table 14: Ensemble results on sentence fusion for individual markers (test set)

Ensemble Learning Similarly to the interpretation task, an ensemble of classifiers was built in order to take advantage of the complementarity of different models. The second-level decision tree learner was again trained on the development set using 10-fold cross-validation. We experimented with 77 conjunctive and 44 different disjunctive models; the component models for which we obtained the best results on the development set are shown in Table 13 and formed the ensemble whose performance was evaluated on the test set. The conjunctive ensemble reached an accuracy of 80.8%. The latter was significantly outperformed by the disjunctive ensemble whose accuracy was 97.3% (see Table 11). In comparison, the best performing model's accuracy on the test set (AR-STV, disjunctive) was 80.1%. Table 14 shows how well the ensembles are performing the fusion task for individual markers. We only report accuracy since the recall is always one. The conjunctive ensemble performs poorly on the fusion task when the temporal marker is *once*. This is to be expected, since *once* is the least frequent marker in our data set, and as we have already observed the conjunctive model is particularly prone to sparse data.

Not surprisingly, the features V and S are also important for the fusion task (see Table 14). Adjectives (A), nouns (N and N_W) and temporal signatures (T), all seem to play more of a role in this task than they did in the interpretation task. This is perhaps to be expected given that the differences between main and subordinate clauses are rather subtle (semantically and structurally) and more information is needed to perform the inference.

Although for the interpretation and fusion tasks the ensemble outperformed the single best model, it is worth noting that the best individual models (ARSTV and SV for fusion and interpretation, respectively) rely on features that can be simply extracted from the parse trees without recourse to taxonomic information. Removing from the disjunctive ensemble the feature combinations that rely on corpus external resources (i.e., Levin, WordNet) yields an overall accuracy of 65.0% for interpretation and 95.6% for fusion.

7. Experiment 3: Human Evaluation

Method We further compared our model's performance against human judges by conducting two separate studies, one fore interpretation an one one for fusion. In the first study, 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 models' task.

In the second study, participants were presented with a series of three sentence fragments and were asked to arrange them so that a coherent sentence is formed. The fragments were a main clause, a subordinate clause and a marker. Punctuation was removed so as not to reveal any ordering clues. Participants saw 40 such triples randomly selected from our test set. The set of items was different from those used in the interpretation task; again five items were selected for each marker. Examples of the materials our participants saw are given in Apendix A.

Both studies were conducted remotely over the Internet. Subjects first saw a set of instructions that explained the task, and had to fill in a short questionnaire including basic demographic information. For the interpretation task, a random order of main-subordinate pairs and a random order of markers per pair was generated for each subject. For the fusion task, a random order of items and a random order of fragments per item was generated for each subject. The interpretation study was completed by 198 volunteers, all native speakers of English. 100 volunteers participated in the fusion study, again all native speakers of English. Subjects were recruited via postings to local Email lists.

Results Our results are summarised in Table 15. We measured how well human subjects (H) agree with the gold standard (G)—i.e., the corpus from which the experimental items were selected—and how well they agree with each other. We also show how well the disjunctive ensembles (E) for the fusion and interpretation task respectively agree with the humans (H) and the gold standard (G). 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.

As shown in Table 15 there is less agreement among humans for the interpretation task than the sentence fusion task. This is expected 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. Also note that neither the model nor the subjects have access to the context surrounding the sentence whose marker must be inferred (we discuss this further in Section 8). Additional analysis of the interpretation data revealed that the majority of disagreements arose for as and once clauses. Once was also problematic for the ensemble model (see Table 10). Only 33% of the subjects agreed with the gold standard for as clauses; 35% of the subjects agreed with the gold standard was

	Interpretation		Fusion	
	K %		K	%
Н-Н	.410	45.0	.490	70.0
H-G	.421	46.9	.522	79.2
E-H	.390	44.3	.468	70.0
E-G	.413	47.5	.489	75.0

Table 15: Agreement figures for subjects and disjunctive ensemble (H-H: intersubject agreement, H-G: agreement between subjects and gold standard, E-H: agreement between ensemble and subjects, E-G: agreement between ensemble and gold standard)

	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 16: Confusion matrix based on percent agreement between subjects

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 16.

The ensemble's agreement with the gold standard approximates human performance on the interpretation task (.413 for E-G vs. .421 for H-G). The agreement of the ensemble with the subjects is also close to the upper bound, i.e., inter-subject agreement (see, E-H and H-H in Table 15). A similar pattern emerges for the fusion task: comparison between the ensemble and the gold standard yields an agreement of .489 (see E-G) when subject and gold standard agreement is .522 (see H-G); agreement of the ensemble with the subjects is .468 when the upper bound is .490 (see E-H and H-H, respectively).

8. General Discussion

In this paper we proposed a data intensive approach for inferring the temporal relations in text. We introduced models that learn temporal relations from sentences where temporal information is made explicit via temporal markers. These models could potentially be used in cases where overt temporal markers are absent. We also evaluated our models against a sentence fusion task. The latter is relevant for applications such as summarisation or question answering where sentence fragments (extracted from potentially multiple documents) must be combined into a fluent sentence. For

the fusion task our models determine the appropriate ordering among a temporal marker and two clauses.

Previous work on temporal inference has focused on the automatic tagging of temporal expressions (e.g., Wilson et al., 2001) or on learning the ordering of events from manually annotated data (e.g., Mani et al., 2003, Boguraev & Ando, 2005). Our models bypass the need for manual annotation by focusing 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 utilising 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 real world knowledge (e.g., Hobbs et al., 1993; Lascarides & Asher, 1993; Kehler, 2002). Our best model achieved an F-score of 69.1% on the interpretation task and 93.4% on the fusion task. This performance is a significant improvement over the baseline and compares favourably with human performance on the same tasks. Our experiments further revealed that not only lexical but also syntactic information is important for both tasks. 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. In sum, we have shown that it is possible to infer temporal information from corpora even if they are not semantically annotated in any way.

An important future direction lies in modelling 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 modelling 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 3 revealed, this is a rather difficult task; even humans cannot easily make decisions regarding temporal relations out-of-context. We plan to take into account contextual (lexical and syntactic) as well as discourse-based features (e.g., coreference resolution). 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).

The approach presented in this paper can be also 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 is 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 Penntreebank project. At each iteration the model is retrained on progressively more accurate and representative data.

Finally, temporal relations and discourse structure are co-dependent (Kamp & Reyle, 1993b; Asher & Lascarides, 2003). It is a matter of future work to devise models that integrate discourse

and temporal relations, with the ultimate goal of performing full-scale text understanding. In fact, the two types of knowledge may be mutually benefitial, thus improving both temporal and discourse text analysis.

Acknowledgments

The authors acknowledge the support of EPSRC (grant numbers GR/R40036/01 and GR/T04540/01). Thanks to Regina Barzilay and Frank Keller for helpful comments and suggestions.

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* (CoNLL).
- Barzilay, R. (2003). *Information Fusion for Multi-Document Summarization: Praphrasing and Generation*. Ph.D. thesis, Columbia University.
- 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*, Denmark, September 2001.
- 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 *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*, 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 these 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 *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.
- 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.
- Kamp, H., & Reyle, U. (1993a). From Discourse to the Lexicon: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer Academic Publishers.
- Kamp, H., & Reyle, U. (1993b). From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer, Dordrecht.
- 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.
- Lapata, M., & Brew, C. (1999). Using subcategorisation to resolve verb class ambiguity. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 1999)*, pp. 266–274.
- 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 Puste-jovsky, 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, Edmonton, Canada.
- Marcu, D. (1999). A formal and computational synthesis of grosz and sidner's and mann and thompson's theories. In *Workshop on the Levels of Representation of Discourse*, pp. 101–107, Edinburgh.
- 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 Pustejovsky, I. M. 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: Smeantic 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 the 16th Irish Conference on Artificial Intelligence and Cognitive Science*.
- van Halteren, H., Zavrel, J., & Daelemans, W. (2001). Improving accuracy in wordclass 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*, p. 81:87, Toulouse, France.
- Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5, 241-259.

Appendix A. Experimental Materials for Human Evaluation

The following is the list of materials used in the human evaluation studies reported in Experiment 3 (Section 7). 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 places 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
	<i>6</i>
11	And a lot of them want to get out they get kicked out. before
11 12	·
	And a lot of them want to get out they get kicked out. before
	And a lot of them want to get out they get kicked out. before prices started falling, the market was doing \$1.5 billion a week in new issues, says the head of
12	And a lot of them want to get out they get kicked out 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
12	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers.
12 13	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before
12 13	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf
12 13 14	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before
12 13 14 15	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before Are more certified deaths required the FDA acts? before
12 13 14 15	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before Are more certified deaths required the FDA acts? before Currently, a large store can be built only smaller merchants in the area approve it, a difficult and
12 13 14 15 16	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets. before Are more certified deaths required the FDA acts? before Currently, a large store can be built only smaller merchants in the area approve it, a difficult and time consuming process. after
12 13 14 15 16 17	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before Are more certified deaths required the FDA acts? before Currently, a large store can be built only smaller merchants in the area approve it, a difficult and time consuming process after The review began last week Robert L. Starer was named president after
12 13 14 15 16 17	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before Are more certified deaths required the FDA acts? before Currently, a large store can be built only smaller merchants in the area approve it, a difficult and time consuming process after The review began last week Robert L. Starer was named president after The lower rate came the nation's central bank, the Bank of Canada, cut its weekly bank rate to
12 13 14 15 16 17 18	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before Are more certified deaths required the FDA acts? before Currently, a large store can be built only smaller merchants in the area approve it, a difficult and time consuming process after The review began last week Robert L. Starer was named president after 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
12 13 14 15 16 17 18	And a lot of them want to get out they get kicked out 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 But you start feeling sorry for the fair sex, note that these are the Bundys, not the Bunkers. before The Organization of Petroleum Exporting Countries will travel a rocky road its Persian Gulf members again rule world oil markets before Are more certified deaths required the FDA acts? before Currently, a large store can be built only smaller merchants in the area approve it, a difficult and time consuming process after The review began last week Robert L. Starer was named president after 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 Black residents of Washington's low-income Anacostia section forced a three-month closing of a

Table 17: Materials for the interpretation task; markers in bodlface indicate the goldstandard completion; subjects were asked to select the missing word from a set of eight temporal markers.

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	And 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.
31	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
30	keeps trying to flick non-existent ashes into an ashtray.
37	Britain's airports were disrupted a 24-hour strike by air traffic control assistants resulted in the
37	cancellation of more thank 500 flights and lengthy delays for travelers.
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.
40	Polish strikers shut Gdansk's port Warsaw rushed riot police to the city. as
10	Total state of the Country of the Transfer Tubled Hot police to the city.

Table 17: (continued)

1 When

it turned nearly sideways and bounced again,

I broke into a cold sweat.

2. When

you get into one of these types of periods,

it can go on for a while.

3 When

two apples touch one another at a single point of decay, the mould spreads over both of them.

4 Republicans get very nervous

when

other Republicans put deals together with the Russians.

5 He sounded less than enthusiastic

when

he announced his decision to remain and lead the movement.

6 Democrats are sure to feast on the idea of Republicans cutting corporate taxes

while

taking a bot out of the working man's pension.

7 While

the representative of one separatist organisation says it has suspended its bombing activities, Colimbo authorities recently found two bombs near a government office.

8 Under Chapter 11, a company continues to operate with protection from creditor' lawsuits

while

it works out a plan to pay debt.

9 Investors in most markets sat out

while

awaiting the U.S. trade figures.

10 The top story received 374 points,

while

the 10th got 77.

11 The dollar firmed in quiet foreign exchange trading

after

the U.S report on consumer prices showed no sign of a rumored surge of inflation last month.

12 The strike, which lasted six days, was called by a group of nine rail unions

after

contract negotiations became deadlocked over job security and other issues.

13 The results were announced

after

the market closed.

Marines and sailors captured five Korean forts

after

a surveying party was attacked.

15 Tariffs on 3,500 kinds of imports were lowered yesterday by an average 50%

after

the cuts received final approval on Saturday from President Lee Teng-hui.

Table 18: Materials for the fusion task displaying the goldstandard order of temporal markers, main and subordinate clauses; subjects were presented with the three fragments in random order and asked to create a well-formed sentence.

Soviet diplomats have been dropping hints all over the world that Moscow wants a deal **before**the Reagan administration ends.

17 **Before**

credit card interest rates are reduced across-the-board you will see North buying a subscription to Pravda.

18 Leonard Shane, 65 years old, held the post of president

before

William Shane, 37, was elected to it last year.

19 The protests came exactly a year

before

the Olympic Games are to begin in Seoul.

20 This matter also must be decided by the regulators

before

the Herald takeover can be resolved.

The exact amount of the loss will not be known

unti

a review of the company's mortgage portfolio is completed.

22 A piece of sheet metal was prepared for installation over the broken section of floor

unti

the plane came out of service for a scheduled maintenace

23 The defective dresses are held

until

the hems can be fixed

24 It buys time

until

the main problem can be identified and repaired.

The last thing cut off was the water, for about a week

until

he came up with some money.

26 Once

the treaty is completed,

both Mr. Reagan and Mr. Corbachev probably will want to take credit for it.

27 The borrower is off the hook

once

a bank accepts such drafts and tries to redeem them.

28 Once

the state controls all credit,

a large degree of private freedom is lost.

29 Skeptics doubt BMW can maintain its highfluing position

once

the Japanese join the fray.

30 Once

that notice is withdrawn,

the companies wouldn't be in a position to call in their bonds.

Table 18: (continued)

31 Mr. Bush watched Discovery land congratulated the astronauts **as** they stepped out of the spaceship.

32 Most other papers wound up lower

as

some investors took profits on Tuesday's sharp gains.

33 The announcement comes

as

Congress is completing action on its spending bills for fiscal 1989.

34 Stock prices took a beating yesterday

ลร

trading strategies related to stock-index futures caused widespread selling of the underlying stocks.

35 Grand melodies poured out of him,

96

he contemplated Caesar's conquest of Egypt.

36 Morale in the corporate-finance department has suffered

since

the Union Bank talks broke down.

37 Japanese auto exports to the U.S. almost certainly fell of their annual quiota for the first time since

export controls were inposed in 1981.

38 Soo Line has cut 1,900 jobs

since

it acquired the core assets of the Mulwaukee Road trail line in February 1985.

39 Since

so many parents report the same symptoms,

it occurred to me that these teen-agers must be suffering from the same malady.

40 Foster children have been placed with openly gay parents

since

the new system wen into effect later in 1985.

Table 18: (continued)

Inferring Sentence-internal Temporal Relations

Mirella Lapata

Department of Computer Science University of Sheffield Regent Court, 211 Portobello Street Sheffield S1 4DP, UK

mlap@dcs.shef.ac.uk

Alex Lascarides

School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
alex@inf.ed.ac.uk

Abstract

In this paper we propose a data intensive approach for inferring sentence-internal temporal relations, which relies on a simple probabilistic model and assumes no manual coding. We explore various combinations of features, and evaluate performance against a gold-standard corpus and human subjects performing the same task. The best model achieves 70.7% accuracy in inferring the temporal relation between two clauses and 97.4% accuracy in ordering them, assuming that the temporal relation is known.

1 Introduction

The ability to identify and analyse temporal information is crucial for a variety of practical NLP applications such as information extraction, question answering, and summarisation. In multidocument summarisation, information must be extracted, potentially fused, and synthesised into a meaningful text. Knowledge about the temporal order of events is important for determining what content should be communicated (*interpretation*) but also for correctly merging and presenting information (*generation*). In question answering one would like to find out when a particular event occurred (e.g., *When did X resign?*) but also to obtain information about how events relate to each other (e.g., *Did X resign before Y?*).

Although temporal relations and their interaction with discourse relations (e.g., Parallel, Result) have received much attention in linguistics (???), the automatic interpretation of events and their temporal relations is beyond the capabilities of current open-domain NLP systems. While corpus-based methods have accelerated progress in other areas of NLP, they have yet to make a substantial impact on the processing of temporal information. This is partly due to the absence of readily available corpora annotated with temporal information, although

efforts are underway to develop treebanks marked with temporal relations (?) and devise annotation schemes that are suitable for coding temporal relations (??). Absolute temporal information has received some attention (???) and systems have been developed for identifying and assigning referents to time expressions.

Although the treatment of time expressions is an important first step towards the automatic handling of temporal phenomena, much temporal information is not absolute but relative and not overtly expressed but implicit. Consider the examples in (1) taken from ?. Native speakers can infer that John first met and then kissed the girl and that he first left the party and then walked home, even though there are no overt markers signalling the temporal order of the described events.

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

In this paper we describe a data intensive approach that automatically captures information pertaining to the temporal order and relations of events like the ones illustrated in (1). Of course trying to acquire temporal information from a corpus that is not annotated with temporal relations, tense, or aspect seems rather futile. However, sometimes there are overt markers for temporal relations, the conjunctions before, after, while, and when being the most obvious, that make relational information about events explicit:

- (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.
 - 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 order in which events occurred when there are no obvious markers signalling temporal ordering. We will assess the feasibility of such an approach by initially focusing on sentence-internal temporal relations. 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 learned by gathering informative features from the two clauses. This framework can then be used for interpretation in cases where overt temporal markers are absent (see the examples in (1)).

Practical NLP applications such as text summarisation and question answering place increasing demands not only on the analysis but also on the generation of temporal relations. For instance, non-extractive summarisers that generate sentences by fusing together sentence fragments (e.g., Barzilay 2003) must be able to determine whether or not to include an overt temporal marker in the generated text, where the marker should be placed, and what lexical item should be used. We assess how appropriate our approach is when faced with the information fusion task of determining the appropriate ordering among a temporal marker and two clauses. We infer probabilistically which of the two clauses is introduced by the marker, and effectively learn to distinguish between main and subordinate clauses.

2 The Model

Given a main clause and a subordinate clause attached to it, our task is to infer the temporal marker linking the two clauses. Formally, $P(S_M, t_j, S_S)$ 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 maximises $P(S_M, t_j, S_S)$:

(3)
$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(S_M, t_j, S_S)$$
$$= \underset{t_j \in T}{\operatorname{argmax}} P(S_M) P(t_j | S_M) P(S_S | S_M, t_j)$$

We ignore the term $P(S_M)$ in (3) as it is a constant and use Bayes' Rule to derive $P(S_M|t_i)$ from $P(t_i|S_M)$:

(4)
$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(t_j | S_M) P(S_S | S_M, t_j)$$
$$= \underset{t_j \in T}{\operatorname{argmax}} P(t_j) P(S_M | t_j) P(S_S | S_M, t_j)$$

We will further assume that the likelihood of the subordinate clause S_S is conditionally independent of the main clause S_M (i.e., $P(S_S|S_M,t_j)\approx P(S_S|t_j)$). 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.

(5)
$$t^* \approx \underset{t_j \in T}{\operatorname{argmax}} P(t_j) P(S_M | t_j) P(S_S | t_j)$$

 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 3.2). By making the simplifying assumption that these features are conditionally independent given the temporal marker, the probability of observing the conjunctions $a_{(M,1)} \cdots a_{(M,n)}$ and $a_{(S,1)} \cdots a_{(S,n)}$ is:

(6)
$$t^* = \underset{t_j \in T}{\operatorname{argmax}} P(t_j) \prod_i \left(P(a_{\langle M, i \rangle} | t_j) P(a_{\langle S, i \rangle} | t_j) \right)$$

We effectively treat the temporal interpretation problem as a disambiguation task. From the (confusion) set T of temporal markers {after, before, while, when, as, once, until, since}, we select the one that maximises (6). We compiled a list of temporal markers from ?. Markers with corpus frequency less than 10 per million were excluded from our confusion set (see Section 3.1 for a description of our corpus).

The model in (6) is simplistic in that the *relation-ships* between the features across the clauses are not captured 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. Also note that some of these markers are ambiguous with respect to their meaning: one sense of *while* denotes overlap, another contrast; *since* can indicate a sequence of events in which the main clause occurs after the subordinate clause or cause, *as* indicates overlap or cause, and *when* can denote overlap, a sequence of events, or contrast. Our model selects the appropriate markers on the basis of distributional evidence while being agnostic to their specific meaning when they are ambiguous.

For the sentence fusion task, the identity of the two clauses is unknown, and our task is to infer which clause contains the marker. This can be expressed as:

(7)
$$p^* = \underset{p \in \{M,S\}}{\operatorname{argmax}} P(t) \prod_{i} \left(P(a_{\langle p,i \rangle} | t) P(a_{\langle \overline{p},i \rangle} | t) \right)$$

where p is generally speaking a sentence fragment to be realised as a main or subordinate clause ($\{\overline{p} = S | p = M\}$ or $\{\overline{p} = M | p = S\}$), and t is the temporal marker linking the two clauses.

We can estimate the parameters for the models in (6) and (7) from a parsed corpus. We first identify clauses in a hypotactic relation, i.e., main clauses of which the subordinate clause is a constituent. Next, in the training phase, we estimate the probabilities $P(a_{\langle M,i\rangle}|t_j)$ and $P(a_{\langle S,i\rangle}|t_j)$ by simply counting the occurrence of the features $a_{\langle M,i\rangle}$ and $a_{\langle S,i\rangle}$ with marker t. For features with zero counts, we use add-k smoothing (?), where k is a small number less than one. In the testing phase, all occurrences of the relevant temporal markers are removed for the interpretation task and the model must decide which member of the confusion set to choose. For the sentence fusion task, it is the temporal order of the two clauses that is unknown

and must be inferred. A similar approach has been advocated for the interpretation of discourse relations by ?. They train a set of naive Bayes classifiers on a large corpus (in the order of 40 M sentences) representative of four rhetorical relations using word bigrams as features. The discourse relations are read off from explicit discourse markers thus avoiding time consuming hand coding. Apart from the fact that we present an alternative model, our work differs from ? in two important ways. First we explore the contribution of linguistic information to the inference task using considerably smaller data sets and secondly apply the proposed model to a generation task, namely information fusion.

3 Parameter Estimation

3.1 Data Extraction

Subordinate clauses (and their main clause counterparts) were extracted from the BLLIP corpus (30 M words), a Treebank-style, machine-parsed version of the Wall Street Journal (WSJ, years 1987–89) which was produced using ?'s (?) parser. From the extracted clauses we estimate the features described in Section 3.2.

We first traverse the tree top-down until we identify the tree node bearing the subordinate clause label we are interested in and 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. 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). 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. Note also 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. We are relying on the parser for providing relatively accurate information about attachment sites, but unavoidably there is some noise in the data.

3.2 Model Features

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

```
(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

Temporal Signature (T) It is well known that verbal tense and aspect impose constraints on the temporal order of events but also on the choice of temporal markers. These constraints are perhaps best illustrated in the system of ? 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.

Although we can't 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 infers the aspectual class 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 1. 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 sequencies 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, future, imperfective, active, affirmative}, whereas the latter is mapped to {present, 0, imperfective, passive, affirmative}, where 0 indicates the absence of a modal. In Table 2 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

FINITE	=	{past, present}
NON-FINITE	=	{infinitive, ing-form, en-form}
MODALITY	=	$\{\emptyset$, future, ability, possibility, obligation $\}$
ASPECT	=	{imperfective, perfective, progressive}
VOICE	=	{active, passive}
NEGATION	=	{affimative, negative}

Table 1: Temporal signatures

Feature	$once_{M}$	onces	$since_{M}$	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 2: Relative frequency counts for temporal features

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.

Verb Identity (V) Investigations into the interpretation of narrative discourse have shown that specific lexical information plays an important role in determing temporal interpretation (e.g., Asher and Lascarides 2003). For example, the fact that verbs like *push* can cause movement of the patient and verbs like *fall* describe the movement of their subject can be used to predict that the discourse (8) is interpreted as the pushing causing the falling, making the linear order of the events mismatch their temporal order.

(8) Max fell. John pushed him.

We operationalise lexical relationships among verbs in our data by counting their occurrence in main and subordinate clauses from a lemmatised 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 explictly represented in our temporal signatures. Table 3 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 in Table 3).

Verb Class (V_W, **V**_L) The verb identity feature does not capture meaning regularities concerning the types of verbs entering in temporal relations. For example, in Table 3 *sell* and *pay* are possession verbs, *say* and *announce* are communication verbs, and *come* and *rise* are motion verbs. We use a semantic classification for obtaining some degree of generalisation over the extracted verb occurrences. We experimented with WordNet (?) and the verb classification proposed by ?.

TMark	Verb _M	Verb _S	Noun _N	Nouns	Adj_M	Adj _S
after	sell	leave	year	company	last	new
as	come	acquire	market	dollar	recent	previous
before	say	announce	time	year	long	new
once	become	complete	stock	place	more	new
since	rise	expect	company	month	first	last
until	protect	pay	president	year	new	next
when	make	sell	year	year	last	last
while	wait	complete	chairman	plan	first	other

Table 3: Verb, noun, and adjective occurrences in main and subordinate clauses

Verbs in WordNet are classified in 15 general semantic domains (e.g., verbs of change, verbs of cognition, etc.). We mapped the verbs occurring in main and subordinate clauses to these very general semantic categories (feature V_W). Ambiguous verbs in WordNet will correspond to more than one semantic class. We resolve ambiguity heuristically by always defaulting to the verb's prime sense and selecting the semantic domain for this sense. In cases where a verb is not listed in WordNet we default to its lemmatised form.

? 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 organised into semantically coherent classes (200 in total). ? 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 in ?. Again, for verbs not included in Levin, the lemmatised verb form is used.

Noun Identity (N) It is not only verbs, but also nouns that can provide important information about the semantic relation between two clauses (see ? for detailed motivation). In our domain for example, the noun *share* is found in main clauses typically preceding the noun *market* which is often found in subordinate clauses. Table 3 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 quite frequent in this data set.

Nouns were extracted from a lemmatised version of the parser's output. 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 organisations (e.g., *United Laboratories Inc.*), person names (e.g., *Jose Y. Campos*), and locations (e.g., *New England*) which were subsequently substituted by the general categories person, organisation, and location.

Noun Class ($N_{\rm W}$). As in the case of verbs, nouns were also represented by broad semantic classes 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 semantic classes (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.

Adjective (A) Our motivation for including adjectives in our feature set is twofold. First, we hypothesise that temporal adjectives will be frequent in subordinate clauses introduced by strictly temporal markers such as before, after, and until and therefore may provide clues for the marker interpretation task. 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 (9)).

(9) 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 3.

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].

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, whether it is modified by a preposition or an adverbial. 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 intitial (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 intitial whereas 90% of the *after* clauses are found in the second position.

In the following sections we describe our experiments with the model introduced in Section 2. We first investigate the model's accuracy on the temporal interpretation and fusion tasks (Experiment 1) and then describe a study with humans (Experiment 2). The latter enables us to examine in more depth the model's classification accuracy when compared to human judges.

4 Experiment 1: Interpretation and Fusion

4.1 Method

The model was trained on main and subordinate clauses extracted from the BLLIP corpus as detailed in Section 3.1. 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.

4.2 Results

In order to assess the impact of our features on the interpretation task, the feature space was exhaustively evaluated on the development set. We have nine features, which results in $\frac{9!}{(9-k)!}$ feature combinations where k is the arity of the combination (unary, binary, ternary, etc.). We measured the accuracy of all feature combinations (1023 in total) on the develoment set. From these, we selected the most informative combinations for evaluating the model on the test set. The best accuracy (61.4%) on the development set was observed with the combination of verbs (V) with syntactic signatures (S). We also observed that some feature combinations performed reasonably well on individual markers, even though their overall accuracy was not better than V and S combined. Some accuracies for these combinations are shown in Table 4. For example, NPRSTV was one of the best combinations for generating after, whereas SV was better for before (feature abbreviations are as introduced in Section 3.2).

Given the complementarity of different model parametrisations, 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 classifiers are sufficiently uncor-

	Interpretat	Fusio	on	
TMark	Feat	Acc	Feat	Acc
after	NPRSTV	69.9	AV_WV	77.9
as	ANN_WPSV	57.0	AV	75.8
before	SV	42.1	ANSTV	85.4
once	PRS	40.7	RT	100
since	PRST	25.1	T	85.2
when	V_LPS	85.5	RST	86.9
while	PST	49.0	V_WS	79.4
until	$V_L V_W RT$	69.4	TV	90.5

Table 4: Best feature combinations for individual markers (development set)

	Interpretation				F	usion
	Е		SV		E	ARSTV
TMark	Prec	Rec	Prec	Rec	Prec	Prec
after	61.5	66.5	51.6	55.2	96.7	75.2
as	61.5	62.6	57.0	52.8	93.2	70.5
before	50.0	51.5	32.0	39.1	96.8	84.1
once	60.0	25.0	12.7	15.0	100	88.3
since	69.4	26.3	25.4	12.0	98.2	81.0
when	83.0	91.1	84.7	85.0	99.3	83.8
while	71.5	28.9	38.0	25.8	97.7	82.8
until	57.8	52.4	38.5	47.7	97.8	87.8
Acc	70).7	62	2.6	97.3	80.1
Baseline	42.6	42.6	42.6	42.6	50.0	50.0

Table 5: Results on interpretation and fusion (test set)

related (?). In this paper an ensemble was constructed by combining classifiers resulting from training different parametrisations of our model on the same data. A decision tree (?) was used for selecting the models with the least overlap and for combining their output.

The decision tree was trained and tested on the development set using 10-fold cross-validation. We experimented with 65 different models; out of these, the best results on the development set were obtained with the combination of 12 models: AN_WNPSV, APSV, ASV, V_WPRS, V_NPS, V_LS, NPRSTV, PRS, PRST, PRSV, PSV, and SV. These models formed the ensemble whose accuracy was next measured on the test set. Note that the features with the most impact on the interpretation task are 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). The argument structure feature (R) seems to have some influence (it is present in five of the 12 combinations), however we suspect that there is some overlap with S. Nouns, adjectives and temporal signatures seem to have less impact on the interpretation task, for the WSJ domain at least. Our results so far point to the importance of the lexicon (represented by V, N, and A) for the marker interpretion task but also indicate that the syntactic complexity of the two clauses is crucial for inferring their semantic relation.

The accuracy of the ensemble (12 feature combinations) was next measured on the unseen test set using

10-fold cross-validation. Table 5 shows precision (Prec) and recall (Rec). For comparison we also report precision and recall for the best individual feature combination on the test set (SV) and the baseline of always selecting when, the most frequent marker in our data set (42.6%). The ensemble (E) classified correctly 70.7% of the instances in the test set, whereas SV obtained an accuracy of 62.6%. The ensemble performs significantly better than SV ($\chi^2 = 102.57$, df = 1, p < .005) and both SV and E perform significantly better than the baseline ($\chi^2 = 671.73$, df = 1, p < .005 and $\chi^2 = 1278.61$, df = 1, p < .005, respectively). The ensemble has difficulty inferring the markers since, once and while (see the recall figures in Table 5). Since is often confused with the semantically similar while. Until is not ambiguous, however it is relatively infrequent in our corpus (6.3% of our data set). We suspect that there is simply not enough data for the model to accurately infer these markers.

For the fusion task we also explored the feature space exhaustively on the development set, after removing the position feature (P). Knowing the linear precedence of the two clauses is highly predictive of their type: 80.8% of the main clauses are sentence initial. However, this type of positional information is typically not known when fragments are synthesised into a meaningful sentence.

The best performing feature combinations on the development set were ARSTV and ANWRSV with an accuracy of 80.4%. Feature combinations with the highest accuracy (on the development set) for individual markers are shown in Table 4. Similarly to the interepretation task, an ensemble of classifiers was built in order to take advantage of the complementarity of different model parameterisations. The decision tree learner was again trained and tested on the development set using 10fold cross-validation. We experimented with 44 different model instantiations; the best results were obtained when the following 20 models were combined: AV_WNRSTV, ANWNSTV, ANWNV, ANWRS, ANV, ARS, ARSTV, ARSV, ARV, AV, VWHS, VWRT, VWTV, NWRST, NWS, N_WST, V_WT, V_WTV, RT, and STV. Not surprisingly V and S are also important for the fusion task. Adjectives (A), nouns (N and N_W) and temporal signatures (T), all seem to play more of a role in the fusion rather than the interpretation task. This is perhaps to be expected given that the differences between main and subordinate clauses are rather subtle (semantically and structurally) and more information is needed to perform the inference.

The ensemble (consisting of the 20 selected models) attained an accuracy of 97.4% on the test. The accuracy of the the best performing model on the test set (ARSTV) was 80.1% (see Table 5). Precision for each individual marker is shown in Table 5 (we omit recall as it is always one). Both the ensemble and ARSTV significantly outperform the simple baseline of 50%, amounting to always guessing main (or subordi-

nate) for both clauses ($\chi^2 = 4848.46$, df = 1, p < .005 and $\chi^2 = 1670.81$, df = 1, p < .005, respectively). The ensemble performed significantly better than ARSTV ($\chi^2 = 1233.63$, df = 1, p < .005).

Although for both tasks the ensemble outperformed the single best model, it is worth noting that the best individual models (ARSTV for fusion and PSTV for interpretation) rely on features that can be simply extracted from the parse trees without recourse to taxonomic information. Removing from the ensembles the feature combinations that rely on corpus external resources (i.e., Levin, WordNet) yields an overall accuracy of 65.0% for the interpretation task and 95.6% for the fusion task.

5 Experiment 2: Human Evaluation

5.1 Method

We further compared our model's performance against human judges by conducting two separate studies, one for the interpretation and one for the fusion task. In the first study, 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.

In the second study, participants were presented with a series of sentence fragments and were asked to arrange them so that a coherent sentence can be formed. The fragments were a main clause, a subordinate clause and a marker. Participants saw 40 such triples randomly selected from our test set. The set of items was different from those used in the interpretation task; again five items were selected for each marker.

Both studies were conducted remotely over the Internet. Subjects first saw a set of instructions that explained the task, and had to fill in a short questionnaire including basic demographic information. For the interpretation task, a random order of main-subordinate pairs and a random order of markers per pair was generated for each subject. For the fusion task, a random order of items and a random order of fragments per item was generated for each subject. The interpretation study was completed by 198 volunteers, all native speakers of English. 100 volunteers participated in the fusion study, again all native speakers of English. Subjects were recruited via postings to local Email lists.

5.2 Results

Our results are summarised in Table 6. We measured how well subjects agree with the gold-standard (i.e., the corpus from which the experimental items were selected) and how well they agree with each other. We also show how well the ensembles from Section 4 agree with the humans and the gold-standard. We measured agreement using the Kappa coefficient (?) but also report percentage agreement to facilitate comparison with our model. In all cases

	Interp	retation	Fusion		
	K %		K	%	
H-H	.410	45.0	.490	70.0	
H-G	.421	46.9	.522	79.2	
E-H	.390	44.3	.468	70.0	
E-G	.413	47.5	.489	75.0	

Table 6: Agreement figures for subjects and ensemble (inter-subject agreement is shown in boldface)

we compute pairwise agreements and report the mean. In Table 6, H refers to the subjects, G to the gold-standard, and E to the ensemble.

As shown in Table 6 there is less agreement among humans for the interpretation task than the sentence fusion task. This is expected given that some of the markers are semantically similar and in some cases more than one marker are compatible with the meaning of the two clauses. Also note that neither the model nor the subjects have access to the context surrounding the sentence whose marker must be inferred (we discuss this further in Section 6). Additional analysis of the interpretation data revealed that the majority of disagreements arose for as and once clauses. Once was also problematic for our model (see the Recall in Table 5). Only 33% of the subjects agreed with the gold-standard for as clauses; 35% of the subjects agreed with the gold-standard for once clauses. For the other markers, the subject agreement with the gold-standard was around 55%. The highest agreement was observed for since and until (63% and 65% respectively).

The ensemble's agreement with the gold-standard approximates human performance on the interpretation task (.413 for E-G vs. .421 for H-G). The agreement of the ensemble with the subjects is also close to the upper bound, i.e., inter-subject agreement (see, E-H and H-H in Table 6). A similar pattern emerges for the fusion task: comparison between the ensemble and the gold-standard yields an agreement of .489 (see E-G) when subject and gold-standard agreement is .522 (see H-G); agreement of the ensemble with the subjects is .468 when the upper bound is .490 (see E-H and H-H, respectively).

6 Discussion

In this paper we proposed a data intensive approach for inferring the temporal relations of events. We introduced a model that learns temporal relations from sentences where temporal information is made explicit via temporal markers. This model then can be used in cases where overt temporal markers are absent. We also evaluated our model against a sentence fusion task. The latter is relevant for applications such as summarisation or question answering where sentence fragments must be combined into a fluent sentence. For the fusion task our model determines the appropriate ordering among a temporal marker

and two clauses.

We experimented with a variety of linguistically motivated features and have shown that it is possible to extract semantic information from corpora even if they are not semantically annotated in any way. We achieved an accuracy of 70.7% on the interpretation task and 97.4% on the fusion task. This performance is a significant improvement over the baseline and compares favourably with human performance on the same tasks. Previous work on temporal inference has focused on the automatic tagging of temporal expressions (e.g., ?) or on learning the ordering of events from manually annotated data (e.g., ?). Our experiments further revealed that not only lexical but also syntactic information is important for both tasks. This result is in agreement with? who find that syntax trees encode sufficient information to enable accurate derivation of discourse relations.

An important future direction lies in modelling the temporal relations of events across sentences. The approach presented in this paper can be used to support the "annotate automatically, correct manually" methodology used to provide high volume annotation in the Penntreebank project. An important question for further investigation is the contribution of linguistic and extra-sentential information to modelling temporal relations. Our model can be easily extended to include contextual features and also richer temporal information such as tagged time expressions (see ?). Apart from taking more features into account, in the future we plan to experiment with models where main and subordinate clauses are not assumed to be conditionally independent and investigate the influence of larger data sets on prediction accuracy.

Acknowledgments

The authors are supported by EPSRC grant number GR/R40036. Thanks to Regina Barzilay and Frank Keller for helpful comments and suggestions.

÷

1