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 (Kamp and Reyle, 1993; Webber, 1991; Asher and Lascarides, 2003), 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 (Katz and Arosio, 2001) and devise annotation schemes that are suitable for coding temporal relations (Ferro et al., 2000; Setzer and Gaizauskas, 2001). Absolute temporal information has received some attention (Wilson et al., 2001; Schilder and Habel, 2001; Wiebe et al., 1998) 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 Katz and Arosio (2001). 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.
c. 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.
c. Investors in most markets sat out while awaiting the U.S. trade figures.

It is precisely this type of data that we will exploit for

making predictions about the 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 $P(S_M, t_j, S_S)$:

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

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

$$(4) \quad \begin{aligned} t^* &= \operatorname*{argmax}_{t_j \in T} P(t_j|S_M) P(S_S|S_M, t_j) \\ &= \operatorname*{argmax}_{t_j \in T} P(t_j) P(S_M|t_j) P(S_S|S_M, t_j) \end{aligned}$$

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.

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

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

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

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 Quirk et al. (1985). 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 *relationships* 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:

$$(7) \quad p^* = \operatorname*{argmax}_{p \in \{M,S\}} P(t) \prod_i \Big(P(a_{\langle p,i\rangle}|t) P(a_{\langle \overline{p},i\rangle}|t) \Big)$$

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 (Johnson, 1932), 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 mem-

ber 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 Marcu and Echihabi (2002). 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 Marcu and Echihabi (2002) 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 Charniak's (2000) 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 (Asher and Lascarides, 2003). By selecting features that represent, albeit indirectly and imperfectly,

```
(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

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.

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 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 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, ∅, imperfective, passive, affirmative}, where ∅ 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*.

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

Table 1: Temporal signatures

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

Table 2: Relative frequency counts for temporal features

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 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

TMark	$Verb_M$	$Verb_S$	$Noun_N$	$Noun_S$	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

degree of generalisation over the extracted verb occurrences. We experimented with WordNet (Fellbaum, 1998) and the verb classification proposed by Levin (1993).

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.

Levin (1993) 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). 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.

Noun Identity (N) It is not only verbs, but also nouns that can provide important information about the semantic relation between two clauses (see Asher and Lascarides 2003 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_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 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 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 de-

TMark	Interpretation Feat	Acc	Fusion 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_LV_WRT	69.4	TV	90.5

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

	Interpretation				Fusion	
	E		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		62.6		97.3	80.1
Baseline	42.6	42.6	42.6	42.6	50.0	50.0

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

cisions 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). 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 (Quinlan, 1993) 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 AN_WRSV 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 10-fold cross-validation. We experimented with 44 different model instantiations; the best results were obtained when the following 20 models were combined: AV_WNRSTV, AN_WNSTV, AN_WNV, AN_WRS, ANV, ARS, ARSTV, ARSV, ARV, AV, V_WHS, V_WRT, V_WTV, N_WRST, N_WS, 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 subordinate) 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 (Siegel and Castellan, 1988) but also report percentage agreement to facilitate comparison with our model. In all cases 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.

	Interpretation		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)

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 rele-

vant 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., Wilson et al. 2001) or on learning the ordering of events from manually annotated data (e.g., Mani et al. 2003). 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.

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 Mani et al. 2003). 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.

References

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Barzilay, Regina. 2003. *Information Fusion for Multi-Document Summarization: Praphrasing and Generation*. Ph.D. thesis, Columbia University.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*. Seattle, WA, pages 132–139.

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

Dorr, Bonnie and Terry Gaasterland. 1995. Selecting tense aspect and connecting words in language generation. In *Proceedings of 14th International Joint Conference on Artificial Intelligence*. Montréal, Canada, pages 1299–1307.

Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Ferro, Lisa, Inderjeet Mani, Beth Sundheim, and George Wilson. 2000. Tides temporal annotation guidelines. Technical report, The MITRE Corporation.

Johnson, W. E. 1932. Probability: The deductive and inductive problems. *Mind* 49:409–423.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Katz, Graham and Fabrizio Arosio. 2001. The annotation of temporal information in natural language sentences. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*. Toulouse, France, pages 104–111.

Lapata, Maria and Chris Brew. 1999. Using subcategorization to resolve verb class ambiguity. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. College Park, MD, pages 266–274.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Mani, Inderjeet, Barry Schiffman, and Jianping Zhang. 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, Daniel and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA, pages 368–375.

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

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

Schilder, Frank and Christopher Habel. 2001. From temporal expressions to temporal information: Semantic tagging of news messages. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*. Toulouse, pages 65–72.

Setzer, Andrea and Robert Gaizauskas. 2001. A pilot study on annotating temporal relations in text. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*. Toulouse, France, pages 73–80.

Siegel, Sidney and N. John Castellan. 1988. *Non Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.

Soricut, Radu and Daniel Marcu. 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*. Edmonton, Canada, pages 228–235.

Webber, Bonnie Lynn. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes* 6(2):107–135.

Wiebe, Janyce M., Thomas P. O'Hara, Thorsten Öhrström Sandgren, and Kenneth J. McKeever. 1998. An empirical approach to temporal reference resolution. *Journal of Artifical Intelligence Research* 9:247–293.

Wilson, George, Interjeet Mani, Beth Sundheim, and Lisa Ferro. 2001. A multilingual approach to annotating and extracting temporal information. In *Proceedings of ACL Workshop on Temporal and Spatial Information Processing*. Toulouse, France, pages 81–87.