Comparison of different algebras for inducing the temporal structure of texts

Pascal Denis[†]
† Alpage Project-Team
INRIA & Université Paris 7
pascal.denis@inria.fr

Philippe Muller[†,◇]
◇ IRIT
Université de Toulouse
muller@irit.fr

Abstract

This paper investigates the impact of using different temporal algebras for learning temporal relations between events. Specifically, we compare three interval-based algebras: Allen (1983) algebra, Bruce (1972) algebra, and the algebra derived from the TempEval-07 campaign. These algebras encode different granularities of relations and have different inferential properties. They in turn behave differently when used to enforce global consistency constraints on the building of a temporal representation. Through various experiments on the TimeBank/AQUAINT corpus, we show that although the TempEval relation set leads to the best classification accuracy performance, it is too vague to be used for enforcing consistency. By contrast, the other two relation sets are similarly harder to learn, but more useful when global consistency is important. Overall, the Bruce algebra is shown to give the best compromise between learnability and expressive power.

1 Introduction

Being able to recover the temporal relations (e.g., precedence, inclusion) that hold between events and other time-denoting expressions in a document is an essential part of natural language understanding. Success in this task has important implications for other NLP applications, such as text summarization, information extraction, and question answering.

Interest for this problem within the NLP community is not new (Passonneau, 1988; Webber, 1988; Lascarides and Asher, 1993), but has been recently revived by the creation of the TimeBank corpus (Pustejovsky et al., 2003), and the organization of the TempEval-07 campaign (Verhagen et al., 2007). These have seen the development of machine learning inspired systems (Bramsen et al., 2006; Mani et al., 2006; Tatu and Srikanth, 2008; Chambers and Jurafsky, 2008).

Learning the temporal stucture from texts is a difficult problem because there are numerous information sources at play (in particular, semantic and pragmatic ones) (Lascarides and Asher, 1993). An additional difficulty comes from the fact that temporal relations have logical properties that restrict the consistent graphs that can be built for a set of temporal entities (for instance the transitivity of inclusion and temporal precedence). Previous work do not attempt to directly predict globally coherent temporal graphs, but instead focus on the the simpler problem of labeling pre-selected pairs of events (i.e., a task that directly lends itself to the use of standard classification techniques). That is, they do not consider the problem of *linking* pairs of events (i.e., of determining which pairs of events are related).

Given the importance of temporal reasoning for determining the temporal structure of texts, a natural question is how to best use it within a machine-based learning approach. Following (Mani et al., 2006), prior approaches exploit temporal inferences to enrich the set of training instances used for learning. By contrast, (Bramsen et al., 2006) use temporal relation compositions to provide constraints in a global inference problem (on the slightly different task of ordering passages in medical history records). (Tatu and Srikanth, 2008) and (Chambers and Jurafsky, 2008) combine both approaches and use temporal reasoning both during training and decoding. Interestingly, these approaches use different inventories of relations: (Mani et al., 2006) use the TimeML 13 relation set, while (Chambers and Jurafsky, 2008;

Bramsen et al., 2006) use subset of these relations, namely precedence and the absence of relation.

This paper adopts a more systematic perspective and directly assesses the impact of different relation sets (and their underlying algebras) in terms of learning and inferential properties. Specifically, we compare three interval-based algebras for building classification-based systems, namely: Allen (1983)'s 13 relation algebra, Bruce (1972)'s 7 relations algebra, and the algebra underlying Tempeval-07 3 relations (henceforth, TempEval algebra). We wish to determine the best trade-off between: (i) how easy it is to learn a given set of relations, (ii) how informative are the representations produced by each relation set, and (iii) how much information can be drawn from the predicted relations using knowledge encoded in the representation. These algebras indeed differ in the number of relations they encode, and in turn in how expressive each of these relations is. From a machine learning point of view of learning, it is arguably easier to learn a model that has to decide among fewer relations (i.e., that has fewer classes). But from a representational point of view, it is better to predict relations that are as specific as possible, for composing them may restrict the prediction to more accurate descriptions of the situation. However, while specific relations potentially trigger more inferences, they are also more likely to predict inconsistent constraints. In order to evaluate these differences, we design a set of experiments on the Timebank/AQUAINT corpus, wherein we learn precise relations and vaguer ones, and evaluate them with respect to each other (when a correspondence is possible).

Section 2 briefly presents the Timebank/AQUAINT corpus. In section 3, we describe the task of temporal ordering through an example, and discuss how it should be evaluated. Section 4 then goes into more detail about the different representation possibilities for temporal relations, and some of their formal properties. Section 5 presents our methods for building temporal structures, that combines relation classifiers with global constraints on whole documents. Finally, we discuss our experimental results in section 6.

2 The Timebank/AQUAINT corpus

Like (Mani et al., 2006) and (Chambers and Jurafsky, 2008), we use the so-called OTC corpus, a corpus of 259 documents obtained by combining the Timebank corpus (Pustejovsky et al., 2003) (we use version 1.1 of the corpus) and the AQUAINT corpus.[1] The Timebank corpus consists of 186 newswire articles (and around $65,000$ words), while AQUAINT has 73 documents (and around $40,000$ words).

Both corpora are annotated using the TimeML scheme for tagging eventualities (events and states), dates/times, and their temporal relations. Eventualities can be denoted by verbs, nouns, and some specific constructions. The temporal relations (i.e., the so-called TLINKS) encode topological information between the time intervals of occurring eventualities. TimeML distinguishes three types of TLINKS: event-event, event-time, and time-time, giving rise to different subtasks. In this paper, we will focus on predicting event-event relations (see (Filatova and Hovy, 2001; Boguraev and Ando, 2005) for work on the other tasks). The set of temporal relations used in TLINKS mirrors the 13 Allen relations (see next section), and includes the following six relations: *before*, *begins*, *ends*, *ibefore*, *includes*, *simultaneous* and their inverses. The combined OTC corpus comprises a total of $6,139$ annotated event-event TLINKS. We also make use of the additional TLINKS independently provided by (Bethard et al., 2007) for 129 of the 186 Timebank documents.

3 Task presentation and evaluation

3.1 An example

We illustrate the task of event ordering using a small fabricated, simplified example:

> Fortis bank invested$_{e_1}$ in junk bonds before the financial crisis$_{e_2}$, but got rid$_{e_3}$ of most of them during the crisis$_{e_{2bis}}$. However, the institution still went bankrupt$_{e_4}$ a year later.

[1] Both corpora are freely available from `http://www.timeml.org/site/timebank/timebank.html`.

The annotation for this temporal structure would include the following relations: e_1 is temporally before e_2, e_3 is temporally included in e_2, and e_3 is before e_4. The coreference relation between e_2 and e_{2bis} implies the equality of their temporal extension. Of course all these events may in theory be related temporally to almost any other event in the text. Events are also anchored to temporal expressions explicitly, and this is usually considered as a separate, much easier task. We will use this example throughout the rest of our presentation.

3.2 Comparing temporal annotations

Due to possible inferences, there are often many equivalent ways to express the same ordering of events, so comparisons between annotation and reference event-event pairs cannot rely on simple precision/recall measures.

Consider the above example and assume the following annotation: e_1 is before e_2, e_3 is included in e_2, and e_3 is before e_3. Without going into too much detail about the semantics of the relations used, one expects annotators to agree with the fact that it entails that e_1 is before e_3, among other things. So the annotation is equivalent to a larger set of relations. In some cases, the inferred information is disjunctive (the relation holding between two events is a subset of possible "simple" relations, such as "before or included").

Nowadays, the given practice is to compute some sort of transitive closure over the network of constraints on temporal events (usually expressed in the well-studied Allen algebra (Allen, 1983)), and compute agreements over the saturated structures. Specifically, we can compare the sets of *simple* temporal relations that are deduced from it (henceforth, the "strict" metric), or measure the agreement between the whole graphs, including disjunctions (Verhagen et al., 2007) (henceforth, the "relaxed" metric).[2] Under this latter metric, precision (resp. recall) of a prediction for a pair of events consisting of a set S of relations with respect to a set of relations R inferred from the reference, is computed as $|S \cap R|/|S|$ (resp. $|S \cap R|/|R|$).

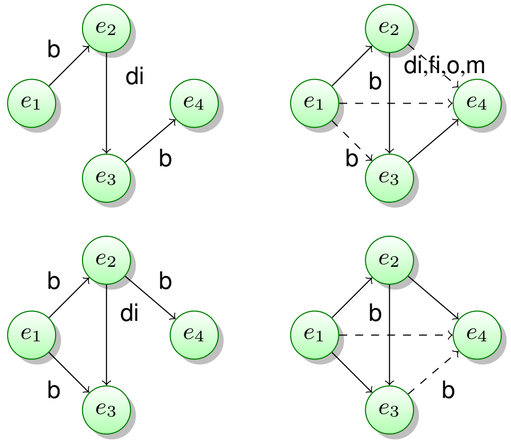


Figure 1: Two non-equivalent annotations of the same situations (left) and their transitive closure in Allen's algebra (right, with new relations only). b stands for Allen's *before* relation, m for *meet*, o for *overlap*, di and fi for the inverses of *during* and *finish*, respectively.

Figure 1 illustrates the point of these "saturated" representations, showing two raw annotations of our example on the left (top and bottom) and their closures on the right. The raw annotations share only 2 relations (between e_1 and e_2, and e_3 and e_4), but their transitive closures agree also on the relations between e_1 and e_3, e_1 and e_4, and e_3 and e_4. They still differ on the relation between e_2 and e_4, but only because one is much more specific than the other, something that can only be taken into account by a partial credit scoring function.

For this example, the "strict" metric yields precision and recall scores of 5/5 and 5/6, when comparing the top annotation against the bottom one. By contrast, the "relaxed" metric (introduced in the TempEval-07) yields precision and recall scores of (5+0.2)/6 and 6/6, respectively.

We now turn to the issue of the set of relations chosen for the task of expressing temporal information in texts.

4 Temporal representations

Because of the inferential properties of temporal relations, we have seen that the same situation can be expressed in different ways, and some relations can be deduced from others. The need for

[2] Taking into account disjunctions means giving partial credit to disjunctions approximating the reference relation (possibly disjunctive itself), see next section.

a precise reasoning framework has been present in previous attempts at the task (Setzer et al., 2006), and people have moved to a set of handmade rules over ad hoc relations to more widely accepted temporal reasoning frameworks, such as algebras of temporal relations, the most famous being Allen's interval algebra.

An algebra of relations can be defined on any set of relations that are mutually exclusive (two relations cannot hold at the same time between two entities) and exhaustive (at least one relation must hold between two given entities). The algebra starts from a set of simple, atomic, relations $U = \{r_1, r_2, ...\}$, and a general relation is a subset of U, interpreted as a disjunction of the relations it contains. From there, we can define union and intersection of relations as classical set union and intersection of the base relations they consist of. Moreover, one can define a composition of relations as follows:

$$(r_1 \circ r_2)(x, z) \leftrightarrow \exists y \; r_1(x, y) \wedge r_2(y, z)$$

In words, a relation between x and z can be computed from what is known between (x and y) and (y and z). By computing beforehand the $n \times n$ compositions of base relations of U, we can compute the composition of any two general relations (because $r \cap r' = \emptyset$ when r, r' are basic and $r \neq r'$):

$$\{r_1, r_2, ...r_k\} \circ \{s_1, s_2, ...s_m\} = \bigcup_{i,j} (r_i \circ s_j)$$

Saturating the graph of temporal constraints means applying these rules to all compatible pairs of constraints in the graph and iterating until a fixpoint is reached. In Allen's algebra there are 13 relations, determined by the different relations that can hold between two intervals endpoints (before, equals, after). These relations are: b (*before*), m (*meet*), o (*overlap*), s (*start*), f (*finish*), d (*during*), their inverses (bi, mi, oi, si, fi, di) and = (*equal*), see figure 2.[3]

It is important to see that a general approach to temporal ordering of events cannot restrict itself to a subset of these and still use the power of inferences to complete a situation, because composition of information is stable only on restricted subsets. And using all of them means generating numerous disjunctions of relations.

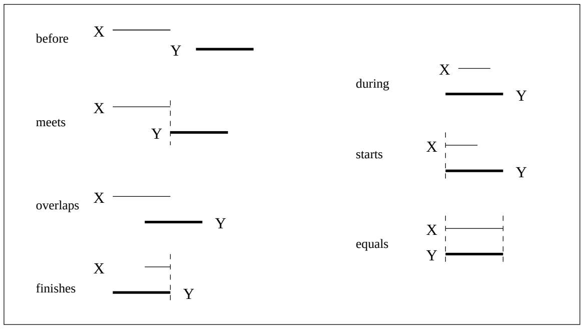


Figure 2: Allen's thirteen relations between two temporal intervals

Allen relations are convenient for reasoning purposes, but might too precise for representing natural language expressions, and that's why recent evaluation campaigns such as TempEval-07 have settled on vaguer representations. TempEval-07 uses three relations called *before*, *overlaps* and *after*, which we note b_t, o_t, and bi_t.[4] These all correspond to disjunctions of Allen relations: $\{\mathsf{b},\mathsf{m}\}_a$, $\{\mathsf{o},\mathsf{d},\mathsf{s},=,\mathsf{f}\}_a$ and its inverse, and $\{\mathsf{bi},\mathsf{mi}\}_a$, respectively. These representations can be converted to Allen relations, over which the same inference procedures can be applied, and then expressed back as (potentially disjunctive) TempEval relations. They thus form a sub-algebra of Allen's algebra, if we add their possible disjunctions.

In fact, starting from the base relations, only $\{\mathsf{b},\mathsf{o}\}_t$, $\{\mathsf{bi},\mathsf{o}\}_t$, and *vague* (i.e., the disjunction of all relations) can be inferred (besides the base relations). This is a consequence of the stability of so-called convex relations in Allen algebra. Note that an even simpler schema is used in (Chambers and Jurafsky, 2008), where only TempEval *before* and *after* and the *vague* relation are used.

We propose to consider yet another set of relation, namely relations from (Bruce, 1972). These provide an intermediate level of representation, since they include 7 simple relations. These are

[3] TimeML uses somewhat different names, with obvious mappings, except *ibefore* ("immediately before") for m, and *iafter* ("immediately after") for mi.

[4] When it is not obvious, we will use subscript symbols to indicate the particular algebra that is used (e.g., b_t is the before relation in TempEval).

also expressible as disjunctions of Allen relations; they are: *before* (b_b), *after* (bi_b) (with the same semantics as TempEval's b_t and bi_t), *equals* ($=_b$, same as $=_a$), *includes* (i, same as Allen's $\{\mathsf{s},\mathsf{d},\mathsf{f}\}_a$), *overlaps* ($\mathsf{o}_b$, same as o_a), *included* (ii) and *is-overlapped* (oi_b), their inverse relations. The equivalences between the three algebras is shown table 1.

Allen	Bruce	Tempeval
before meet	before	before
overlaps	overlaps	overlaps
starts during finishes	included	
overlapsi	is-overlapped	
startsi duringi finishesi	includes	
meeti beforei	after	after
equals	equals	equals

Table 1: Correspondances between temporal algebras. A relation ranging over multiple cells is equivalent to a disjunction of all the relations within these cells.

Considering a vaguer set is arguably more adequate for natural language expressions while at the same time this specific set preserves at least the notions of temporal order and inclusion (contrary to the TempEval scheme), which have strong inferential properties: they are both transitive, and their composition yields simple relations; overlap allows for much weaker inferences. Figure 3 shows part of our example from the introduction expressed in the three cases: with Allen relations, the most precise, with Bruce relations and TempEval relations, with dotted lines showing the extent of the vagueness of the temporal situations in each case (with respect to the most precise Allen description). We can see that TempEval relations lose quickly all information that is not before or after, while Bruce preserves inference combining precedence and temporal inclusion.

Information can be converted from one algebra to the other, since vaguer algebras are based on relations equivalent to disjunctions in Allen algebra. But conversion from a precise relation to a vaguer one and back to a more precise algebra leads to information loss. Hence on figure 3, the original Allen relation: $e_3\ \mathsf{d}_a\ e_2$ is converted to: $e_3\ \mathsf{o}_t\ e_2$ in TempEval, which converts back into the much less informative: $e3\ \{\mathsf{o}, \mathsf{d}, \mathsf{s}, =, \mathsf{f}, \mathsf{oi}, \mathsf{si}, \mathsf{fi}, \mathsf{di}\}_a\ e_2$. We will use these translations during our system evaluation to have a common comparison point between representations.

5 Models

5.1 Algebra-based classifiers

In order to compare the impact of the different algebras described in section 4, we build three event pair classification models corresponding to each relation set. The resulting Allen-based, Bruce-based, and Tempeval-based models therefore contain 13, 7, and 3 class labels, respectively.[5] For obvious sparsity issues, we did not include classes corresponding to disjunctive relations, as there are $2^{|R|}$ possible disjunctions for each relation set R.

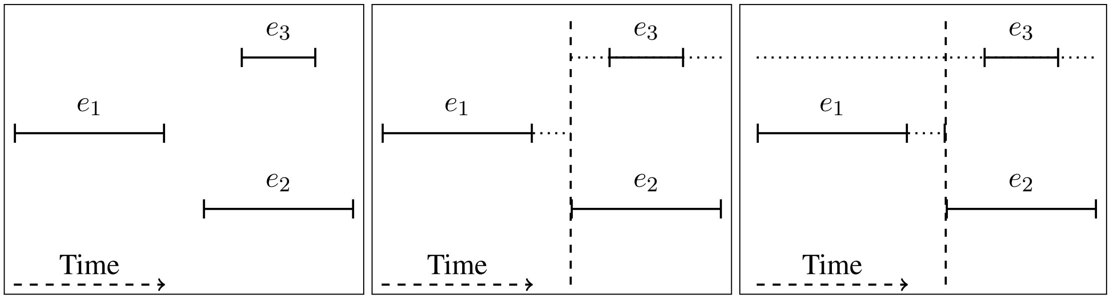
For training our models, we experiment with 4 various configurations that correspond to ways of expanding the set of training examples. Specifically, these configurations vary in: (i) whether or not we added the additional "Bethard relations" to the initial OTC annotations (Bethard et al., 2007), (ii) whether or not we applied saturation over the set of annotated relations.

5.2 Features

Our feature set for the various models is similar to that used by previous work, including binary features that encode event string as well as the five TimeML attributes and their possible values:

- **aspect**: none, prog, perfect, prog perfect
- **class**: report, aspectual, state, I-state I-action, perception, occurrence
- **modality**: none, to, should, would, could can, might
- **polarity**: positive, negative
- **tense**: none, present, past, future

[5] Our TempEval model actually has a fourth label for the *identity* relation. The motivations behind the inclusion of this extra label are: (i) this relation is linguistically motivated and comparatively easy to learn (for a lot of instances of this relation are cases of anaphora, which are often signaled by identical strings) (ii) this relation triggers a lot of specific inferences.



(a) Allen:
$(e_1 \mathsf{b}_a e_2 \wedge e_3 \mathsf{d}_a e_2) \rightarrow e_1 \mathsf{b}_a e_3$

(b) Bruce:
$(e_1 \mathsf{b}_b e_2 \wedge e_3 \mathsf{d}_b e_2) \rightarrow e_1 \mathsf{b}_b e_3$

(c) Tempeval:
$(e_1 \mathsf{b}_t e_2 \wedge e_3 \mathsf{o}_t e_2) \rightarrow e_1 \{\mathsf{b}_t, \mathsf{o}_t\} e_3$

Figure 3: Comparing loss of inferential power in algebras: hard lines show the actual temporal model, exactly expressed in Allen relations (a); dotted lines show the vagueness induced by alternative schemes, and the inference that can or cannot still be made in each algebra, (b) and (c).

Additional binary features check agreement for same attribute (e.g., the same tense). Finally, we add features that represent the distance between two events (in number of sentences, and in number of intervening events). [6]

5.3 Training set generation

Our generic training procedure works as follows. For each document, we scan events in their order of appearance in the text. We create a training instance $inst_{(e_i,e_j)}$ for each *ordered* pair of events (e_i, e_j): if (e_i, e_j) (resp. (e_j, e_i)) corresponds to an annotated relation r, then we label $inst_{(e_i,e_j)}$ with the label r (resp. its inverse r^{-1}).

5.4 Parameter estimation

All of these classifiers are maximum entropy models (Berger et al., 1996). Parameter estimation was performed with the Limited Memory Variable Metric algorithm (Malouf, 2002) implemented in the Megam package.[7]

5.5 Decoding

We consider two different decoding procedures. The first one simply mirrors the training procedure just described, scanning pairs of events in the order of the text, and sending each pair to the classifier. The pair is then labeled with the label outputted by the classifier (i.e., the label receiving the highest probability). No attempt is made to guarantee the consistency of the final temporal graph.

Our second inference procedure works as follows. As in the previous method, we scan the events in the order of the text, and create ordered pairs of events that we then submit to the classifier. But the difference is that we saturate the graph after each classification decision to make sure that the graph created so far is coherent. In case where the classifier predicts a relation whose addition results in an incoherent graph, we try the next highest probability relation, and so on, until we find a coherent graph. This greedy procedure is similar to the Natural Reading Order (NRO) inference procedure described by (Bramsen et al., 2006).

6 Experiments and results

We perform two main series of experiments for comparing our different models. In the first series, we measure the accuracy of the Allen-, Bruce-, and Tempeval-based models on predicting the correct relation for the event-event TLINKS annotated in the corpus. In the second series, we saturate the event pair relations produced by the classifiers (combined with NRO search to enforce global coherence) and compare the predicted graphs against the saturated event-event TLINKS.

6.1 Experiment settings

All our models are trained and tested with 5-fold cross-validation on the OTC documents. For eval-

[6] These were also encoded as binary features, and the various feature values were binned in order to avoid sparseness.

[7] Available from `http://www.cs.utah.edu/~hal/megam/`.

uation, we use simple accuracy for the first series of experiments, and two "strict" and "relaxed" precision/recall measures described in section 3 for the other series. For each type of measures, we report scores with respect to both Allen and TemEval relation sets. All scores are reported using macro-averaging. Out of the 259 temporal graphs present in OTC, we found that 54 of them were actually inconsistent when saturated; the corresponding documents were therefore left out of the evaluation.[8] Given the rather expensive procedure involved in the NRO decoding (saturating an inconsistent graph "erases" all relations), we skipped 8 documents wich were much longer than the rest, leaving us with 197 documents for our final experiments.

6.2 Event-event classification

Table 2 summarizes the accuracy scores of the different classifiers on the event-event TLINKS of OTC. We only report the best configuration for each model. For the TempEval-based model, we found that the best training setting was when Bethard annotations were added to the original TimeML annotations, but with no saturation.[9] For Allen and Bruce models, neither Bethard's relations nor saturation helps improve classification accuracy. In fact, saturation degrades performance, which can be explained by the fact that saturation reinforces the bias towards already over-represented relations.[10] The best accuracy performances are obtained by the Allen-based and TempEval-based classifiers, each one performing better in its own algebra (with 47.0% and 54.0%). This is not surprising, since these classifiers were specifically trained to optimize their respective metrics. The Bruce-based classifier is slightly better than the Allen-based one in TempEval, but also slightly worse than TempEval-based classifier in Allen.

	Allen Acc.	TempEval Acc.
Allen	47.0	48.9
Bruce	N/A	49.3
TempEval	N/A	54.0

Table 2: Accuracy scores for Allen, Bruce, and TempEval classifiers on event-event TLINKS, expressed in Allen or TempEval algebra. Scores for Bruce and TempEval models into Allen are left out, since they predict (through conversion) disjunctive relations for all relations but equality.

Our accuracy scores for Allen, and TempEval-based classifiers are somewhat lower than the ones reported for similar systems by (Mani et al., 2006) and (Chambers and Jurafsky, 2008), respectively. These differences are likely to come from the fact that: (i) (Mani et al., 2006) perform a 6-way classification, and not a 13-way classification[11], and (ii) (Chambers and Jurafsky, 2008) use a relation set that is even more restrictive than TempEval's.

6.3 Saturated graphs

Table 3 summarizes the various precision/recall scores of the graph obtained by saturating the classifiers predictions (potentially altered by NRO) against the event-event saturated graph. These results contrast with the accuracy results presented in table 2: while the TempEval-based model was the best model in classification accuracy in TempEval, it is now outperformed by both the Allen- and Bruce-based systems (this with or with using NRO). The best system in TempEval is actually Bruce-based system, with 52.9 and 62.8 for the strict/relaxed metrics, respectively. The results suggest that this algebra might actually offer the best trade-off between learnanility and expressive power. The use of NRO to restore global coherence yields important gains (10 points) in the relaxed metric for both Allen- and Bruce-based systems (although they do not convert into gains in the strict metric). Unsuprisingly, the best model on the Allen set remains Allen-based model (and this time the use of NRO results in gains on the strict metric). Predictions without

[8] Because there is no way to trace the relation(s) responsible for an inconsistency without analysing the whole set of annotations of a text, and considering that it usually happens on very long texts, we did not attempt to manually correct the annotations.

[9] This is actually consistent with similar findings made by (Chambers and Jurafsky, 2008).

[10] For instance, for Allen relations, there are roughly 50% of *before-after* relations before saturation but 73% of them after saturation.

[11] This is only possible because they order the event-event pairs before submitting them to the classifier.

System	Allen						Tempeval					
	RELAX			STRICT			RELAX			STRICT		
	R	P	F1	R	P	F1	R	P	F1	R	P	F1
Allen	57.5	46.7	**51.5**	49.6	56.2	**52.7**	62.0	50.3	55.5	50.4	57.1	53.6
Bruce	46.0	39.0	42.1	18.0	44.0	25.9	62.9	52.6	**57.3**	50.9	57.0	**53.8**
Tempeval	37.1	35.9	36.5	14.0	44.0	21.2	49.3	47.1	48.2	21.7	44.2	29.1
Allen$_{NRO}$	44.8	60.1	**51.3**	57.2	62.9	**59.9**	63.8	67.0	65.3	45.2	60.6	51.8
Bruce$_{NRO}$	46.3	53.1	49.5	13.9	45.3	21.2	65.5	71.8	**68.5**	46.6	61.1	**52.9**
Tempeval$_{NRO}$	37.1	35.9	36.5	13.9	44.3	21.2	49.3	47.1	48.2	21.7	44.2	29.1

Table 3: Comparing Allen-, Bruce-, Tempeval-based classifiers saturated predictions on saturated event-event graph. The $_{NRO}$ subscript indicates whether the system uses NRO or not. Evaluation are given with respect to both Allen and Tempeval relation sets.

NRO yielded between 7.5 and 9% of inconsistent saturated graphs that were ignored by the evaluation, which means this impacted recall measures only.

7 Related work

Early work on temporal ordering (Passonneau, 1988; Webber, 1988; Lascarides and Asher, 1993) concentrated on studying the knowledge sources at play (such as tense, aspect, lexical semantics, rhetorical relations). The development of annotated resources like the TimeBank corpus (Pustejovsky et al., 2003) has triggered the development of machine learning systems (Mani et al., 2006; Tatu and Srikanth, 2008; Chambers and Jurafsky, 2008).

More recent work uses automatic classification methods, based on the TimeBank and Acquaint corpus, either as is, with inferential enrichment for training (Mani et al., 2006; Chambers et al., 2007), or supplied with the corrections of (Bethard et al., 2007), or are restricted to selected contexts, such as intra-sentential event relations (Li et al., 2004; Lapata and Lascarides, 2006). All of these assume that event pairs are preselected, so the task is only to determine what is the most likely relation between them. The best scores are obtained with the added assumption that the event-event pair can be pre-ordered (thus reducing the number of possible labels by 2).

More recently, (Bramsen et al., 2006) and subsequently (Chambers and Jurafsky, 2008) propose to use an Integer Linear Programming solver to enforce the consistency of a network of constraints while maximizing the score of local classification decisions. But these are restricted to the relations BEFORE and AFTER, which have very strong inference properties that cannot be generalised to other relations. The ILP strategy is not likely to scale up very well for richer relation sets, for the number of possible relations between two events (and thus the number of variables to put in the LP solver for each pair) is the order of $2^{|R|}$ (where R is the relation set), and each transitivity constraints generates an enormous amount of constraints.

8 Conclusion

We have investigated the role played by ontological choices in temporal representations by comparing three algebras with different granularities of relations and inferential powers. Our experiments on the Timebank/AQUAINT reveal that the TempEval relation set provides the best overall classification accuracy, but it provides much less informative temporal structures, and it does not provide enough inferences for being useful for enforcing consistency. By contrast, the other two relation sets are significantly harder to learn, but provide more richer inferences and are therefore more useful when global consistency is important. Bruce's 7 relations-based model appears to perform best in the TempEval evaluation, suggesting that this algebra provides the best trade-off between learnability and expressive power.

References

Allen, James. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, pages 832–843.

Berger, A., S. Della Pietra, and V. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Bethard, Steven, James H. Martin, and Sara Klingenstein. 2007. Timelines from text: Identification of syntactic temporal relations. In *International Conference on Semantic Computing*, pages 11–18, Los Alamitos, CA, USA. IEEE Computer Society.

Boguraev, Branimir and Rie Ando. 2005. TimeML-compliant text analysis for temporal reasoning. In Kaelbling, Leslie Pack and Fausto Giunchiglia, editors, *Proceedings of IJCAI05*, pages 997–1003.

Bramsen, Philip, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198, Sydney, Australia, July. Association for Computational Linguistics.

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

Chambers, Nathanael and Daniel Jurafsky. 2008. Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 698–706, Honolulu, Hawaii, October. Association for Computational Linguistics.

Chambers, Nathanael, Shan Wang, and Daniel Jurafsky. 2007. Classifying temporal relations between events. In *ACL*. The Association for Computer Linguistics.

Filatova, Elena and Eduard Hovy. 2001. Assigning time-stamps to event-clauses. In Mani, I., J. Pustejovsky, and R Gaizauskas, editors, *The Language of Time: A Reader*. Oxford University Press.

Lapata, Maria and Alex Lascarides. 2006. Learning sentence-internal temporal relations. *J. Artif. Intell. Res. (JAIR)*, 27:85–117.

Lascarides, Alex and Nicholas Asher. 1993. Temporal interpretation, discourse relations and common sense entailment. *Linguistics and Philosophy*, 16:437–493.

Li, Wenjie, Kam-Fai Wong, Guihong Cao, and Chunfa Yuan. 2004. Applying machine learning to chinese temporal relation resolution. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 582–588, Barcelona, Spain, July.

Malouf, Robert. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Workshop on Natural Language Learning*, pages 49–55, Taipei, Taiwan.

Mani, Inderjeet, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia, July. Association for Computational Linguistics.

Passonneau, Rebecca J. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2):44–60.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics*, pages 647–656, Lancaster University, UK, March.

Setzer, Andrea, Robert Gaizauskas, and Mark Hepple. 2006. The Role of Inference in the Temporal Annotation and Analysis of Text. *Language Resources and Evaluation*, 39:243–265.

Tatu, Marta and Munirathnam Srikanth. 2008. Experiments with reasoning for temporal relations between events. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 857–864, Manchester, UK, August. Coling 2008 Organizing Committee.

Verhagen, Marc, Robert Gaizauskas, Franck Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 - 15: TempEval Temporal Relation Identification. In *Proceedings of SemEval workshop at ACL 2007*, Prague, Czech Republic, June. Association for Computational Linguistics, Morristown, NJ, USA.

Webber, Bonnie Lynn. 1988. Tense as discourse anaphor. *Computational Linguistics*, 14(2):61–73.