Using Semantic Inference for Temporal Annotation Comparison

Andrea Setzer, Robert Gaizauskas, Mark Hepple

University of Sheffield
Department of Computer Science
{andrea,robertg,hepple}@dcs.shef.ac.uk

Abstract

Temporal information plays an important role in natural language processing applications such as Question Answering, Information Extraction, and Topic Detection and Tracking. The development of these applications can be supported by high quality temporally annotated corpora. In this paper, we introduce a method for comparing temporal annotations and supporting the creation of gold standard annotations, both of which are important for the development of such annotated corpora. The method we propose is based on comparing different annotations in semantic, rather than syntactic terms. We introduce the notion of a temporal closure for temporal annotations as well as formulæ for calculating precision and recall.

1 Introduction

The automatic recognition and annotation of temporal expressions as well as event expressions has become an active area of research in computational linguistics, as evidenced by workshops such as the ACL2001 workshop on Temporal and Spatial Information Processing [4], the LREC 2002 workshop on Annotation Standards for Temporal Information in Natural Language [8], and the TERQAS [1] workshop on Time and Event Recognition for Question Answering Systems [6]. Temporal information is important to many application areas, including Question Answering, Information Extraction (IE), and Topic Detection and Tracking.

To enable the development and evaluation of systems that recognise and annotate temporal information, annotated corpora must be created — and often the goal is to do this automatically. Before this is possible, the annotation scheme must be validated, which is often done by hand-annotating a trial corpus and analysing the results. To overcome the potential inconsistencies to which hand-annotated corpora are prone and to ensure the quality of the

[1] www.time2002.org

description of the scheme used by the annotators during annotation, it is necessary to be able to compare annotations and to assess their 'goodness'. This is frequently done by manually creating a 'gold-standard' annotation against which other manual annotations are compared. These comparisons are used to calculate inter-annotator agreement figures, which are one way of judging how well the scheme is defined and with how much agreement.

In this paper we will present a way of comparing temporal annotations and an approach to support the creation of a gold standard annotation. We introduce the notion of the temporal closure that can be computed over an annotation as well as methods for computing precision and recall figures for different annotations.

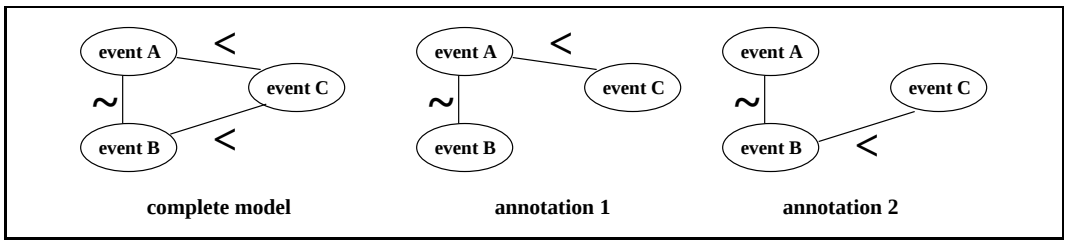

Fig. 1. Comparing annotations

2 An Approach for Comparing Temporal Annotations

Annotation schemes for temporal information usually define the temporal entities, i.e. events and times, that are recognised by the approach and a scheme for annotating them, as well as a scheme for annotating the temporal relations that hold between different temporal entities. (We will not discuss these details here — see [6] and [7] for more information.) The set of temporal relations used by different approaches varies. Allen [1], for example, uses 13 different temporal relations. A similar set of relations is used in the annotation scheme developed during the TERQAS workshop [6]. In this paper, we consider only a small set of uncontroversial relations, to simplify presentation. However, the approach for comparing annotations that we present can as well be used with larger sets of temporal relations. The relations used here are: BEFORE, INCLUDES and SIMULTANEOUS.[2] For convenience, we also assume that events and times are assigned unique identifiers.

2.1 The Equivalence of Temporal Annotations

Temporal annotations, like for example co-reference chain annotations (see section 3), are of a semantic nature and should be compared in semantic

[2] In [7], a larger set is used, which includes also the relations AFTER and IS-INCLUDED. Since these additional relations are the inverses of BEFORE and INCLUDES, their omission does not reduce expressiveness.

terms. Two annotations are equivalent if they convey the same 'temporal information', even if different ways of annotating this information are chosen. For example, consider the first diagram in figure 1, which shows two simultaneous events A and B, which both precede a third event C. The two further diagrams in the figure, representing possible partial annotations, differ from the first but are nonetheless equivalent to it, as the omitted relations are implied by simple inference rules. For Version 1, for example, we can infer the omitted fact that *B is before C* from *A is simultaneous to B* and *A is before C*. Any comparison of annotations should take into account this observation that annotations can be distinct but equivalent.

$\forall x, y, z \in (E \cup T)$:

(1)	$(x,y) \in S$	$\Rightarrow$	$(y,x) \in S$
(2)	$(x,y) \in B \wedge (y,z) \in B$	$\Rightarrow$	$(x,z) \in B$
(3)	$(x,y) \in I \wedge (y,z) \in I$	$\Rightarrow$	$(x,z) \in I$
(4)	$(x,y) \in B \wedge (y,z) \in I$	$\Rightarrow$	$(x,z) \in B$
(5)	$(x,y) \in I \wedge (x,z) \in B$	$\Rightarrow$	$(y,z) \in B$
(6)	$(x,y) \in S \wedge (y,z) \in S$	$\Rightarrow$	$(x,z) \in S$
(7)	$(x,y) \in B \wedge (y,z) \in S$	$\Rightarrow$	$(x,z) \in B$
(8)	$(x,y) \in I \wedge (y,z) \in S$	$\Rightarrow$	$(x,z) \in I$
(9)	$(x,y) \in S \wedge (y,z) \in I$	$\Rightarrow$	$(x,z) \in I$
(10)	$(x,y) \in B \wedge (x,z) \in S$	$\Rightarrow$	$(z,y) \in B$
(11)	$(x,y) \in S \wedge (x,z) \in I$	$\Rightarrow$	$(y,z) \in I$

Fig. 2. Inference Rules

2.2 Calculating the Temporal Closure

The above discussion suggests an approach in which the temporal relations made explicit in a particular annotation of a text might be expanded, using an appropriate set of inference rules, to provide a complete (i.e. maximal) representation of the temporal consequences of that annotation. Such a representation will be termed the *temporal closure* of the annotation. Two annotations of a text can then be compared in terms of the equivalence or overlap of their temporal closures. To formalise this idea, let us allow that identifiers for annotated event and time expressions form two sets, E and T, respectively. Our temporal relations are all binary relations between event and time expressions, and so the denotation of each is a subset of $(E \cup T) \times (E \cup T)$. A set of inference rules for our temporal relations is given in figure 2, using S, B and I to denote the extension of the relations SIMULTANEOUS, BEFORE and INCLUDES, respectively. Some of the inference rules concern only one of the relations, and follow logically from the formal properties of the relation, e.g. that SIMULTANEOUS is an equivalence relation, whilst BEFORE and INCLUDES are

transitive, asymmetric and irreflexive. Other inference rules capture interactions between relations that follow naturally from their intuitive meaning, e.g. if x and y are simultaneous, and x is before z, then y also is before z.

Let S_t denote the simultaneity pairs explicitly specified by a temporally annotated text t, and likewise for B_t and I_t. These components combine to give the overall temporal model of the text $\mathcal{M}_t = \langle S_t, B_t, I_t \rangle$. The inference rules can be applied to this model to generate its deductive closure $\mathcal{M}_t^{\models}$. Let $S_t^{\models}$ denote the SIMULTANEOUS relation that results in $\mathcal{M}_t^{\models}$, and likewise for $B_t^{\models}$ and $I_t^{\models}$. For this approach, we can say that two alternative annotations t and t' of a text are equivalent just in case the deductive closure of their models are equivalent, i.e. $\mathcal{M}_t^{\models} = \mathcal{M}_{t'}^{\models}$. Furthermore, we can say that a model $\mathcal{M}_t$ is a *minimal* model if it has no proper subset which has an equivalent temporal closure. Minimal models need not be unique, as the example in figure 1 shows.

2.3 Recall and Precision

This approach allows comparison between alternative annotations of a text in terms of the degree of overlap between their temporal closures. In particular, a given annotation of a text can be compared to a 'gold standard' annotation of the same text by computing figures of precision and recall between their temporal closures. Let k and r denote key (i.e 'gold standard') and response (i.e. system generated) annotations of the same text. The precision and recall for the SIMULTANEOUS relation S_r as compared to S_k is given by:

$$R = \frac{|S_k^{\models} \cap S_r^{\models}|}{|S_k^{\models}|} \qquad P = \frac{|S_k^{\models} \cap S_r^{\models}|}{|S_r^{\models}|}$$

Parallel definitions can be provided for the other temporal relations. Precision and recall measures for the overall temporal model $\mathcal{M}_t$ can be defined as: [3]

$$R = \frac{|S_k^{\models} \cap S_r^{\models}| + |B_k^{\models} \cap B_r^{\models}| + |I_k^{\models} \cap I_r^{\models}|}{|S_k^{\models}| + |B_k^{\models} + |I_k^{\models}|} \qquad P = \frac{|S_k^{\models} \cap S_r^{\models}| + |B_k^{\models} \cap B_r^{\models}| + |I_k^{\models} \cap I_r^{\models}|}{|S_r^{\models}| + |B_r^{\models} + |I_r^{\models}|}$$

3 Comparison to Co-reference Scoring Approaches

Temporal annotation is not the only case that should be compared in semantic terms. As mentioned earlier, a similar situation arises for comparing differently annotated co-reference chains. We will briefly introduce two co-reference annotation scoring methods, and then compare them to our approach.

3.1 The MUC co-reference scoring scheme

Co-reference is a form of equivalence relation, and co-reference annotations generate equivalence classes. Distinct annotations can generate identical equivalence classes, i.e. we see a situation similar to that discussed for temporal

[3] This method can be compared to Crowe's [3] way of evaluating clause-event grids.

annotation. For example, the co-reference linkages <A-B,B-C> and <A-B,A-C> both generate the equivalence class $\{A, B, C\}$.

The co-reference scoring approach used for MUC6 [5] was developed to handle this fact of distinct annotations being semantically equivalent. The approach exploits the fact that co-reference is an equivalence relation, and provides precision and recall metrics that are computed relative to the minimal size of linkage required to generate the given equivalence classes. An equivalence class with n elements minimally requires $n - 1$ to generate it.

We will introduce the MUC6 precision and recall metrics using a simple example (from [5]). Assume that elements $\{A, B, C, D\}$ are present, of which B-C-D corefer, giving equivalence classes $\{\{A\}, \{B, C, D\}\}$, as captured by a key annotation <B-C, C-D>. The response annotation is <B-C,A-D>, giving equivalence classes $\{\{A, D\}, \{B, C\}\}$

Recall: To calculate Recall, we take each equivalence class S of the key, and compute the minimum number of links that must be added to the response to place all the elements of S in the same equivalence class. The sum of these counts gives an overall number m of missing links. For the example, the key equivalence class $\{B, C, D\}$ would require one link to be added to the response to bring together its elements in the response equivalence classes, whilst none are required for the key equivalence class $\{A\}$, so $m = 1$. Let c be the minimal number of links needed to generate the key equivalence classes ($c = 2$ for the example), then the Recall is:

$$Recall = \frac{c - m}{c} \qquad (e.g.\ \ Recall = \frac{2 - 1}{2} = 0.5)$$

Precision: To calculate Precision, we simply reverse the roles of the key and the response. The number of links that must be added to the key to put together elements as required by the equivalence classes of the response gives a value m'. For the example, the response equivalence class $\{B, C\}$ requires no additional links (as it is a subset of one of the key equivalence classes), whereas the response equivalence class $\{A, D\}$ requires one additional link to the key, so $m' = 1$. If c' is the minimal number of links to generate the response equivalence classes ($c' = 2$ for the example), then the formula for Precision is:

$$Precision = \frac{c' - m'}{c'} \qquad (e.g.\ \ Precision = \frac{2 - 1}{2} = 0.5)$$

3.2 The B-CUBED Scoring Algorithm

Bagga and Baldwin [2] make the following three criticisms of the MUC6 co-reference scoring approach:

(i) Separating out singletons (i.e. entities that occur only in chains of which they are the only member) from other chains is not given credit by the MUC scoring algorithm.

(ii) All errors are considered equal, i.e. precision is penalised equally for any type of error, though some are more damaging than others, e.g. it is

more damaging to incorrectly link together two long chains than to link a long chain with a short one, i.e. as more entities are incorrectly made co-referent. This distinction is not reflected in the MUC6 algorithm.

(iii) The MUC6 approach gives equal weight to all instances of co-reference. Bagga and Baldwin argue that this is appropriate for IE (where the MUC6 metric has been used), but that alternative weighting schemes are more suitable in other contexts, such as Information Retrieval.

Bagga and Baldwin propose an alternative scoring method for co-reference, the *B-CUBED algorithm*, which looks at the absence/presence of entities relative to each of the other entities in the equivalence classes produced, rather than concentrating on the links produced. Precision and recall for each entity are defined as follows:

$$Precision_i = \frac{number\ of\ correct\ elements\ in\ the\ output\ chain\ containing\ entity_i}{number\ of\ elements\ in\ the\ output\ chain\ containing\ entity_i}$$

$$Recall_i = \frac{number\ of\ correct\ elements\ in\ the\ output\ chain\ containing\ entity_i}{number\ of\ elements\ in\ the\ key\ (true)\ chain\ containing\ entity_i}$$

Final recall and precision are computed by the following formulæ:

$$FinalRecall = \sum_{i=1}^{N} w_i * Recall_i \qquad FinalPrecision = \sum_{i=1}^{N} w_i * Precision_i$$

In these formulæ, N is the number of entities in the document and w_i the weight assigned to entity i in the document. In a case where all entities are equally important, the value $1/N$ could be assigned for all w_i, so that the *Final* scores are just the mean of the scores for each entity. However, the possibility of varying these weights allows the third criticism above to be addressed. The method also addresses the other criticisms, i.e. it gives merit for correctly identified singletons, and it will more severely punish a linkage that incorrectly co-refers more elements than one co-referring less.

3.3 Comparing the Co-reference and Temporal Annotation Scoring Methods

The two co-reference scoring methods both avoid the problem of distinct annotations being semantically equivalent by looking beyond the specific linkages to the equivalence classes they they induce. This is possible of course only because co-reference is an equivalence relation which induces such classes. For temporal annotation, such an approach is not possible, because although the SIMULTANEOUS relation is an equivalence relation, the BEFORE and INCLUDES are not (being asymmetric and irreflexive), and so we cannot put aside the unimportant differences between distinct but equivalent annotations by 'flattening' them down to a representation based purely on equivalence classes. Instead our approach overcomes irrelevant differences by expanding each annotation's model to its temporal closure, allowing comparisons to be made between annotations in terms of their full informational consequences. The question of

whether a scoring scheme could be developed for temporal annotations which is based on some more complicated notion of a minimal model is non-trivial and one that deserves research attention in its own right.

Despite this important difference between our approach and the co-reference scoring methods, we would observe that our approach exhibits some similar characteristics to that of Bagga and Baldwin [2] in regard to their point that some errors are more damaging than others, for which we noted the specific case that an incorrect linking two long chains should be more damaging to precision than an incorrect linking of a long and a short chain. We can construct a similar example for temporal annotation rather than co-reference in terms of 'chains' of entities deemed *simultaneous*. Imagine a key forming three such 'simultaneity chains' such as the following (where $\sim$ denotes simultaneity):

$$\text{KEY:}\quad e_1 \sim e_2 \sim e_3 \sim e_4 \sim e_5 \quad e_6 \sim e_7 \sim e_8 \sim e_9 \sim e_{10} \quad e_{11} \sim e_{12}$$

A response A might incorrectly link the two larger chains, whilst a response B might incorrectly link a large chain to a small one:

$$\text{Resp-A:}\quad e_1 \sim e_2 \sim e_3 \sim e_4 \sim e_5 \sim e_6 \sim e_7 \sim e_8 \sim e_9 \sim e_{10} \qquad e_{11} \sim e_{12}$$
$$\text{Resp-B:}\quad e_1 \sim e_2 \sim e_3 \sim e_4 \sim e_5 \sim e_{11} \sim e_{12} \qquad e_6 \sim e_7 \sim e_8 \sim e_9 \sim e_{10}$$

Our temporal closure-based precision metric does treat the mistake in A as worse than that in B, assigning a precision score of 0.45 for A and 0.67 for B. This is because connecting the two long chains results in a greater number of false inferences linking entities from the two chains during deductive closure. (The recall for both responses is 1, because all the relations that are in the key are also in the response.) Using the MUC scoring method, the precision is 0.9 for both responses, as only one additional link is introduced in each case. For more information about this example refer to [7].

4 Using Temporal Inference to Facilitate Annotation

We now turn to the process of creating gold-standard annotations manually, and consider how this process might be facilitated using insights from the approach described above. A gold standard annotation of temporal information should be such as to determine correctly all the temporal relations holding between all entities (events and times) within the text, insofar as the text does support this information. Note that there is a separation between an annotation that supports this information and one that makes it explicit, i.e. because an annotation that does not explicitly state all the relations may yet imply them under deductive closure, and, with our approach to temporal evaluation, a lesser annotation that supports the full information is functionally equivalent to one that makes it all explicit.

The central question here is how the manual annotation process should be formulated so as to enable annotators to produce gold-standard annotations that are of high quality, with the least effort. Assume that annotators will

firstly identify and mark temporal entities, i.e. times and events. Then, for temporal relations, we might ask annotators simply to mark those relations that are salient to them from the text. Such an approach, however, runs the risk that many relation instances that are supported by the text but which are not immediately salient will be missed, so that incomplete annotations are produced. An alternative strategy would be to ask annotators to consider *every* possible pairing amongst the identified entities, and consider which if any of the temporal relations might hold between the paired elements. Such an approach seems much more likely to elicit an annotation capturing the full temporal content of the text, perhaps even one that makes all relations explicit. However, even if the annotation software were to provide support for the enumeration of pairings to be addressed (which could easily be done), such an approach threatens to impose too high a burden upon the annotator, e.g. a text containing only 20 events and times (which would be quite a small text), would give rise to around 200 entity pairs for consideration (assuming we treat $a+b$ and $b+a$ as the same pair), whilst a text with 50 entities would yield over 1000 pairs.

We propose an alternative annotation strategy which combines the above two approaches and additionally uses temporal inference to facilitate the process and reduce the amount of work required of the annotator. In this approach, the annotator initially marks up some portion of the temporal relations within a text, presumably the most salient ones, or those signalled explicitly by temporal prepositions or temporal subordinating conjunctions. In a second, interactive, stage, the annotator is questioned regarding the temporal relationship between pairs of entities for which this information is unknown. (Note that a valid response at this point is that the relation is not determined by the text.) Crucially, throughout this second phase, temporal inference is used to compute the deductive closure of the annotation done so far. Such inference will resolve the relational status of many entity pairs which are not explicitly linked by the annotation, so that the annotator need only be questioned about those pairs whose relationship is unknown. Each additional annotation that results from questioning may itself have consequences under inference, further reducing the number of unresolved pairs. The process terminates when the relational status of all pairs of entities has been resolved.

The effectiveness of this approach in reducing the work required to produce complete annotations is supported by the results of the following experiment, which is discussed in greater detail in [7]. Six texts were used, of average length 312 words and containing on average 31 temporal entities. The texts were annotated by 2 or 3 subjects each, producing 16 annotations in total. In the initial phase, 200 temporal relations were marked up across all 16 annotated texts. In the second phase, annotators were prompted with a total of 865 questions until the phase reached completion, giving an average of 54 prompts per annotated text. The temporal closure of these annotations contained in total 5288 relational pairs, of which 3.8% (i.e. the 200) were manually

annotated in the initial phase, 16.3% were generated under prompting, and the remaining 79.9% were automatically inferred. While an average of 54 prompts to the annotator is not ideal, it shows a significant improvement over the alternative without temporal inference, which would have required 318 prompts per document on average during the second phase.

In the above experiment, the selection of the next unresolved entity pair to question the annotator about was made randomly. This approach could be improved by using a more intelligent method for selecting these pairs, with a view to minimising the number of prompts that must be made. To illustrate this possibility, consider an example where the only relation is BEFORE (denoted $<$), and where the current annotation of a text produces an ordering of entities of $a < b < c$ and $d < e < f$. The selection of different pairs from amongst a–f will determine the relation of more or less unresolved pairs, depending on the answer given. Prompting for (b, e), for example, will resolve 4 relation pairs whether the answer is $b < e$ or $e < b$ (we ignore here the possibility of a 'not related' response), whilst prompting for (a, d) will resolve 3 pairs, again with either answer. This suggests that (b, e) is a better prompt pair than (a, d). A prompt for (c, d) will resolve 9 relation pairs if the answer is $c < d$, but only 1 if the answer is $d < c$, so the preference between (c, d) and (b, e) is less obvious. If we assume that answers are equally likely, then (c, d) is better, resolving 5 pairs *on average*, but we might instead go for the best guarenteed result, which favours (b, e). A further possibility is that we might find cues to suggest that one answer is more likely than another for a given pair (e.g. perhaps from the position of temporal entities within the text), and use this to determine the likely outcome for each pair as a basis for choosing. This topic merits further investigation.

Two annotation tools have been developed which incorporate the above approach (neither using the idea of 'intelligent' prompt selection just discussed). The first, described in [7], prompts the user textually for missing relations, while at the same time highlighting the entities involved (the text itself being shown in a separate window). The annotation tool developed as part of the TERQAS workshop[4] employs temporal closure within an algorithm known as *text segmented closure*, in which a window covering just a few sentences is slid over the text and the annotator is only prompted to ensure a complete interrelation of entities within the window. This move is a response to the fact that for larger documents containing many entities, the number of prompts required may be unacceptably high, even using temporal closure. This tool also uses a more sophisticated visual method of aiding the user in asserting missing temporal relations.

[4] www.time2002.org

5 Conclusions and Future Work

Comparing and evaluating temporal annotations is an important part of developing annotation schemes and annotated corpora. In this paper, we have described a method to do this, based on computing the temporal closure of temporal annotations, using inference rules. We have have also described how this approach can be used to facilitate the process of creating gold standard annotated corpora, and we have reported some preliminary results that indicate the effectiveness of the method.

In future work, we hope to further investigate how 'intelligent' selection of prompt pairs can improve the effectiveness of the annotation method (as discussed in the preceding section), and to extend the inference rules to cover annotation schemes employing larger sets of temporal relations. Another area of future work would address semantic relations that are not temporal, but that may have temporal consequences, such as cause and effect and sub-eventness.

References

[1] Allen, J., *Towards a General Theory of Action and Time*, Artificial Intelligence **23** (1984), pp. 123–154.

[2] Bagga, A. and B. Baldwin, *Algorithms for Scoring Coreference Chains*, in: *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, 1998, pp. 563–566.

[3] Crowe, J., "Constraint Based Event Recognition for Information Extraction," Ph.D. thesis, University of Edinburgh (1997).

[4] Harper, L., I. Mani and B. Sundheim, editors, "Proceedings of ACL-EACL 2001, Workshop for Temporal and Spatial Information Processing," Toulouse, France, 2001.

[5] "Proceedings of the Sixth Message Understanding Conference (MUC-6)," Morgan Kaufman, 1995.

[6] Pustejovsky, P., J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer and G. Katz, *TimeML: Robust Specification of Event and Temporal Expressions in Text*, in: *Proceedings of the IWCS-5 Fifth International Workshop on Computational Semantics*, 2003.

[7] Setzer, A., "Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study," Ph.D. thesis, University of Sheffield (2001).

[8] Setzer, A., editor, "Proceedings of LREC 2002, Workshop on Annotation Standards for Temporal Information in Natural Language," Pas Palmas, Gran Canaria, 2002.