Evaluation Metrics for Automatic Temporal Annotation of Texts

Xavier Tannier*, **Philippe Muller**†

*LIMSI-CNRS
University Paris-Sud 11
B.P. 133 - F-91403 ORSAY Cedex
Xavier.Tannier@limsi.fr

†Toulouse University
118 Route de Narbonne
F-31062 TOULOUSE CEDEX 9
muller@irit.fr

Abstract

Recent years have seen increasing attention in temporal processing of texts as well as a lot of standardization effort of temporal information in natural language. A central part of this information lies in the temporal relations between events described in a text, when their precise times or dates are not known. Reliable human annotation of such information is difficult, and automatic comparisons must follow procedures beyond mere precision-recall of local pieces of information, since a coherent picture can only be considered at a global level. We address the problem of evaluation metrics of such information, aiming at fair comparisons between systems, by proposing some measures taking into account the globality of a text.

1. Introduction

Recent years have seen increasing attention in temporal processing of texts (see Mani et al. (2005), or the dedicated track at SemEval 2007 (Verhagen et al., 2007)), justifying the need for some standardization effort (Pustejovsky et al., 2005). Temporal information is an essential piece of knowledge for many applications like summarisation, question-answering or information extraction (Hagège and Tannier, 2008). Automatic temporal annotation is generally two-fold:

- Events and temporal adjuncts are extracted from the text. Several definitions of what an event is can be given, but most of the state-of-the-art systems consider mainly events introduced by finite verb phrases, and sometimes by certain noun or adjectival phrases;
- Ideally, a time-stamp is assigned to each event when possible, or a temporal ordering between them is computed. This is done using linguistic and extra-linguistic information such as temporal markers, verb tenses and aspects but also lexical and pragmatic knowledge.

This second task is fairly hard since temporal information is not local, but spread out in a coherent manner throughout the text. There are many equivalent ways to express the same ordering of events. As a consequence, consensual human annotation is difficult, and automatic evaluation must follow procedures beyond mere precision-recall of local pieces of information (Setzer et al., 2006).

We address the problem of evaluation metrics of such information, aiming at fair comparisons between systems, regardless of certain bias that are artificially introduced in current practices. We first address the issues of temporal annotation and problems that must be solved (Section 2.) and then describe a few original metrics and their behaviour on a corpus of temporally annotated texts (Sections 2.2. and 3.).

2. Temporal Processing and Evaluation

It is difficult to reach a good agreement between human annotators on event ordering, for two reasons (Setzer et al., 2006): First, human subjects can express relations between events in different, yet equivalent, ways. For instance, they can say that an event e_1 happens during another one e_2, and that e_2 happens before e_3, leaving implicit that e_1 is before e_3 too, while another might list explicitly all relations. This makes it hard to reach an exhaustive list of temporal relations, and harder to verify such relations.

The second problem is that some relations can be described in more or less precise ways (for instance "e_1 *is before* e_2" is more precise but consistent with "e_1 *is before* e_2 *or* e_1 *overlaps* e_2"), making necessary the handling of partial relevance if only a subset or a inclusive set of relations have been found in another annotation.

We have addressed this latter issue in (Muller and Tannier, 2004). Taking disjunctions into account was also part of the evaluation of a temporal task at SemEval 2007 (Verhagen et al., 2007).

The first issue implies the definition of a referent to which each annotation should be compared, and this is the focus of this paper. What is usually done (see among others (Setzer et al., 2006)) is to use inference rules capturing the formal links between relations, such as Allen's algebra on relations (Allen, 1983), and compute a temporal closure on the graph of temporal relations on events.

Temporal closure is a reasoning mechanism that consists in composing known pairs of temporal relations in order to obtain new relations (e.g.: if A is *before* B and B *contains* C then A is *before* C[1]). These new relations do not really bring

[1] A table of all composition rules can be found for example in (Allen, 1983) or (Rodríguez et al., 2004).

new intrinsic constraints, but allow to produce new explicit information. The temporal closure generally leads to incomplete information, *i.e.* disjunctives relations[2]. Therefore, 2^n different relations can hold between two nodes.

Only closures are compared, with potentially n^2 relations if there are n events in a text. Using inference rules capturing the formal links between relations, such as Allen's algebra (Allen, 1983), and computing a temporal closure, is now widely accepted as necessary (Setzer et al., 2006).

2.1. Importance of relations

However, in a temporal graph, all relations do not have all the same importance, since some are crucial while others can be deduced from the others, but not the other way around. Metrics used so far in temporal evaluation do not deal with this aspect; the final values of recall and precision (or equivalent) on a graph are just an average of measures for each relation. To see why this is a problem, consider the very simple graph examples of Figure 1, in which the first graph K is the gold standard. S_1 contains only two relations against six in K. But it seems unfair to consider a recall score of $\frac{2}{6}$, since adding only one relation (B *before* C) would be enough to infer all others. An intuitive recall would be around $\frac{2}{3}$. This is very similar to the problem of measure agreement on coreference chains, as in the MUC campaigns (Vilain et al., 1995). In coreference chains, only an equivalence relation is used, so a good measure can be made by restricting the evaluations to minimal spanning trees of annotations. Things are more complex in the temporal case, however.

Back to the example, in S_2, the relation "B *before* D" is found. This relation exists but is "minor" in K (*i.e.* useless, because it can be deduced by the transitivity of $<$ from $B < C$ and $C < D$); But in S_2, it is not the case, and this relation must then be rewarded. However, even if the amount of temporal information brought by S_2 and S_3 seem equivalent (two "major" relations and one "minor"), S_3 should get a higher score. Indeed, the amount of *missing* relations (to come to the full graph) is much lower in S_3 (only "C *before* D" is missing) than in S_2.

Note that the issues described here concern only the recall measure, since they are related to the importance of *missing* information. Precision or precision-like measure would not be affected.

2.2. Measuring information in a text

In order to estimate a good way of measuring temporal information in a text, one has first to decide if one focuses on simple relations that can be extracted directly (e.g. event e_1 is before event e_2) and then precision and recall on triplets of the form (event,event,relation) is enough; or all that can be inferred from the text and which is relevant to the temporal structure of a text, and then we have to deal with more complex information such as disjunctions.

Then there is the question of the information provided by a text from a global point of view. When inference is used on a representation, a lot of information is potentially added.

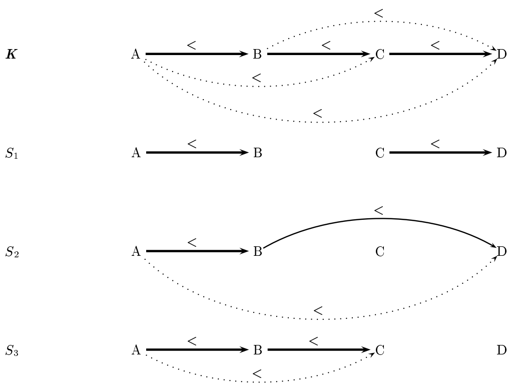

Figure 1: Examples of temporal graphs and relations

A fundamental problem is that when there are n entities extracted (events or dates), there can be up to $n*(n-1)/2$ relations between them (provided also that we do not make a difference between $r(x,y)$ and $r^{-1}(y,x)$). If we just count the total number of relations found, as is usually done, the importance of a text is not a linear function of its "size" (in terms of events introduced, which is generally proportional to its size in word tokens), but a square function of the number of events (because the number of relations can be up to $n*(n-1)/2$).

For the aforementioned co-reference task of MUC-6 (Vilain et al., 1995), there was a similar problem to estimate recall. The scoring proposed estimated the minimal number of missing links necessary to complete a co-reference chain in order to make it match the human annotation, with respect to the minimal number of links necessary to generate the whole annotation (a minimal spanning tree). We want to design something similar for temporal annotation.

The metric that we propose here will be similar in spirit, although temporal graphs are more complex since relations between events can have different values, and a change on an edge propagates constraints possibly through the entire graph.

The choice of the measure applied at the relation level is independent. In other words, it can be used with strict or relaxed recall, or any measure existing to compare two temporal relations.

We will call a "minimal graph" a maximal graph from which no relation can be removed without losing any temporal information after temporal closure. In Figure 1, minimal graphs are composed with bold (not dotted) relations. Unfortunately, a unique minimal graph does not exist in the general case, and in particular for Allen relations. Rodríguez et al. (2004) propose a way to find all minimal graphs for a given temporal graph. The algorithm suggested by (Rodríguez et al., 2004) first finds the *core* relations by intersecting all derivations[3], and then computes all possible remaining combinations in order to find those composing a

[2] For example, with Allen relations, if A *includes* B and B is *before* C, then A *includes* or *overlaps* or *meets* or *overlaps* or *is finished by* C.

[3] For example, for the relation $R_{A,B}$ between A and B, derivations are $R_{A,C} \circ R_{C,B}$, $R_{A,D} \circ R_{D,B}$, etc. If the intersection of all these derived relations equals $R_{A,B}$, it means that $R_{A,B}$ is not a *core* relation, since it can be obtained by composing some other ones. Otherwise, the relation is a *core* one, since removing it always leads to a loss of information. This operation is computationally feasible. The way this kernel is obtained ensures its

minimal graph.

This latter procedure is computationally heavy, and Rodríguez et al. (2004) do not detail much their empirical investigations. To solve these problems we decided for now to focus on a "kernel", the set of *core* relations, *i.e.* the relations found in every minimal graph. These are easier to find and form a unique set: they are the relations whose removal yields an incomplete graph after closure (with respect to the original key).

The evaluation procedure is the following:

- We close both key and candidate graphs, respectively called K and G.
- We look how many core relations from K have been found in G. This is value r_c, or kernel recall.
- We check how many core relations from G are also in K. These may not be crucial in the key but are essential in the candidate graph, and make up for unfound core information (if r_c is one, this set is empty and hence the measure is vacuous). This is kernel precision.

Since core relations does not contain all information provided by closed graphs, this measure is only an approximation of what should be assessed. However, it gives a good idea of how important the relations are in a same graph. More importantly, we expect these measures to grow more linearly with the number of events in a text.

2.3. Defining general measures

We try here to unify the possible measures of temporal annotation on a text. An annotation is a set of constraints on a set of entities, a graph of binary constraints with labels on edges. Let H be the human annotation, with a set E_h of (temporal) entities and A_h a set of constraints. A constraint is a triplet with two events and a relation.

$H = (E_h = \{e_1, e_2, ...\}, A_h = \{(e_4, e_6, before), ...\})$

Let $S = (E_s, A_s)$ be the similar annotation by a system to evaluate.

Let $E = E_s \cup E_h$ the set of all events extracted, either by the first or the second annotation.

On a saturated graph we do not include edges with no information on them (no information means any relation can hold between the two events).

Let g be a measure of comparison between two relations. We can define a generic measure on two representations of a text as depending on two filters and a measuring function, $measure(H, S, f_1, f_2, g)$ as

$$\sum_{a \in f_1(A_h) \cup f_2(A_s)} (g(A_h(a), A_s(a)))$$

A filter f is of the kind:

$$f : (E \times E \times R)^2 \to (E \times E \times R)^2$$

The two filter functions f_1 and f_2 pick out which entities we want to focus on in each representation, and g is the way we compare atomic representations. Here, $A_h(a)$ (respectively $A_s(a)$) stands for the relation holding on the edge with the same endpoints as a in the human annotation (respectively the system annotation). We can then define more specific measures, with various focuses:

- With f1=Id (identity), f2=($x \mapsto \emptyset$), we have a family of measure focusing on the recall of edges in the human annotation;
- With f2=Id, f1=($x \mapsto \emptyset$), we have a family of measure focusing on the precision of the edges contained in a system's annotation.
- If f1 or f2 only consider edges with a "single relation": $f(A) = \{(e_i, e_j, R) \in A/|R| = 1\}$ and with $g = (\lambda x, y.x = y)$, we have the type of comparison of (Mani and Wilson, 2000), precision and recall on "single" (precise) relations. We have then the other filter mapping to the null set.
- We can also filter event-event relations (event ordering) or events-dates (anchoring)

We can use gradual measure of comparisons for atomic representations, like Jaccard, Finesse or Coherence (Muller and Tannier, 2004):

- Jaccard = $(a_H \cap a_S)/(a_H \cup a_S)$, a global overlap measure between the set of possible relations in the human annotation (a_H) and the system's (a_S) on a single edge. Other similar measures, such as the Dice coefficient could be used, but since they are monotonically transformable into one another, we can focus on the study of only one of them.[4]
- Finesse = $(a_H \cap a_S)/a_S$ applied to a recall on human edges was proposed in (Muller and Tannier, 2004) to estimate how informative was a system's annotation.
- Coherence = $(a_H \cap a_S)/a_H$ combined to a measure of precision of edges was introduced in (Muller and Tannier, 2004) to give an estimation of the correction of the system's annotation.

The global finesse score of an annotation is the average of a measure on all edges that have information according to the human annotation once the graph is saturated, while coherence must be averaged on the set of edges that bear information according to the system annotation.

Finesse is intended to measure the quantity of information the system gets, while coherence gives an estimate of errors the system makes with respect to information in the text. Finesse and coherence thus are somewhat similar respectively to recall and precision, but in a gradual setting.

We can also estimate every measure with respect to any set of relation: in our case, the set of annotation relation (AR) to focus on the task, or the set of Allen relations to focus on the underlying computation.

uniqueness.

[4] If j is a Jaccard score and d a corresponding Dice coefficient, $d = 2j * (1 + j)$.

3. Evaluation of the evaluation

All the previous measures focus on one aspect of an evaluation and all seem plausible ways of estimating the similarities between representations. Obviously, many more could be devised, so we must ask ourselves what kind of criteria we can have to estimate what are good measures. A few commonsense criteria could include:

- A measure is decreasing with the decreasing of information (in a monotonic and if possible regular way, even linearly with the level of information or correctness provided);
- We have to deal with different number of edges in a representation (see above): we must look at how a measure behaves on texts of different sizes (in terms of events and dates; as was stated above, the number of relations grows in the order of n^2 if n is the number of temporal entities).

In order to do this, we have designed a few empirical tests to see what measures seem the most useful. They are the following:

- Compare an annotation with the same annotation where information is taken out one piece at a time ;
- Compare an annotation with the same annotation where some noise is introduced: a relation is changed to something else ;

We are going to show a sample of what we have found among all the possible combinations from our proposal. We tested these on the TimeBank corpus[5], which is a set of news article in English.

This corpus is annotated with relations similar to Allen's relations, so it was directly usable, if not free from errors. A few texts had inconsistent temporal annotations, so they do not appear in the evaluation (since temporal closure will fail on these texts).

3.1. Consequences of removing information

In the first experiment, we try to see the behavior of a measure with respect to the quantity of information contained in a text in the following way: we take a human annotation of a text, and we remove (at random) an annotated relation before saturating the graph of constraints, and then we compare it with the initial graph. We average this on a number of different runs. We then do the same by removing 1, 2, ... n-1 relations if there are n relations in the initial annotation. We report the behavior of different measures by putting scores in relation with the percentage of the initial annotation that was removed (and averaging on texts with the same amount removed), so that we can see all texts together. One disadvantage of this is that text with few relations give an overall impression of having high scores because they lack points in the low-information area.

Precision on single measures is always going to be 1 given our protocol, so this experiment is only useful to observe the behaviour of "recall"-types measures: recall of all single relations, recall of relations only from the kernel of relations as explained above, and a gradual relation on each edge (finesse).

Figure 2 compares various measures with respect to the percentage of information (the number of annotated relations) removed in the first experiment. Scores decrease quickly with the number of relations, following an inverted square root (For each measure we show the best parabolic fit).

The relations include a simple measure of the recall of single (non disjunctive) relations, a measure of finesse on each informative edge, and a recall on relations that are in the kernel of the key.

It is interesting to note that the measure based on the kernel relations has a more linear fit than the other measures, as it concentrates on the relevant set of relations. That was the intended effect of focusing on a set of relations considered as more "central" to the text.

We have noticed that averaging on the number of events for each texts does not change much the results, something we expected for the kernel measure, but not for the other ones. The curves remain parabolic in nature, probably because averaging over texts while removing events at random even out the results anyway, while a measure on kernels stay stable.

We have nonetheless noticed very different results on small texts (less than 10 relations) than on larger texts (some have more than 120 relations in TimeBank). We hope these measures can help find annotation biases in such corpora, but obviously this needs to be investigated in more detail.

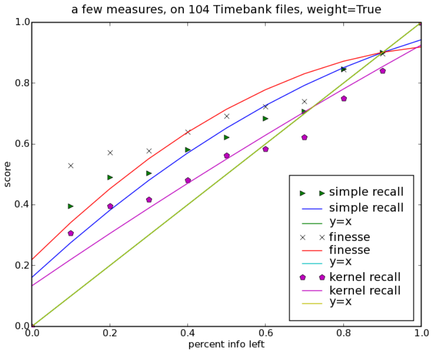

Figure 2: Behaviour of measures when removing information from annotation

3.2. Consequences of some incorrect information

A second test ("small change") was done in a similar way, but instead of removing relations from the original unsaturated human annotation, it was slightly modified by picking out a few relations that were randomly changed, provided this preserved consistency of the global representation. When no change is possible preserving consistency after a number of tries, we stop, (so the points near 0 average on less texts, and are less regular). This experiment is done to observe the behaviour of various measures when confronted to variations of precision of an annotation (as opposed to variations of recall in the previous experiment).

[5] See http://corpora.dutchboy.net/timebank/.

In order to be able to compare texts that had not the same number of events to begin with, we plot the scores against a ratio of undisturbed events with respect to the total number of events in a text, in a way similar to the experiment above. As we average over texts of various sizes, we smoothed the plot by putting points in bins around every 10% of information left unchanged. It has to be noted that the variance around each of this points is sometimes high.

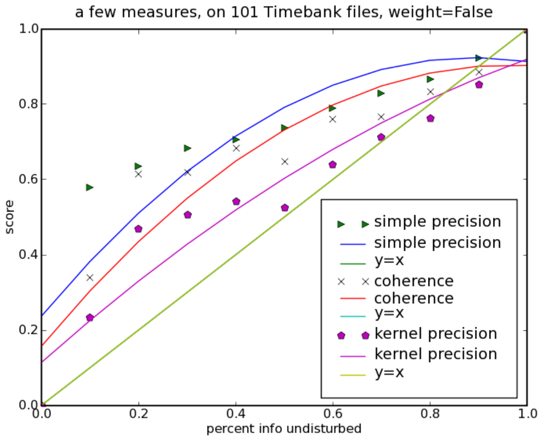

Figure 3: Behaviour of precision measures when disturbing information (unweighted)

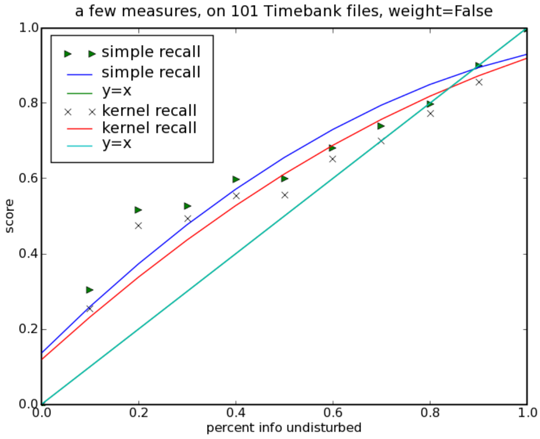

Figure 4: Behaviour of recall measures when disturbing information (unweighted)

The precision-type measures that we have tested here are: the precision of single relations, a measure on coherence of each edge with information in the changed annotation (thus including disjunctive relations), and a precision on the relations that are part of the changed annotation.
The behaviour of all masures are rather far from the ideal $y = x$ (a linear decrease, shown also on the graph). This is shown again by their best parabolic fit (figure 3). This time the weighting of texts is crucial to reaching a sort of balance, most notably with the kernel-related measures (figure 4). This shows that considering only the kernel of relations in the annotation to be compared to the key is not enough to be safe from the "parabolic" effect. Indeed, since kernels are different, we must be sure that they are close enough so the degrading of information is only proportional to the number of events considered. This appears not to be the case. However, averaging each text over its number of events reinstate a better behaviour for all measures, all the more for the kernel-based one.

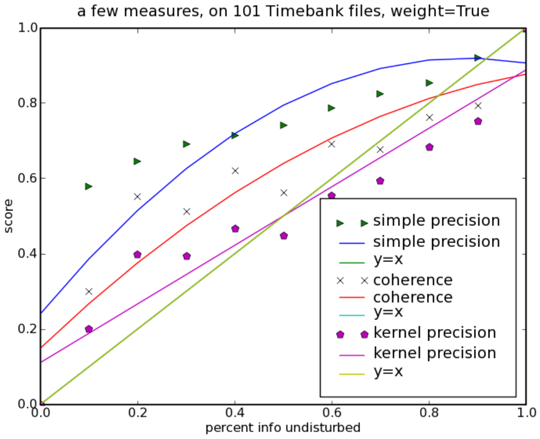

Figure 5: Difference of precision measures when disturbing information (weighted)

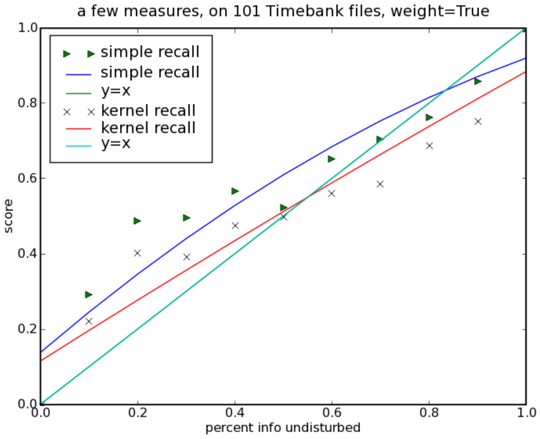

Figure 6: Difference of recall when disturbing information (weighted)

For recall measure when degrading annotations, we see a somewhat less dramatic but similar result (figures 5 and 6), showing this experiment is less relevant to that type of measures.

4. Conclusion

In this paper, we have tried to give some empirical background to the design of the measure of temporal information in written texts. A lot of evaluations and measures in the literature are based on assumptions that are not always explicit, or lack a study of their properties with respect to the few problems we mention in the beginning: that temporal information is global, and that some relations are more important than others. Although much work is still to be done (especially for finding minimal graphs), we hope that that kind of study will help improve evaluation of annotating temporal information in texts. In short what we have found is we can have a sort of recall measure that is immune to the bias introduced by the size of a text in a given corpus by restricting the evaluation to a kernel of central relations, and that the precision on that same kernel can be used also as a more stable measure than the others that we tested, once we average the result on a text by the number of events it contains. Our work has a number of biases that need to be further studied: we have only experiment on news-related texts. Moreover, they vary in size in a way that can have an influence on the stability of our results: a lot of them are very short, and the longer are much longer than the small ones. It is very likely that threshold effects are hidden in our first attempt at this kind of investigation.

5. References

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

Caroline Hagège and Xavier Tannier. 2008. XTM: A Robust Temporal Text Processor. In *Computational Linguistics and Intelligent Text Processing, proceedings of 9th International Conference CICLing 2008*, pages 231–240, Haifa, Israel, February. Springer Berlin / Heidelberg.

I. Mani and G. Wilson. 2000. Robust temporal processing of news. In *Proceedings of ACL 2000*.

Inderjeet Mani, James Pustejovsky, and Robert Gaizauskas, editors. 2005. *The Language of Time: A Reader*. Oxford University Press.

P. Muller and X. Tannier. 2004. Annotating and measuring temporal relations in texts. In *Proceedings of Coling 2004*, volume I, pages 50–56, Genève.

James Pustejovsky, Robert Ingria, Roser Sauri, Jose Castano, Jessica Littman, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. In I. Mani, J. Pustejovsky, and R Gaizauskas, editors, *The Language of Time: A Reader*. Oxford University Press.

Andrea Rodríguez, Nico Van de Weghe, and Philippe De Maeyer. 2004. Simplifying Sets of Events by Selecting Temporal Relations. In *Geographic Information Science, Third International Conference, GIScience 2004*, volume 3234/2004 of *Lecture Notes in Computer Science*, pages 269–284, Adelphi, MD, USA, October. Springer Berlin / Heidelberg.

Andrea Setzer, 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.

Marc Verhagen, 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.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52, Morristown, NJ, USA. Association for Computational Linguistics.