An Evaluation Framework for Aggregated Temporal Information Extraction

Enrique Amigó
UNED NLP & IR group
Madrid, Spain
enrique@lsi.uned.es

Javier Artiles, Qi Li, Heng Ji
Computer Science Department
Queens College and Graduate Center (CUNY)
New York, USA
javier.artiles@qc.cuny.edu

ABSTRACT

This paper focusses on the representation and evaluation of temporal information about a certain event or entity. In particular, we consider temporal information that can be normalized to specific dates. This task requires the aggregation of temporal relations between events and dates extracted from multiple texts. Given that the resulting temporal information can be vague, it is necessary that an evaluation framework captures and compares the temporal uncertainty of system outputs and human assessed gold-standard data. Current representation models and measures are not suitable for this scenario. In this paper, we propose a novel representation model and assess its properties and limitations. In order to compare extracted information against a gold standard, we define an evaluation metric based on a set of formal constraints. Finally, we present experiments that show the behavior of the proposed metric. The task setting and the evaluation measure presented here have been introduced in the TAC 2011 Knowledge Base Population evaluation for the Temporal Slot Filling task.

1. INTRODUCTION

Recent research on Information Extraction (IE) has seen an evolution from single document to cross-document extraction tasks in an effort to build structured knowledge bases from large collections of unstructured texts. Among the most successful efforts in this direction is the Knowledge Base Population (KBP) track, which has gathered researchers around the problem of extracting of information about entities linked to an external knowledge source [13, 9]. Another example is the Web People Search (WePS) task, where systems are requested to cluster multiple search results about the same person and extract biographical attributes [3, 4, 2].

Several recent studies have stressed the benefits of using information redundancy on estimating the correctness of IE output [7], improving disease event extraction [22] and MUC event extraction [12, 14]. However, the information obtained on IE tasks such as KBP or WePS is viewed as static, ignoring the temporal dimension that is relevant to many types of attributes. While this is a reasonable simplification in many situations, it is unsatisfactory for applications that require some awareness of the time span when a fact was valid. Furthermore, considering a temporal dimension on a cross-document task adds the challenging problem of managing vague or incomplete information about the temporal boundaries of a particular fact.

This work focuses on the development of an evaluation framework for extracting temporal information about a particular event in order to estimate its specific temporal boundaries. Temporal information can be scattered across documents (e.g. in one document "John joined Microsoft in Sept. 1990" and in another document "Microsoft resigned his contract yesterday "), and expressed with different granularities (e.g. "He started working for Microsoft in the 90s", "He began his contract in September of this year"). Several evaluation frameworks for temporal information extraction have been defined; although, most of the previous work has focused on temporal relations extracted from single texts. Other work addresses information aggregation, but they do not provide a formal representation and evaluation framework.

In this paper, we present a model for representing temporal information that has been derived from the task requirements. After this, we present a novel evaluation measure that allows the comparison of system outputs against human assessed solutions. Given that the representation model is new, it is not possible to compare our approach to a previously stablished standard evaluation measure. However, we do define a set of formal constraints which state the properties that any measure should satisfy on this task. In addition, we describe some experiments that empirically assess the measure behavior and properties. This work has motivated the introduction of the Temporal Slot Filling task in the TAC 2011 Knowledge Base Population (KBP2011) evaluation campaign. In this task, systems are required to add temporal boundaries to IE slots such as "employee of", "country of residence" or "spouse".

In Section 2 we formally describe our proposed evaluation framework. In Section 3 we compare our proposal with previous related work. Section 4 describes our experiments and in Section 5 we discuss our conclusions.

2. THE EVALUATION FRAMEWORK

2.1 Time Representation

According to our task definition, we assume that: (i) events are not discontinuous in time; (ii) the temporal information is distributed across several documents; and (iii) both the gold standard and system outputs can contain uncertainty. This uncertainty can be due to the variable levels of granularity of temporal information (e.g. years, months) or to the

Copyright is held by the author/owner(s).
EOS, SIGIR 2011 workshop, July 28, Beijing, China

reasoning based on temporal order relations ("He worked for Microsoft before working for Apple").

Given the previous assumptions the representation model should consider temporal ranges for both the beginning and ending points. For simplicity, we assume that uncertainty follows a uniform distributions over time ranges. Our representation model consists of a tuple $< t_1, t_2, t_3, t_4 >$, which represents the set S of possible beginnings and endings of an event such that:

$$S = \{< t_{init}, t_{end} > | (t_1 < t_{init} < t_2) \wedge (t_3 < t_{end} < t_4)\}$$

In other words, t_1 and t_3 represent the lower bounds for the beginning and ending points respectively, while t_2 and t_4 represent the upper bounds.

Interesting properties of this temporal representation model include:

Aggregation: Each tuple represents a set of possible pairs $< t_{init}, t_{end} >$, allowing a straightforward aggregation of temporal information:

$$S \cap S' \equiv < max(t_1, t'_1), min(t_2, t'_2), max(t_3, t'_3), min(t_4, t'_4) >$$

Temporal relations event-time: This model is able to represent Allen's relations [1] between an event and a time interval. For instance, if an event S happens before a certain time interval $< t_a, t_b >$, then the latest time boundaries in S must be previous to the beginning of the time interval (t_a)[1] (see [1] for a graphical depiction of these relations):

$$S_{(BEFORE < t_a, t_b >)} \equiv < -\infty, t_a, -\infty, t_a >$$

Temporal relations event-event: Although our model does not capture directly time relations between events, it allows the reduction of time uncertainty when these relationships are detected. Given two events represented by the tuples S and S' and an Allen relation, there exists a simple operation over tuples that reduces the temporal uncertainty for both events. For instance, if S happens before S', the upper bounds for S must be previous to the S' beginning upper bound:

$$S \ BEFORE \ S' \rightarrow$$
$$(S =< t_1, min(t_2, t'_2), t_3, min(t_4, t'_2) >)$$
$$\wedge (S' =< max(t'_1, t_3), t'_2, max(t'_3, t_3), t'_4 >)$$

Temporal inconsistencies: The representation model provides a straightforward method to detect inconsistencies when aggregating temporal information in a tuple. An event tuple is inconsistent if a lower bound exceeds its corresponding upper bound:

$< t_1, t_2, t_3, t_4 >$ is inconsistent if and only if

$$t_1 > t_2 \vee t_3 > t_4 \vee t_1 > t_4$$

For instance, according to the BEFORE operator previously defined, given to (consistent) events S and S', a temporal relation $S \ BEFORE \ S'$ is inconsistent if the boundary lower bounds in S exceeds the S' beginning upper bound:

$S \ BEFORE \ S'$ is inconsistent if $t_1 > t'_2 \vee t_3 > t'_2$

[1]Due to the space limitations we do not list all relations.

The main limitation of assuming that events are continuous is that our representation model is not able to capture some relations such as regularly recurring events ("each Friday") or some fuzzy relations ("lately", "recently") that are encoded with the SET type in TimeML [15].

In short, besides providing the starting point for our evaluation framework, this representation model allows us to: (i) aggregate temporal information about an event; (ii) represent temporal relations between events and dates (time interval); (iii) reduce the temporal uncertainty about events by considering temporal relations between them; and (iv) detect errors during the aggregation process by checking for inconsistencies between tuples.

2.2 Formal Constraints for an Evaluation Measure

In order to complete the evaluation framework, we need to define a metric $Q(S)$ that compares a system's output S against a gold standard tuple $S_g =< g_1, g_2, g_3, g_4 >$. We now define a set of formal constraints that any quality metric based on our representation model should satisfy.

Best System Constraint: The maximum score should be achieved only by a tuple that exactly matches the gold standard:

$$Q(S) = 1 \leftrightarrow t_i = g_i \ \forall i \in \{1..4\}$$

Worst System Constraint: When the gold standard tuple does not contain any ∞ or $-\infty$ component, an uninformative tuple, such that all elements are ∞ or $-\infty$, always achieves the minimum score. This constraint prevents over-scoring tuples that do not produce any useful information.

$$t_i \in \{\infty, -\infty\} \ \forall i \in \{1..4\} \rightarrow Q(S) = 0$$

Quality Decrease: Any increment of the distance between a tuple component and the corresponding gold standard component implies a quality decrease. This constraint ensures that the metric is sensitive to any change in the tuple.

$$\triangle |t_i - g_i| \rightarrow \triangledown Q(S)$$

Temporal Boundary Independence: The correct estimation of any of the gold standard components g_i has an intrinsic value with no dependency on the rest of components. In other words, estimating correctly any of the components implies a quality higher than 0. The practical implication is that we cannot infinitely penalize an error in one tuple component.

$$(\exists i \in \{1..4\}.t_i = g_i) \rightarrow Q(S) > 0$$

Parameterization Constraint: According to the *Aggregation* property of our representation model, the more we aggregate temporal information, the less the resulting tuple is vague. However, we risk including incorrect information that is over-constraining the upper and lower bounds in the tuple. Depending on the task, the relative penalization for both aspects should be parameterizable in the metric. A parameter α should determine if a certain amount of *vagueness* is worse or better than a certain amount of *over-constraining*.

Formally, given two tuples S and S' with finite values such that $S \subset S_g \subset S'$, then:

$$\exists v/(\alpha > v) \leftrightarrow (Q_\alpha(S) > Q_\alpha(S'))$$

2.3 The Proposed Evaluation Metric

A simple approach to the evaluation metric would be to estimate the precision and recall between S and S_g sets. However, this approach would not satisfy the *Quality Decrease* constraint when S_g and S are not overlapped; for instance, when it is applied over events that happen at a single point in time ($g_1 = g_2 = g_3 = g_4$).

Our proposal consists of measuring absolute distances between components t_i and g_i.

$$Q(S) = \frac{1}{4} \sum_i \frac{c}{c + |t_i - g_i|}$$

Q is bounded between zero and one. An error of c time units produces a 0.5 score. As the value of c decreases errors are penalized to a greater extent. The c value can be fixed for all the events found in a corpus test set, or it can be set dynamically by considering a certain variable (e.g. event length in the gold standard).

According to the *Best System* constraint, the maximum score is achieved when $|t_i - g_i| = 0$ forall i. As required by the *Temporal Boundary Independence* constraint, estimating correctly any of the tuple components results in a quality higher than zero:

$$t_j = g_j \rightarrow Q(S) = \frac{1}{4}(\frac{c}{c + |t_j - g_j|} + \sum_{i \neq j} \frac{c}{c + |t_i - g_i|}) =$$

$$\frac{1}{4}(1 + \sum_{i \neq j} \frac{c}{c + |t_i - g_i|}) > 0$$

In addition, the metric is sensitive to any difference between tuples, satisfying the *Quality Decrease* constraint. For this, it is enough to demonstrate that $Q(S)$ is strictly decreasing regarding $|t_i - g_i|$. If we derive $Q(S)$ regarding any $|t_i - g_i|$, we obtain:

$$\frac{d\ Q(S)}{d\ |t_i - g_i|} = -\frac{c}{(c + |t_i - g_i|)^2} < 0$$

Finally, when all components in the reference tuple are finite, a non finite value in all the evaluated tuple component implies a minimum quality (*Worst System* constraint).

$$Q(< -\infty, \infty, -\infty, \infty >) = \frac{1}{4} \sum_i \frac{c}{c + \infty} = 0$$

In order to satisfy the *Parameterization* constraint, we just have to refine the c parameter by cases. *Vagueness* ($S \supset S_g$) represents to what extent systems tend to produce longer uncertainty ranges. *Over-constraining* ($S \subset S_g$) represents to what extent the constraints are false or the specificity of the evaluated tuple is higher than that of the gold standard tuple. Both quality criteria are affected by the following time mismatches:

$$S \supset S_g \longrightarrow t_1 \leq g_1 \wedge t_2 \geq g_2 \wedge t_3 \leq g_3 \wedge t_4 \geq g_4$$
$$S \subset S_g \longrightarrow t_1 \geq g_1 \wedge t_2 \leq g_2 \wedge t_3 \geq g_3 \wedge t_4 \leq g_4$$

S_g	**1950**	**1960**	**1970**	**1980**	
S	1950	1960	1970	1980	$Q = 1$
S	1945	1955	1975	1985	$Q = 0.5$
S	1940	1950	1980	1990	$Q = 0.33$
S	0	10000	1970	10000	$Q = 0.25$
S_g	**0**	**1960**	**1970**	**10000**	
S	0	1960	1970	10000	$Q = 1$
S	1950	1960	1970	1980	$Q = 0.5$

Table 1: Evaluation examples ($c_{vag} = c_{cons} = 5$).

These two quality aspects are complete. The two possible ordering relations between g_i and $t : i$ are represented. We can satisfy the *Parameterization* constraint by setting a different c value for vagueness and over-constraining.

$$Q(S) = \frac{1}{4} \sum_i \frac{c_i}{c_i + |t_i - g_i|}$$

Notice that an individual component of a tuple can only have vagueness or over-constraining error, but not both. Since the tuple is comprised of four components the overall tuple can have both of those types of errors.

$$c_i = \begin{cases} c_{vag}, & \text{if } (i \in \{1,3\} \wedge t_i \leq g_i) \vee (i \in \{2,4\} \wedge t_i \geq g_i) \\ c_{cons}, & \text{otherwise} \end{cases}$$

The demonstration for the *Parameterization* constraint is straightforward. If $S \subset S_g$ then:

$$Q(S) = \frac{1}{4} \sum_i \frac{c_{cons}}{c_{cons} + |t_i - g_i|}$$

If $S_g \subset S'$ then:

$$Q(S') = \frac{1}{4} \sum_i \frac{c_{vag}}{c_{vag} + |t_i - g_i|}$$

Therefore, depending on c_{cons} and c_{vag} both tuples can achieve any score between zero and one. Thus, there exists a parameter that determines if $Q(S) > Q(S')$.

Table 1 shows some examples of evaluated tuples[2]. In the first case the tuple is an exact match for all components. In the second case, there is an error of 5 years for each component, producing a final score of 0.5. In the next case, the error is greater (ten years) and the score decreases to 0.33. In the fourth case the system is not able to estimate three out of four component, obtaining a score of 0.25. In the last two examples, we have considered the situation in which there is no estimation in the gold standard for the earliest possible beginning and for the latest possible ending points. We approach the infinite values as extremely small and big values (0 and 10000). As the table exemplifies, if the tuple introduces a false lower bound for the beginning and a false upper bound for the ending, and the other two components are well estimated, the score is 0.5.

[2]For simplicity we consider years in our examples, although more fine-grained dates can be used.

3. RELATED WORK

The TempEval campaigns [21, 17] have evaluated the classification of temporal relations between events, time expressions and document creation dates, without considering cross-document relations. Most current research efforts are focused on these tasks [6, 20, 8, 10], which are evaluated using precision and recall measures over the TimeML [16] annotation scheme.

Other works have considered the aggregation of temporal information from multiple documents. In [23] the logical constraints of different temporal relations are considered jointly; although, evaluation is carried separately over each type of temporal relation. Thus, they do not provide an integrated representation and evaluation measure for the temporal information associated to a certain event. [18] represents and aggregates temporal information using a fuzzy version of Allen's Interval Algebra [1]. In this case the evaluation is carried indirectly, in terms of precision and recall of the classification of overlapped or non-overlapped events. Although they represent the uncertainty of temporal relations, they do not offer an evaluation measure that consider the vagueness of system outputs. In [19, 5] the time span of events is estimated based on the distribution of relevant terms across documents, but only a manual evaluation is provided.

Among the previous work [11] is the closest to our proposal. It estimates the ending and beginning time of an event across documents and their representation model consists of a pair indicating the starting and ending points. However, this representation model does not capture temporal uncertainty. They propose two evaluation measures. The first is a discrete measure of the precision and recall of correctly predicted constraints across sentences. Therefore, the *Quality Decrease* constraint is not satisfied. The second measure represents the tightness of the temporal bounds induced by a system. Although this measure is able to penalize long time intervals in responses, it does not give information about the amount of overlap between the estimated and actual bounds and does not satisfy the *Parameterization* property.

To the best of our knowledge, the previous evaluation frameworks are unable to capture and evaluate the uncertainty of temporal information derived from the aggregation of information from multiple documents.

4. EXPERIMENTS

In order to assess the behavior of the metric, we automatically generate a set of gold standard and "system" tuples. We generate 100 gold standard tuples according to the following procedure: r_{init} and r_{end} are two random values from (0..100) and (100..200):

$$g_1 = r_{init} - rand(0..50) \quad g_2 = r_{init} + rand(0..50)$$
$$g_3 = r_{end} - rand(0..50) \quad g_4 = r_{end} + rand(0..50)$$

4.1 Parameterization Experiment

In order to check the ability of the Q measure to weight the relative relevance of *vagueness* against *over-constraining*, we automatically produce two tuples for each gold standard; one vague tuple and one over-constrained tuple.

The components for the over-constrained and the vague tuples are:

$$S_{Overcons} =< t_1 + a_1, t_2 - a_2, t_3 + a_3, t_4 - a_4 >$$

$$S_{Vague} =< t_1 - a_1, t_2 + a_2, t_3 - a_3, t_4 + a_4 >$$

where $a_1 \in (0..(r_{init} - g_1))$,$a_2 \in (0..(g_2 - r_{init}))$, $a_3 \in (0..(r_{end} - g_3))$, and $a_4 \in (0..(g_4 - r_{end}))$. The limits for the a_i values avoid inconsistent tuples (e.g. $t_1 > t_2$). We have computed the average Q score for both approaches across the 100 gold standard cases. Figure 1 shows that depending on the c_{con} and c_{vag} parameters in the Q measure one system can improve the other and *vice versa*.

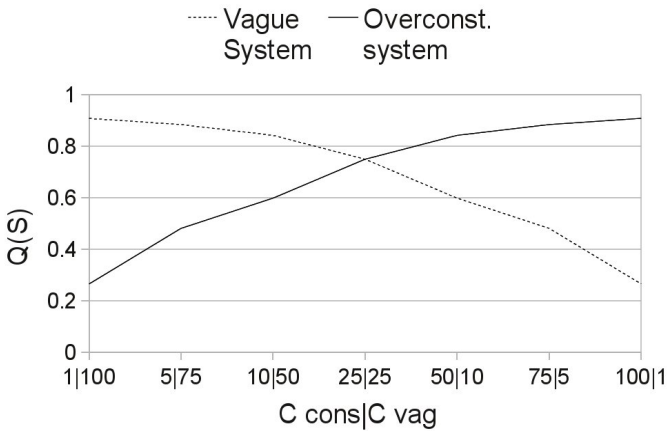

Figure 1: Parameterizing the c_{cons} and c_{vag} values in the Q measure over a vague and an over-constrained approach

4.2 Increasing the Error

The next experiment consists of producing tuples that start as a perfect match with the gold standard, and progressively incorporate a certain amount of error β. First, we generate a HOLD tuple assuming that the system correctly locates points in the time interval, but with an error β in the upper and lower bounds.

$$S_{HOLD} =< g_1 - \beta, g_2, g_3, g_4 + \beta >$$

Next, the BEFORE tuple detects a point before the event and correctly estimates the time lower bound. Analogously, we define the AFTER tuple:

$$S_{BEFORE} =< g_1, g2 + \beta, g_3 + \beta, g_4 + \beta >$$

$$S_{AFTER} =< g_1 - \beta, g2 - \beta, g_3 - \beta, g_4 >$$

The vague tuple progressively increases the uncertainty of the ending and beginning time points:

$$S_{VAGUE} =< g_1 - \beta, g2 + \beta, g_3 - \beta, g_4 + \beta >$$

Notice that it is not possible to emulate an infinitely over-constrained tuple, given that it would not satisfy the basic format constraints (e.g. $t_1 < t_2$). Also, we include a RAND tuple that estimates a single point in the time range with an error β around the center of the event.

$$S_{RAND} \equiv (t_1 = t_2 = t_3 = t_4 = \frac{r_{init} + r_{end}}{2} \pm \beta)$$

Figure 2 shows the Q score for these approaches when β is increased. In this case we have set the constants c_{con} and c_{vag} at 5. As the figure shows, the HOLD tuples approach a 0.5 score in the limit (note that, by definition, this approach captures a half of the tuple component). The AFTER and BEFORE tuples approach 0.3 in the limit, since they capture less information. When the VAGUE tuples are extremely vague their score approaches zero. Finally, the RAND tuples over-constrain the gold standard, achieving a low score even for small values of β.

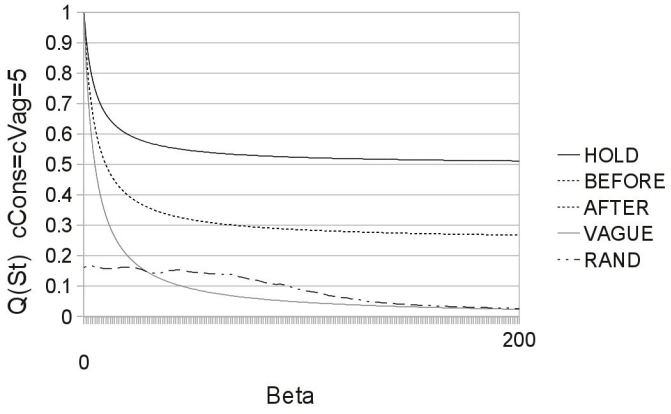

Figure 2: Several synthetic approaches with a increasing error beta

5. CONCLUSIONS

In this work we have proposed a novel framework to represent, aggregate and evaluate temporal information distributed in multiple documents. This framework allows the inclusion of operators that reduce the uncertainty when finding new time relations (events-date or event-event). The second contribution of this paper is the definition of an evaluation metric for this task grounded on a set of formal constraints. The empirical results over synthetic data show that the proposed metric: (i) can effectively assign a relative weight to vagueness vs. over-constraining; and (ii) correctly discriminates the quality of a set of controlled outputs.

One of the straights of the proposed model is its simplicity. Most temporal relations can be captured with a four element tuple or with a simple operator over tuples. Nevertheless two main limitations of our research are that our model can not represent discontinuous events such as regularly recurring events (e.g. "every weekend") and also that the assumption of a uniform distribution for uncertainty ranges does not allow us to capture fuzzy relations (e.g. "it happened at the beginning of the year 2000"). Our future work is to apply the representation and evaluation model over a collection of real data. The ongoing TAC KBP2011 Temporal Slot Filling task will provide a collection of manual assessments as well as testing the proposed metric on a variety of systems.

6. ACKNOWLEDGMENTS

This research was partially supported by the Spanish Ministry of Science and Innovation (Holopedia Project, TIN2010-21128-C02) and the Regional Government of Madrid and the European Social Fund under MA2VICMR (S2009/TIC-1542).

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, the U.S. NSF CAREER Award under Grant IIS-0953149 and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

7. REFERENCES

[1] J. F. Allen. Maintaining Knowledge about Temporal Intervals. In *Communications of the ACM, November 1983, Volume 26, Number 11*, volume 26, pages 832–843, New York, NY, USA, 1983.

[2] J. Artiles, A. Borthwick, J. Gonzalo, S. Sekine, and E. Amigó. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks. In *2nd Web People Search Evaluation Workshop (WePS 2010), CLEF 2010 Conference, Padova Italy*, 2010.

[3] J. Artiles, J. Gonzalo, and S. Sekine. The SemEval-2007 WePS Evaluation: Establishing a benchmark for the Web People Search Task. In *Proc. of the 4th International Workshop on Semantic Evaluations (Semeval-2007)*, 2007.

[4] J. Artiles, J. Gonzalo, and S. Sekine. WePS 2 Evaluation Campaign: overview of the Web People Search Clustering Task. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[5] H. L. Chieu and Y. K. Lee. Query based event extraction along a timeline. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 425–432, New York, NY, USA, 2004. ACM.

[6] L. Derczynski and R. Gaizauskas. Usfd2: Annotating temporal expresions and tlinks for tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 337–340, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[7] D. Downey, O. Etzioni, and S. Soderland. A Probabilistic Model of Redundancy in Information Extraction. In *Proc. IJCAI 2005*, 2005.

[8] E. Y. Ha, A. Baikadi, C. Licata, and J. C. Lester. Ncsu: Modeling temporal relations with markov logic and lexical ontology. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 341–344, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[9] R. G. H. T. D. Heng Ji and K. Griffitt. An Overview of the TAC2010 Knowledge Base Population Track. In *Proc. Text Analytics Conference (TAC2010)*, 2010.

[10] A. K. Kolya, A. Ekbal, and S. Bandyopadhyay. Ju_cse_temp: A first step towards evaluating events, time expressions and temporal relations. In

Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, pages 345–350, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] X. Ling and D. S. Weld. Temporal information extraction. In *Proceedings of the Twenty Fifth National Conference on Artificial Intelligence*, 2010.

[12] G. Mann. Multi-document Relationship Fusion via Constraints on Probabilistic Databases. In *Proc. of HLT-NAACL 2007*, pages 332–339, 2007.

[13] P. McNamee and H. Dang. Overview of the TAC2009 knowledge base population track. In *Text Analysis Conference (TAC)*, 2009.

[14] S. Patwardhan and E. Riloff. A Unified Model of Phrasal and Sentential Evidence for Information Extraction. In *Proc. EMNLP 2009*, pages 151–160, 2009.

[15] J. Pustejovsky, J. M. Castaño, R. Ingria, R. Sauri, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proc. of IWCS-5, Fifth International Workshop on Computational Semantics (IWCS-5)*, 2003.

[16] J. Pustejovsky, J. CastaÃśo, R. Ingria, R. Saura, R. Gaizauskas, A. Setzer, and G. Katz. Timeml: Robust specification of event and temporal expressions in text. In *in Fifth International Workshop on Computational Semantics (IWCS-5*, 2003.

[17] J. Pustejovsky and M. Verhagen. Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In *Proc. of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 112–116, 2009.

[18] S. Schockaert, M. De Cock, and E. E. Kerre. Reasoning about fuzzy temporal information from the web: towards retrieval of historical events. *Soft Comput.*, 14:869–886, June 2010.

[19] R. Swan and J. Allan. Automatic generation of overview timelines. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, New York, NY, USA, 2000. ACM Press.

[20] N. UzZaman and J. F. Allen. Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 276–283, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[21] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 75–80, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[22] R. Yangarber. Verification of Facts across Document Boundaries. In *Proc. International Workshop on Intelligent Information Access*, 2006.

[23] K. Yoshikawa, S. Riedel, M. Asahara, and Y. Matsumoto. Jointly identifying temporal relations with markov logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 405–413, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.