Extending TimeML with Typical Durations of Events

Feng Pan, Rutu Mulkar, and Jerry R. Hobbs
Information Sciences Institute (ISI), University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292, USA
{pan, rutu, hobbs}@isi.edu

Abstract

In this paper, we demonstrate how to extend TimeML, a rich specification language for event and temporal expressions in text, with the implicit typical durations of events, temporal information in text that has hitherto been largely unexploited. Event duration information can be very important in applications in which the time course of events is to be extracted from text. For example, whether two events overlap or are in sequence often depends very much on their durations.

1 Introduction

Temporal information processing has become more and more important in many natural language processing (NLP) applications, such as question answering (Harabagiu and Bejan, 2005; Moldovan et. al., 2005; Saurí et. al., 2005), summarization (Mani and Schiffman, 2005), and information extraction (Surdeanu et. al., 2003).

Temporal anchoring and event ordering are among the most important kinds of temporal information needed for NLP applications. Although there has been much work on extracting and inferring such information from texts (Hitzeman et al., 1995; Mani and Wilson, 2000; Filatova and Hovy, 2001; Boguraev and Ando, 2005), none of this work has exploited the implicit event duration information from the text.

Consider the sentence from a news article:

George W. Bush met with Vladimir Putin in Moscow.

How long was the meeting? Our first reaction to this question might be that we have no idea. But in fact we do have an idea. We know the meeting was longer than 10 seconds and less than a year. How much tighter can we get the bounds to be? Most people would say the meeting lasted between an hour and three days.

There is much temporal information in text that has hitherto been largely unexploited, encoded in the descriptions of events and relying on our knowledge of the range of usual durations of types of events, which can be very important in applications in which the time course of events is to be extracted from news. For example, whether two events overlap or are in sequence often depends very much on their durations. If a war started yesterday, we can be pretty sure it is still going on today. If a hurricane started last year, we can be sure it is over by now.

To extract such implicit event duration information from texts automatically, we developed a corpus annotated with typical durations of events (Pan et al., 2006a) which currently contains all the 48 non-Wall-Street-Journal (non-WSJ) news articles (a total of 2132 event instances), as well as 10 WSJ articles (156 event instances), from the TimeBank corpus annotated in TimeML (Pustejovky et al., 2003).

Because the annotated corpus is still fairly small, we cannot hope to learn to make *fine-grained* judgments of event durations that are currently annotated in the corpus, but as we show in greater detail in (Pan et al., 2006b), it is possible to learn useful *coarse-grained* judgments that considerably outperform a baseline and approach human performance.

This paper describes our work on extending TimeML with annotations of typical durations of events, which can enrich the expressiveness of TimeML, and provides NLP applications that exploit TimeML with this additional implicit event duration information for their temporal information processing tasks.

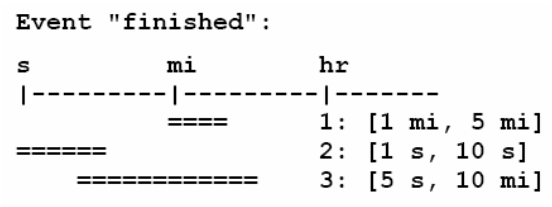
In Section 2 we first describe the corpus of typical durations of events, including the annotation guidelines, the representative event classes with examples, the inter-annotator agreement

study, and the machine learning results. TimeML and its event classes will be described in Section 3, and we will discuss how to integrate event duration annotations into TimeML in Section 4.

2 Annotating and Learning Typical Duration of Events

In the corpus of typical durations of events, every event to be annotated was already identified in the TimeBank corpus. Annotators are asked to provide lower and upper bounds on the duration of the event, and a judgment of level of confidence in those estimates on a scale from one to ten. An interface was built to facilitate the annotation. Graphical output is displayed to enable us to visualize quickly the level of agreement among different annotators for each event. For example, here is the output of the annotations (3 annotators) for the "finished" event (in bold) in the sentence

*After the victim, Linda Sanders, 35, had **finished** her cleaning and was waiting for her clothes to dry,...*



This graph shows that the first annotator believes that the event lasts for minutes whereas the second annotator believes it could only last for several seconds. The third annotates the event to range from a few seconds to a few minutes. A logarithmic scale is used for the output.

2.1 Annotation Instructions

Annotators are asked to identify upper and lower bounds that would include 80% of the possible cases, excluding anomalous cases.

The judgments are to be made in context. First of all, information in the syntactic environment needs to be considered before annotating, and the events need to be annotated in light of the information provided by the entire article. Annotation is made easier and more consistent if coreferential and near-coreferential descriptions of events are identified initially.

When the articles were completely annotated by the three annotators, the results were analyzed and the differences were reconciled. Differences in annotation could be due to the differences in interpretations of the event; however, we found that the vast majority of radically different judgments can be categorized into a relatively small number of classes. Some of these correspond to aspectual features of events, which have been intensively investigated (e.g., Vendler, 1967; Dowty, 1979; Moens and Steedman, 1988; Passonneau, 1988). We then developed guidelines to cover those cases (see the next section).

2.2 Event Classes

Action vs. State: Actions involve change, such as those described by words like "speaking", "gave", and "skyrocketed". States involve things staying the same, such as being dead, being dry, and being at peace. When we have an event in the passive tense, sometimes there is an ambiguity about whether the event is a state or an action. For example,

*Three people were **injured** in the attack.*

Is the "injured" event an action or a state? This matters because they will have different durations. The state begins with the action and lasts until the victim is healed. Besides the general diagnostic tests to distinguish them (Vendler, 1967; Dowty, 1979), another test can be applied to this specific case: Imagine someone says the sentence after the action had ended but the state was still persisting. Would they use the past or present tense? In the "injured" example, it is clear we would say "Three people *were* injured in the attack", whereas we would say "Three people *are* injured from the attack." Our annotation interface handles events of this type by allowing the annotator to specify which interpretation he is giving. If the annotator feels it's too ambiguous to distinguish, annotations can be given for both interpretations.

Aspectual Events: Some events are aspects of larger events, such as their start or finish. Although they may seem instantaneous, we believe they should be considered to happen across some interval, i.e., the first or last *sub-event* of the larger event. For example,

*After the victim, Linda Sanders, 35, had **finished** her cleaning and was waiting for her clothes to dry,…*

The "finished" event should be considered as the last sub-event of the larger event (the "cleaning" event), since it actually involves opening the door of the washer, taking out the clothes, closing the door, and so on. All this takes time. This

interpretation will also give us more information on typical durations than simply assuming such events are instantaneous.

Reporting Events: These are everywhere in the news. They can be direct quotes, taking exactly as long as the sentence takes to read, or they can be summarizations of long press conferences. We need to distinguish different cases:

Quoted Report: This is when the reported content is quoted. The duration of the event should be the actual duration of the utterance of the quoted content. The time duration can be easily verified by saying the sentence out loud and timing it. For example,

*"It looks as though they panicked," a detective **said** of the robbers.*

This probably took between 1 and 3 seconds; it's very unlikely it took more than 10 seconds.

Unquoted Report: This is when the reporting description occurs without quotes that could be as short as just the duration of the actual utterance of the reported content (lower bound), and as long as the duration of a briefing or press conference (upper bound).

If the sentence is very short, then it's likely that it is one complete sentence from the speaker's remarks, and a short duration should be given; if it is a long, complex sentence, then it's more likely to be a summary of a long discussion or press conference, and a longer duration should be given. For example,

*The police **said** it did not appear that anyone else was injured.*

*A Brooklyn woman who was watching her clothes dry in a laundromat was killed Thursday evening when two would-be robbers emptied their pistols into the store, the police **said.***

If the first sentence were quoted text, it would be very much the same. Hence the duration of the "said" event should be short. In the second sentence everything that the spokesperson (here the police) has said is compiled into a single sentence by the reporter, and it is unlikely that the spokesperson said only a single sentence with all this information. Thus, it is reasonable to give longer duration to this "said" event.

Multiple Events: Many occurrences of verbs and other event descriptors refer to multiple events, especially, but not exclusively, if the subject or object of the verb is plural. For example,

*Iraq has **destroyed** its long-range missiles.*

Both single (i.e., destroyed one missile) and aggregate (i.e., destroyed all missiles) events happened. This was a significant source in disagreements in our first round of annotation. Since both judgments provide useful information, our current annotation interface allows the annotator to specify the event as multiple, and give durations for both the single and aggregate events.

Events Involving Negation: Negated events didn't happen, so it may seem strange to specify their duration. But whenever negation is used, there is a certain class of events whose occurrence is being denied. Annotators should consider this class, and make a judgment about the likely duration of the events in it. In addition, there is the interval during which the nonoccurrence of the events holds. For example,

*He was willing to withdraw troops in exchange for guarantees that Israel would not be **attacked**.*

There is the typical amount of time of "being attacked", i.e., the duration of a single attack, and a longer period of time of "*not* being attacked". Similarly to multiple events, annotators are asked to give durations for both the event negated and the negation of that event.

Positive Infinite Durations: These are states which continue essentially forever once they begin. For example,

*He is **dead**.*

Here the time continues for an infinite amount of time, and we allow this as an annotation.

2.3 Inter-Annotator Agreement

Although the graphical output of the annotations enables us to visualize quickly the level of agreement among different annotators for each event, a quantitative measurement of the agreement is needed. The kappa statistic (Krippendorff, 1980; Carletta, 1996) has become the de facto standard to assess inter-annotator agreement. It is computed as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ is the observed agreement among the annotators, and $P(E)$ is the expected agreement,

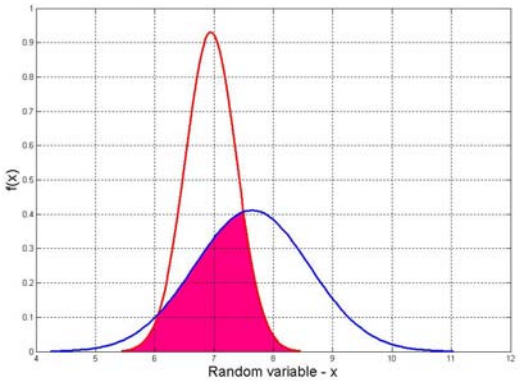


Figure 1: Overlap of Judgments of [10 minutes, 30 minutes] and [10 minutes, 2 hours].

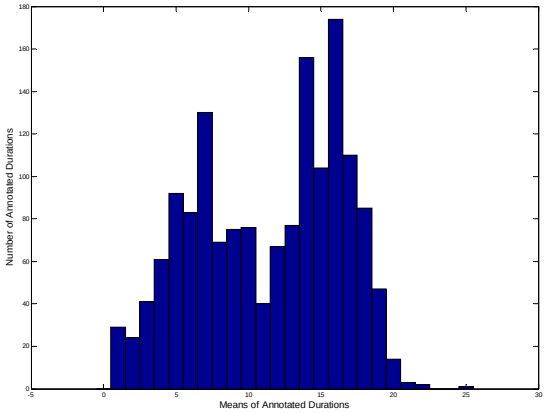


Figure 2: Distribution of Means of Annotated Durations.

which is the probability that the annotators agree by chance.

2.3.1 What Should Count as Agreement?

Determining what should count as agreement is not only important for assessing inter-annotator agreement, but is also crucial for later evaluation of machine learning experiments.

We first need to decide what scale is most appropriate. One possibility is just to convert all the temporal units to seconds. However, this would not correctly capture our intuitions about the relative relations between duration ranges. For example, the difference between 1 second and 20 seconds is significant; while the difference between 1 year 1 second and 1 year 20 seconds is negligible. In order to handle this problem, we use a logarithmic scale for our data. After first converting from temporal units to seconds, we then take the natural logarithms of these values. This logarithmic scale also conforms to the half orders of magnitude (HOM) (Hobbs and Kreinovich, 2001) which was shown to have utility in several very different linguistic contexts.

In the literature on the kappa statistic, most authors address only category data; some can handle more general data, such as data in interval scales or ratio scales (Krippendorff, 1980; Carletta, 1996). However, none of the techniques directly apply to our data, which are ranges of durations from a lower bound to an upper bound.

In fact, what coders were instructed to annotate for a given event is not just a range, but a *duration distribution* for the event, where the area between the lower bound and the upper bound covers about 80% of the entire distribution area. Since it's natural to assume the most likely duration for such distribution is its mean (average) duration, and the distribution flattens out toward the upper and lower bounds, we use the normal or Gaussian distribution to model our duration distributions.

In order to determine a normal distribution, we need to know two parameters: the mean and the standard deviation. For our duration distributions with given lower and upper bounds, the mean is the average of the bounds. Under the assumption that the area between lower and upper bounds covers 80% of the entire distribution area, the lower and upper bounds are each 1.28 standard deviations from the mean.

With this data model, the agreement between two annotations can be defined as the overlapping area between two normal distributions. The agreement among many annotations is the average overlap of all the pairwise overlapping areas. For example, the overlap of judgments of [10 minutes, 30 minutes] and [10 minutes, 2 hours] are as in Figure 1. The overlap or agreement is 0.508706.

2.3.2 Expected Agreement

As in (Krippendorff, 1980), we assume there exists one global distribution for our task (i.e., the duration ranges for all the events), and "chance" annotations would be consistent with this distribution. Thus, the baseline will be an annotator who knows the global distribution and annotates in accordance with it, but does not read the specific article being annotated. Therefore, we must compute the global distribution of the durations, in particular, of their means and their widths. This will be of interest not only in determining expected agreement, but also in terms of what it says about the genre of news articles and about fuzzy judgments in general.

We first compute the distribution of the means of all the annotated durations. Its histogram is shown in Figure 2, where the horizontal axis

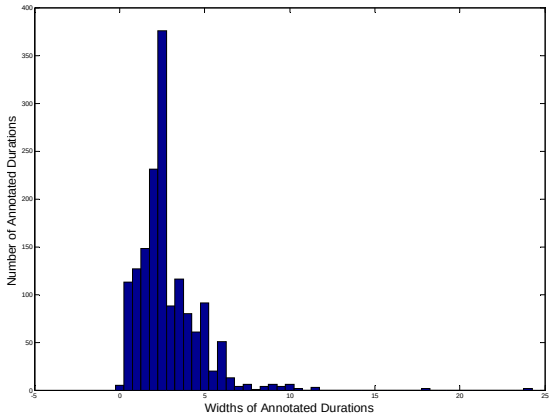


Figure 3: Distribution of Widths of Annotated Durations.

represents the mean values in the natural logarithmic scale and the vertical axis represents the number of annotated durations with that mean.

We also compute the distribution of the widths (i.e., upper bound – lower bound) of all the annotated durations, and its histogram is shown in Figure 3, where the horizontal axis represents the width in the natural logarithmic scale and the vertical axis represents the number of annotated durations with that width.

Two different methods were used to compute the expected agreement (baseline), both yielding nearly equal results. These are described in detail in (Pan et al., 2006a). For both, P(E) is about 0.15.

Experimental results show that the use of the annotation guidelines resulted in about 10% improvement in inter-annotator agreement, measured as described in this section, see (Pan et al., 2006a) for details.

2.4 Machine Learning Experiments

2.4.1 Features

Local Context. For a given event, the local context features include a window of *n* tokens to its left and *n* tokens to its right, as well as the event itself. The best *n* was determined via cross validation. A token can be a word or a punctuation mark. For each token in the local context, including the event itself, three features are included: the original form of the token, its lemma (or root form), and its part-of-speech (POS) tag.

Syntactic Relations. The information in the event's syntactic environment is very important in deciding the durations of events. For a given event, both the head of its subject and the head of its object are extracted from the parse trees generated by the CONTEX parser (Hermjakob and Mooney, 1997). Similarly to the local context features, for both the subject head and the object head, their original form, lemma, and POS tags are extracted as features.

WordNet Hypernyms. Events with the same hypernyms may have similar durations. But closely related events don't always have the same direct hypernyms. We extract the hypernyms not only for the event itself, but also for the subject and object of the event, since events related to a group of people or an organization usually last longer than those involving individuals, and the hypernyms can help distinguish such concepts. For our learning experiments, we extract the first 3 levels of hypernyms from WordNet (Miller, 1990).

2.4.2 Learning Coarse-grained Binary Event Durations

The distribution of the means of the annotated durations in Figure 2 is bimodal, dividing the events into those that take less than a day and those that take more than a day. Thus, in our first machine learning experiment, we have tried to learn this *coarse-grained* event duration information as a binary classification task.

Data. The original annotated data can be straightforwardly transformed for this binary classification task. For each event annotation, the most likely (mean) duration is calculated first by averaging (the logs of) its lower and upper bound durations. If its most likely (mean) duration is less than a day (about 11.4 in the natural logarithmic scale), it is assigned to the "short" event class, otherwise it is assigned to the "long" event class. (Note that these labels are strictly a convenience and not an analysis of the meanings of "short" and "long".)

We divide the total annotated non-WSJ data (2132 event instances) into two data sets: a training data set with 1705 event instances (about 80% of the total non-WSJ data) and a held-out test data set with 427 event instances (about 20% of the total non-WSJ data). The WSJ data (156 event instances) is kept for further test purposes.

Results. The learning results in Figure 4 show that among all three learning algorithms explored (Naïve Bayes (NB), Decision Trees C4.5, and Support Vector Machines (SVM)), SVM with linear kernel achieves the best overall precision (76.6%). Compared with the baseline (59.0%) and human agreement (87.7%), this level of performance is very encouraging, especially as the learning is from such limited training data.

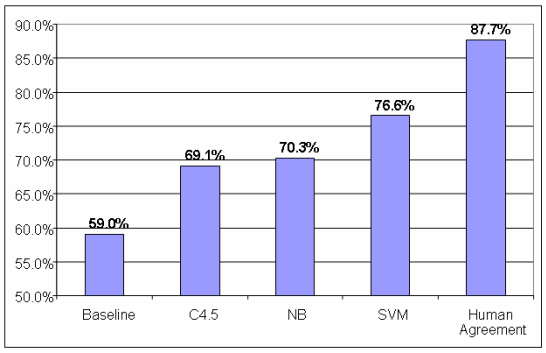


Figure 4: Overall Test Precision on non-WSJ Data.

Feature evaluation in (Pan et al., 2006b) shows that most of the performance comes from event word or phrase itself. A significant improvement above that is due to the addition of information about the subject and object. Local context does not help and in fact may hurt, and hypernym information also does not seem to help. It is gratifying to see that the most important information is that from the predicate and arguments describing the event, as our linguistic intuitions would lead us to expect.

In order to evaluate whether the learned model can perform well on data from different news genres, we tested it on the unseen WSJ data (156 event instances). A precision of 75.0%, which is very close to the test performance on the non-WSJ data, proves the great generalization capacity of the learned model.

Some preliminary experimental results of learning the more fine-grained event duration information, i.e., the most likely temporal unit (cf. (Rieger 1974)'s ORDERHOURS, ORDERDAYS), are shown in (Pan et al., 2006b). SVM again achieves the best performance with 67.9% test precision (baseline 51.5% and human agreement 79.8%) in "approximate agreement" where temporal units are considered to match if they are the same temporal unit or an adjacent one.

3 TimeML and Its Event Classes

TimeML (Pustejovsky et al., 2003) is a rich specification language for event and temporal expressions in natural language text. Unlike most previous attempts at event and temporal specification, TimeML separates the representation of event and temporal expressions from the anchoring or ordering dependencies that may exist in a given text.

TimeML includes four major data structures: EVENT, TIMEX3, SIGNAL, AND LINK. EVENT is a cover term for situations that happen or occur, and also those predicates describing states or circumstances in which something obtains or holds true. TIMEX3, which extends TIMEX2 (Ferro, 2001), is used to mark up explicit temporal expressions, such as time, dates, and durations. SIGNAL is used to annotate sections of text, typically function words that indicate how temporal objects are related to each other (e.g., "when", "during", "before"). The set of LINK tags encode various relations that exist between the temporal elements of a document, including three subtypes: TLINK (temporal links), SLINK (subordination links), and ALINK (aspectual links).

Our event duration annotations can be integrated into the EVENT tag. In TimeML each event belongs to one of the seven event classes, i.e., reporting, perception, aspectual, I-action, I-state, state, occurrence. TimeML annotation guidelines[1] give detailed description for each of the classes:

Reporting. This class describes the action of a person or an organization declaring something, narrating an event, informing about an event, etc (e.g., say, report, tell, explain, state).

Perception. This class includes events involving the physical perception of another event (e.g., see, watch, view, hear).

Aspectual. In languages such as English and French, there is a grammatical device of aspectual predication, which focuses on different facets of event history, i.e., initiation, reinitiation, termination, culmination, continuation (e.g., begin, stop, finish, continue).

I-Action. An I-Action is an Intensional Action. It introduces an event argument (which must be in the text explicitly) describing an action or situation from which we can infer something given its relation with the I-Action (e.g., attempt, try, promise).

I-State. This class of events are similar to the previous class. This class includes states that refer to alternative or possible worlds (e.g., believe, intend, want).

State. This class describes circumstances in which something obtains or holds true (e.g., on board, kidnapped, peace).

Occurrence. This class includes all the many other kinds of events describing something that happens or occurs in the world (e.g., die, crash, build, sell).

[1] http://www.cs.brandeis.edu/~jamesp/arda/time/timeMLdocs/annguide12wp.pdf

4 Integrating Event Duration Annotations into TimeML

Our event duration annotations can be integrated into TimeML by adding two more attributes to the EVENT tag for the lower bound and upper bound duration annotations (e.g., "lowerBoundDuration" and "upperBoundDuration" attributes).

To minimize changes of the existing TimeML specifications caused by the integration, we can try to share as much as possible our event classes as described in Section 2.2 with the existing ones in TimeML as described in Section 3.

We can see that four event classes are shared with very similar definitions, i.e., reporting, aspectual, state, and action/occurrence. For the other three event classes that only belong to TimeML (i.e., perception, I-action, I-state), the I-action and perception classes can be treated as special subclasses of the action/occurrence class, and the I-state class as a special subclass of the state class.

However, there are still three classes that only belong to the event duration annotations (i.e., multiple, negation, and positive infinite). The positive infinite class can be treated as a special subclass of the state class with a special duration annotation for positive infinity.

Each multiple event has two annotations. For example, for

*Iraq has **destroyed** its long-range missiles.*

there is the time it takes to destroy one missile and the duration of the interval in which all the individual events are situated – the time it takes to destroy all its missiles.

Since the single event is usually more likely to be encountered in multiple documents, and thus the duration of the single event is usually more likely to be shared and re-used, to simplify the specification, we can take only the duration annotation of the single events for the multiple event class, and the single event can be assigned with one of the seven TimeML event classes. For example, the "destroyed" event in the above example is assigned with the occurrence class in TimeBank.

The events involving negation can be simplified similarly. Since the event negated is usually more likely to be encountered in multiple documents, we can take only the duration annotation of the negated event for this class. For example, in

*He was willing to withdraw troops in exchange for guarantees that Israel would not be **attacked**.*

the event negated is the "being attacked" event and it is assigned with the occurrence class in TimeBank. Alternatively, TimeML could be extended to treat negations of events as states.

The format used for annotated durations is consistent with that for the value of the DURATION type in TimeML. For example, the sentence

*The official **said** these sites could only be **visited** by a special team of U.N. monitors and diplomats.*

can be marked up in TimeML as:

```
The official <EVENT eid="e63" class="REPORTING"> said </EVENT> these sites <SIGNAL sid="s65" >could</SIGNAL> only be <EVENT eid="e64" class="OCCURRENCE"> visited </EVENT> by a special team of <ENAMEX TYPE="ORGANIZATION"> U.N. </ENAMEX> monitors and diplomats.
```

If we annotate the "said" event with the duration annotation of [5 seconds, 5 minutes], and the "visited" event with [10 minutes, 1 day], the extended mark-up becomes:

```
The official <EVENT eid="e63" class="REPORTING" lowerBoundDuration="PT5S" upperBoundDuration="PT5M"> said </EVENT> these sites <SIGNAL sid="s65" >could</SIGNAL> only be <EVENT eid="e64" class="OCCURRENCE" lowerBoundDuration="PT10M" upperBoundDuration="P1D"> visited </EVENT> by a special team of <ENAMEX TYPE="ORGANIZATION"> U.N. </ENAMEX> monitors and diplomats.
```

5 Conclusion

In this paper we have demonstrated how to extend TimeML with typical durations of events. We can see that the extension is very straightforward. Other interesting temporal information can be extracted or learned. For example, for each event class, we can generate its own mean and widths graphs, and learn their durations separately from other classes, which may capture different duration characteristics associated with each event class.

Acknowledgments

This work was supported by the Advanced Research and Development Activity (ARDA), now the Disruptive Technology Office (DTO), under DOD/DOI/ARDA Contract No. NBCHC040027. The authors have profited from discussions with Hoa Trang Dang, Donghui Feng, Kevin Knight, Daniel Marcu, James Pustejovsky, Deepak Ravichandran, and Nathan Sobo.

References

B. Boguraev and R. K. Ando. 2005. TimeML-Compliant Text Analysis for Temporal Reasoning. In *Proceedings of the International Joint Conferenceon Artificial Intelligence (IJCAI)*.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Lingustics*, 22(2):249–254.

D. R. Dowty. 1979. *Word Meaning and Montague Grammar*, Dordrecht, Reidel.

L. Ferro. 2001. Instruction Manual for the Annotation of Temporal Expressions. Mitre Technical Report MTR 01W0000046, *the MITRE Corporation*, McLean, Virginia.

E. Filatova and E. Hovy. 2001. Assigning Time-Stamps to Event-Clauses. *Proceedings of ACL Workshop on Temporal and Spatial Reasoning*.

S. Harabagiu and C. Bejan. 2005. Question Answering Based on Temporal Inference. In *Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering*, Pittsburgh, PA.

U. Hermjakob and R. J. Mooney. 1997. Learning Parse and Translation Decisions from Examples with Rich Context. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*.

J. Hitzeman, M. Moens, and C. Grover. 1995. Algorithms for Analyzing the Temporal Structure of Discourse. In *Proceedings of EACL*. Dublin, Ireland.

J. R. Hobbs and V. Kreinovich. 2001. Optimal Choice of Granularity in Commonsense Estimation: Why Half Orders of Magnitude, In *Proceedings of Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, Vacouver, British Columbia.

K. Krippendorf. 1980. *Content Analysis: An introduction to its methodology*. Sage Publications.

I. Mani and G. Wilson. 2000. Robust Temporal Processing of News. In *Proceedings of Annual Conference of the Association for Computational Linguistics (ACL)*.

I. Mani and B. Schiffman. 2005. Temporally Anchoring and Ordering Events in News. In J. Pustejovsky and R. Gaizauskas ed. *Time and Event Recognition in Natural Language*. John Benjamins.

G. A. Miller. 1990. WordNet: an On-line Lexical Database. *International Journal of Lexicography* 3(4).

M. Moens and M. Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics* 14(2): 15-28.

D. Moldovan, C. Clark, and S. Harabagiu. 2005. Temporal Context Representation and Reasoning. In *Proceedings of the International Joint Conferenceon Artificial Intelligence (IJCAI)*.

F. Pan, R. Mulkar, and J. R. Hobbs. 2006a. An Annotated Corpus of Typical Durations of Events. To appear in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

F. Pan, R. Mulkar, and J. R. Hobbs. 2006b. Learning Event Durations from Event Descriptions. To appear in *Proceedings of the 44th Conference of the Association for Computational Linguistics (COLING-ACL)*, Sydney, Australia.

R. J. Passonneau. 1988. A Computational Model of the Semantics of Tense and Aspect. *Computational Linguistics* 14:2.44-60.

J. Pustejovsky, J. Castano, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the AAAI Spring Symposium on New Directions in Question-Answering*.

C. J. Rieger. 1974. Conceptual memory: A theory and computer program for processing and meaning content of natural language utterances. *Stanford AIM-233*.

R. Saurí, R. Knippen, M. Verhagen and J. Pustejovsky. 2005. Evita: A Robust Event Recognizer for QA Systems. In *Proceedings of HLT/EMNLP*.

M. Surdeanu, S. Harabagiu, J. Williams, and P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41th Annual Conference of the Association for Computational Linguistics (ACL-03)*, pages 8–15.

Z. Vendler. 1967. *Linguistics in Philosophy*, Ithaca, Cornell University Press.