Annotating and Learning Event Durations in Text

Feng Pan[*]
Bing, Microsoft Corporation

Rutu Mulkar-Mehta[**]
Information Sciences Institute (ISI),
University of Southern California

Jerry R. Hobbs[†]
Information Sciences Institute (ISI),
University of Southern California

This article presents our work on constructing a corpus of news articles in which events are annotated for estimated bounds on their duration, and automatically learning from this corpus. We describe the annotation guidelines, the event classes we categorized to reduce gross discrepancies in inter-annotator judgments, and our use of normal distributions to model vague and implicit temporal information and to measure inter-annotator agreement for these event duration distributions. We then show that machine learning techniques applied to this data can produce coarse-grained event duration information automatically, considerably outperforming a baseline and approaching human performance. The methods described here should be applicable to other kinds of vague but substantive information in texts.

1. Introduction

Consider the sentence from a news article:

> *George W. Bush met with Vladimir Putin in Moscow.*

How long did the meeting last? Our first inclination is to say we have no idea. But in fact we do have some idea. We know the meeting lasted more than ten seconds and less than one year. As we guess narrower and narrower bounds, our chances of being correct go down, but if we are correct, the utility of the information goes up. Just how accurately can we make duration judgments like this? How much agreement can we expect among people? Will it be possible to extract this kind of information from text automatically?

[*] Microsoft Corporation, 475 Brannan St., San Francisco, CA 94107, USA.
E-mail: `fengpan@microsoft.com`.

[**] 4676 Admiralty Way, Marina del Rey, CA 90292, USA. E-mail: `me@rutumulkar.com`.

[†] 4676 Admiralty Way, Marina del Rey, CA 90292, USA. E-mail: `hobbs@isi.edu`.

Submission received: 2 November 2006; revised submission received: 26 January 2011; accepted for publication: 7 March 2011.

Sometimes we are explicitly told the duration of events, as in "a five-day meeting" and "I have lived here for three years." But more often, such phrases are missing, and present-day natural language applications simply have to proceed without them.

There has been a great deal of work on formalizing temporal information (Allen 1984; Moens and Steedman 1988; Zhou and Fikes 2002; Han and Lavie 2004; Hobbs and Pan 2004) and on temporal anchoring and event ordering in text (Hitzeman, Moens, and Grover 1995; Mani and Wilson 2000; Filatova and Hovy 2001; Boguraev and Ando 2005; Mani et al. 2006; Lapata and Lascarides 2006). The uncertainty of temporal durations has been recognized as one of the most significant issues for temporal reasoning (Allen and Ferguson 1994). Chittaro and Montanari (2000) point out by way of example that we have to know how long a battery remains charged to decide when to replace it or to predict the effects of actions which refer to the battery charge as a precondition.

Yet to our knowledge, there has been no serious published empirical effort to model and learn the vague and implicit duration information in natural language, and to perform reasoning over this information. Cyc has some fuzzy duration information, although it is not generally available; Rieger (1974) discusses the issue for less than a page; there has been work in fuzzy logic on representing and reasoning with imprecise durations (Godo and Vila 1995; Fortemps 1997). But none of these efforts make an attempt to collect human judgments on such durations or to extract them automatically from text.

Nevertheless, people have little trouble exploiting temporal information implicitly encoded in the descriptions of events, relying on their knowledge of the range of usual durations of types of events. This hitherto largely unexploited information is part of our commonsense knowledge. We can estimate roughly how long events of different types last and roughly how long situations of various sorts persist. We know that government policies typically last somewhere between one and ten years, and weather conditions fairly reliably persist between three hours and one day. We are often able to decide whether two events overlap or are in sequence by accessing this information. We know that 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.

This article describes an exploration into how this information can be captured automatically. Our results can have a significant impact on computational linguistics applications like event anchoring and ordering in text (Mani and Schiffman 2007), event coreference (Bejan and Harabagiu 2010), question answering (Tao et al. 2010; Harabagiu and Bejan 2005), and other intelligent systems that would benefit from such temporal commonsense knowledge, for example, temporal reasoning (Zhou and Hripcsak 2007).

Our goal is to be able to extract this implicit event duration information from text automatically, and to that end we first annotated the events in news articles with bounds on their durations. The corpus that we have annotated currently contains all 48 non-Wall-Street-Journal (non-WSJ) news articles (2,132 event instances), as well as 10 WSJ articles (156 event instances), from the TimeBank corpus annotated in TimeML (Pustejovsky et al. 2003). The non-WSJ articles (mainly political and disaster news) include both print and broadcast news that are from a variety of news sources, such as ABC, AP, CNN, and VOA. All the annotated data have already been integrated into the TimeBank corpus.[1]

This article is organized as follows. In Section 2 we describe our annotation guidelines, including the annotation strategy and assumptions, and the representative event

[1] The annotated data are available at `http://www.isi.edu/~hobbs/EventDuration/annotations`.

classes we have categorized to minimize discrepant judgments between annotators. The method for measuring inter-annotator agreement when the judgments are intervals on a scale is described in Section 3. We will discuss how to integrate our event duration annotations to TimeML in Section 4. In Section 5 we show that machine learning techniques applied to the annotated data considerably outperform a baseline and approach human performance.

2. Annotation Guidelines and Event Classes

Every event to be annotated was already identified in the TimeBank corpus. In our project, annotators were 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 1 to 10. 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, Figure 1 shows the output of the annotations (three annotators) for the "finished" event in the sentence:

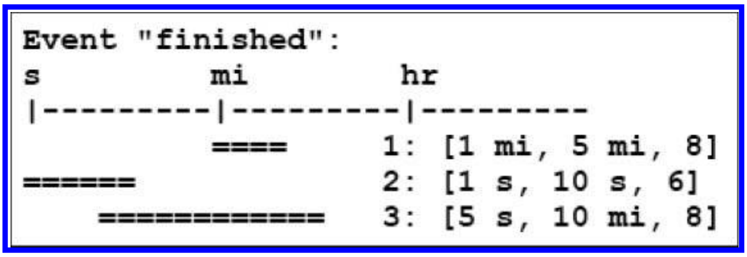
> *After the victim, Linda Sanders, 35, had* **finished** *her cleaning and was waiting for her clothes to dry, ...*

Figure 1 shows that the first annotator believed that the event lasts for minutes whereas the second annotator believed it could only last for several seconds. The third annotated the event as ranging from a few seconds to a few minutes. The confidence level of the annotators is generally subjective but as all three were higher than 5, it shows reasonable confidence. A logarithmic scale is used for the output (see Section 3.1 for details).

2.1 Annotation Instructions

Annotators were asked to make their judgments as intended readers of the article, using whatever world knowledge was relevant to an understanding of the article. They were asked to identify upper and lower bounds that would include 80% of the possible cases. For example, rainstorms of 10 seconds or of 40 days and 40 nights might occur, but they are clearly anomalous and should be excluded. There are two strategies for considering the range of possibilities:

1. Pick the most probable scenario, and annotate its upper and lower bounds.



```
Event "finished":
s          mi          hr
|---------|---------|---------
           ====      1: [1 mi, 5 mi, 8]
======               2: [1 s, 10 s, 6]
    ============     3: [5 s, 10 mi, 8]
```

Figure 1
Example annotation output by three annotators.

2. Pick the set of probable scenarios, and annotate the bounds of their upper and lower bounds.

We deemed the second to be the preferred strategy.

The judgments were to be made in context. First of all, information in the syntactic environment needed to be considered before annotating. For example, there is a difference in the duration of the watching events in the phrases **watch** *a movie* and **watch** *a bird fly*.

Moreover, the events needed to be annotated in light of the information provided by the entire article. This meant annotators were to read the entire article before starting to annotate. One may learn in the last paragraph, for example, that the demonstration event mentioned in the first paragraph lasted for three days, and that information was to be used for annotation.

However, they were not to use knowledge of the future when annotating a historical article. For example, an article from the fall of 1990 may talk about the coming war against Iraq. Today we know exactly how long that lasted. But annotators were asked to try to put themselves in the shoes of the 1990 readers of that article, and make their judgments accordingly. This was because we wanted people's estimates of typical durations of events, rather than the exact durations.

Annotation could be made easier and more consistent if *coreferential* and *nearcoreferential* descriptions of events were identified initially. Annotators were asked to give the same duration ranges for such cases. For example, in the sentence *during the* **demonstration**, *people* **chanted** *antigovernment slogans*, annotators were to give the same durations for the "demonstration" and "chanted" events.

2.2 Analysis

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; we found that the vast majority of radically different judgments could be categorized into a relatively small number of classes, however. Some of these correspond to aspectual features of events, which have been investigated intensively (e.g., Vendler 1967; Dowty, 1979; Moens and Steedman 1988; Passonneau 1988; Giorgi and Pianesi 1997; Madden and Zwaan 2003; Smith 2005). We then developed guidelines to make annotators aware of these cases and to guide them in making the judgments (see the next section). There is a residual of gross discrepancies in annotators' judgments that result from differences of opinion, for example, about how long a government policy is typically in effect. But the number of these discrepancies was surprisingly small.

These guidelines were then used in the annotation of a test set. It was shown that the agreement in the test set was greater than the agreement obtained when annotations were performed without the guidelines. (See Section 3.3 for the experimental results.)

2.3 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, in

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

does the word "injured" describe 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. In the sentence,

> *Farkas was ordered home and* **retired**.

although *retired* usually indicates a state, it looks here more like the action by his company of retiring Farkas.

There are some general diagnostic tests to distinguish actions and states (Vendler 1967; Dowty 1979); for example, action verbs are fine in the progressive form but progressives of stative verbs are usually odd. Another test can be applied to this specific case: Imagine someone says the sentence after the action has ended but the state is 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." Similarly, we would say "Farkas *was* retired" rather than "Farkas *is* retired."

Our annotation interface handles events of this type by allowing the annotators to specify which interpretation they are 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, in

> *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), because 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 if we simply assume such events are instantaneous.

In the following example:

> *General Abacha's supporters* **began** *a two-day rally in the capital.*

the gathering of people marks the "beginning" of the rally, and it generally takes time for a crowd of people to get together to start a rally.

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 duration can easily be verified by saying the sentence out loud and timing it. For example, in

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

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

- **Unquoted Report:** When the reporting description occurs without quotes, the report could be as short as 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, consider

> *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, in

> *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. Because 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.

In the following example:

> *Seventy-five million copies of the rifle have been* **built** *since it entered* **production** *in February 1947.*

"built" and "production" are both multiple events. The annotators were asked to give durations for both the single (i.e., built/produce one rifle) and aggregate (i.e., built/produce 75 million copies of the rifle) 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, in

> *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," namely, the duration of a single attack, and a longer period of time of "not being attacked." The first is probably from seconds to minutes and the second from months to years. Similarly to multiple events, annotators were asked to give durations for both the event negated and the negation of that event.

Appearance Events. Verbs like "seem" and "appear" usually indicate appearance events. The duration of this kind of event depends on the duration of the validity or availability of the evidence that causes one to have some impression at the time. Such an event begins when enough evidence has accumulated for one to make that guess or judgment, and it ends when either the evidence is contradicted or certainty is achieved. For example, in

> *It* **appears** *that the destruction of this city in 2700 B.C. was related to the eruption of the volcano.*

the "appears" event lasts from when the archaeologist discovers enough evidence to make the conjecture until the time the conjecture is refuted or confirmed.

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

> *He is* **dead**.

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

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; Siegel and Castellan 1988; Carletta 1996; Di Eugenio and Glass 2004), which factors out the agreement that is expected by chance, has become the de facto standard to assess inter-annotator agreement. It is computed as follows:

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

$P(A)$ is the observed agreement among the annotators, and $P(E)$ is the expected agreement, which is the probability that the annotators agree by chance.

In order to compute the kappa statistic for our task, we have to compute $P(A)$ and $P(E)$ first. But those computations are not straightforward.

$P(A)$: What should count as agreement among annotators for our task?

$P(E)$: What is the probability that the annotators agree by chance for our task?

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. For example, for a given event with a known gold-standard duration range from 1 hour to 4 hours, if a machine learning program outputs a duration of 3 hours to 5 hours, how should we evaluate this result?

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, whereas the difference between 1 year 1 second and 1 year 20 seconds is negligible. Consider the range from 1 year to 5 years and the range from 1 second to 5 seconds. The distance between 1 year and 5 years in seconds would be much larger than that between 1 second and 5 seconds, but intuitively, they represent the same level of uncertainty. 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 use of a logarithmic scale also conforms to the idea of the importance of half orders of magnitude (HOM) (Hobbs 2000; Hobbs and Kreinovich 2001), which has been shown to have utility in commonsense reasoning and in several very different linguistic contexts.

In the literature on the kappa statistic, most authors address only category data (either in nominal scales or ordinal scales); 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 involves a range of durations from a lower bound to an upper bound.

In fact, what coders 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. It is natural to assume that the most likely duration in such a distribution is the mean or average duration, and that the distribution flattens out toward the upper and lower bounds. Thus, we use the normal or Gaussian distribution to model the distribution of possible durations.

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. Then the standard deviation can be computed using either the upper bound (X_{upper}) or the lower bound (X_{lower}) as follows:

$$\sigma = \frac{X_{upper} - \mu}{1.28} = \frac{X_{lower} - \mu}{-1.28}, \text{ where } \mu = \frac{X_{upper} + X_{lower}}{2} \tag{2}$$

With this data model, the agreement between two annotations can be defined as the overlapping area between two normal distributions.[2] The agreement among many annotations is the average overlap of all the pairwise overlapping areas. For example, for a given event, suppose the two annotations are:

1. Lower: 10 minutes; upper: 30 minutes
2. Lower: 10 minutes; upper 2 hours

After converting to seconds and to the natural logarithmic scale, they become:

1. Lower: 6.39692; upper: 7.49554
2. Lower: 6.39692; upper: 8.88184

We then compute their means and standard deviations:

1. $\mu_1 = 6.94623$; $\sigma_1 = 0.42861$
2. $\mu_2 = 7.63938$; $\sigma_2 = 0.96945$

The distributions and their overlap are then as in Figure 2. The overlap or agreement ($P(A)$) is 0.508706.

3.2 Expected Agreement

What is the probability that the annotators agree by chance for our task? The first quick response to this question may be 0, if we consider all the possible durations from 1 second to 1,000 years or even positive infinity.

However, not all the durations are equally possible. As in Krippendorff (1980) and Siegel and Castellan (1988), 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 3, where the horizontal axis represents the mean values in the natural logarithmic scale and the vertical axis represents the number of annotated durations with that mean.

There are two peaks in this distribution. One is from 5 to 7 in the natural logarithmic scale, which corresponds to about 1.5 minutes to 30 minutes. The other is from 14 to 17 in the natural logarithmic scale, which corresponds to about 8 days to 6 months. One could speculate that this bimodal distribution is because daily newspapers report short events that happened the day before and place them in the context of larger trends. The lowest point between the two peaks occurs at 11, which roughly corresponds to one day.

[2] This idea is due to Hoa Trang Dang.

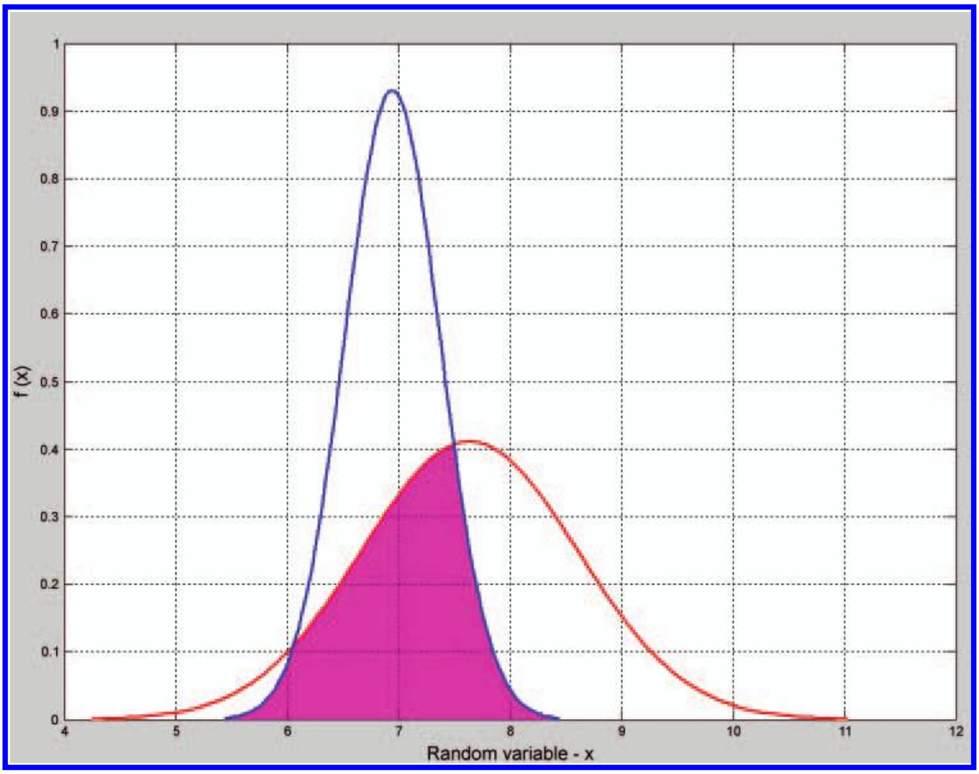


Figure 2
Overlap of judgments of [10 minutes, 30 minutes] and [10 minutes, 2 hours].

We also compute the distribution of the widths (i.e., $X_{upper} - X_{lower}$) of all the annotated durations, and its histogram is shown in Figure 4, 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.

The peak of this distribution occurs at 2.5 in the natural logarithmic scale. This shows that for annotated durations, the most likely uncertainty factor from a mean or average duration is 3.5:

$$\frac{X_{upper}}{\mu} = \frac{\mu}{X_{lower}} = e^{1.25} = 3.5 \tag{3}$$

because

$$\log(X_{upper}) - \log(\mu) = \log(\frac{X_{upper}}{\mu}) = 2.5/2 = 1.25 \tag{4}$$

This is the half orders of magnitude factor that Hobbs and Kreinovich (2001) argue gives the optimal granularity; making something three to four times bigger changes the way we interact with it.

Because the global distribution is determined by these mean and width distributions, we can then compute the expected agreement, that is, the probability that the annotators agree by chance, where the chance is based on this global distribution. Two approaches were used to approximate this probability, both of which use a normal distribution to approximate the global distribution.

The first approach is to compute a fixed global normal distribution with the mean as the mean of the mean distribution and the standard deviation as the mean standard deviation (this can be straightforwardly computed from the width distribution). We then compute the expected agreement by averaging all the agreement scores (overlaps) between this fixed distribution and each of the annotated duration distributions.

The second approach is to generate 1,000 normal distributions whose means are randomly generated from the mean distribution and standard deviations are randomly computed from the width distribution. We then compute the expected agreement by averaging all the agreement scores (overlaps) between these 1,000 random distributions.

In a sense, both of these capture the way an annotator might annotate if he or she did not read the article but only guessed on the basis of the global distribution. As it turns out, the results of the two approaches of computing the expected agreement are very close; they differ by less than 0.01: $P(E)_1 = 0.1439$, $P(E)_2 = 0.1530$. We will use the results of the second approach as the baseline in the next section.

3.3 Inter-Annotator Agreement Experiments

In order to see how effective our guidelines are, we conducted experiments to compare the inter-annotator agreement *before* and *after* annotators read the guidelines.

The data for the evaluation was split into two sets. The first set contained 13 articles (521 events, 1,563 annotated durations) which were all political and disaster news stories from ABC, APW, CNN, PRI, and VOA. The annotators annotated independently

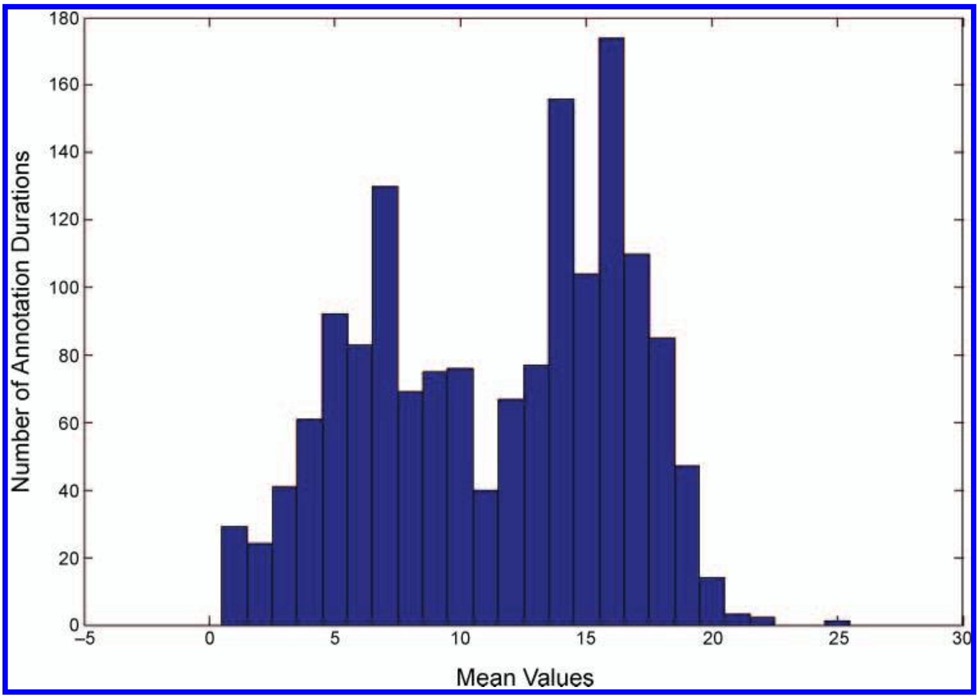


Figure 3
Distribution of means of annotated durations.

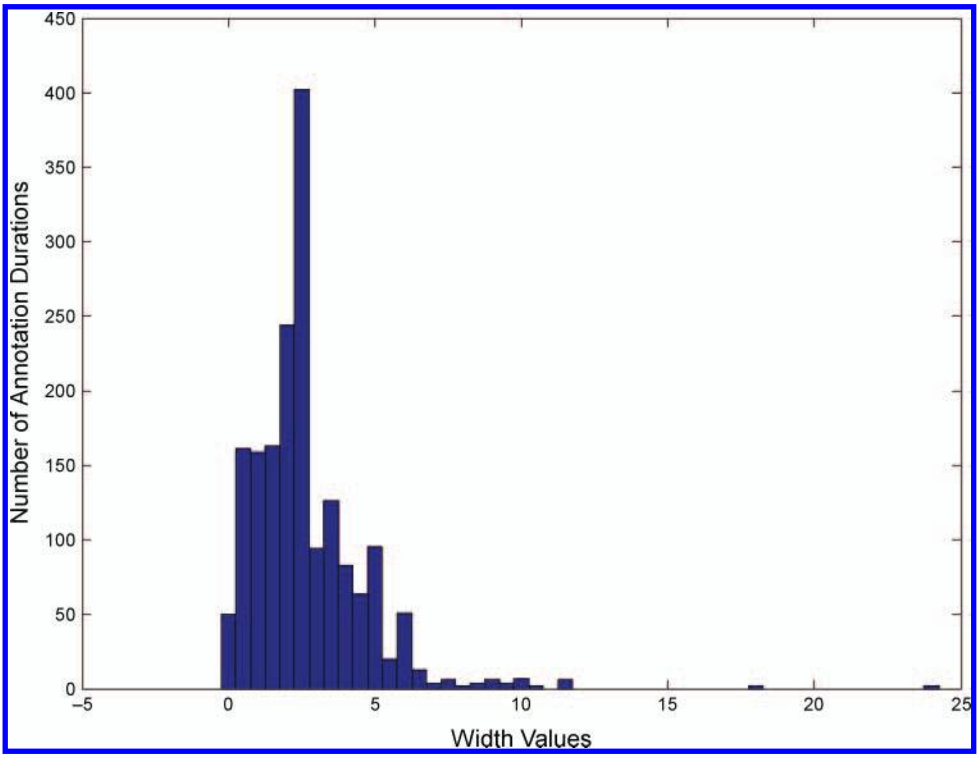


Figure 4
Distribution of widths of annotated durations.

before reading the guidelines. The annotators were only given short instructions on what to annotate and one sample article with annotations. The second set (test set) contained 5 articles (125 events, 375 annotated durations) that were also political and disaster news stories from the same news sources. The annotators annotated independently *after* reading the guidelines.

The comparison is shown in Figure 5. Agreement is measured by the area of overlap in two distributions and is thus a number between 0 and 1. The graphs show the answer to the question "If we set the threshold for agreement at x, counting everything above x as agreement, what is the percentage y of inter-annotator agreement?" The horizontal axis represents the overlap thresholds, and the vertical axis represents the agreement percentage, that is, the percentage of annotated durations that agree for given overlap thresholds. There are three lines in the graph. The top one (with circles) represents the after-guidelines agreement; the middle one (with triangles) represents the before-guidelines agreement; and the lowest one (with squares) represents the expected or baseline agreement. This graph shows that, for example, if we define agreement to be a 10% overlap or better (an overlap threshold of 0.1), we can get 0.8 agreement after reading the guidelines, 0.72 agreement before reading the guidelines, and 0.36 expected agreement with only the knowledge of the global distribution. From this graph, we can see that our guidelines are indeed effective in improving the inter-annotator agreement.

Table 1 shows more detailed experimental results. For each overlap threshold, it shows the expected or baseline agreement, the before-guidelines agreement, and the after-guidelines agreement, as well as the kappa statistic computed from the after-

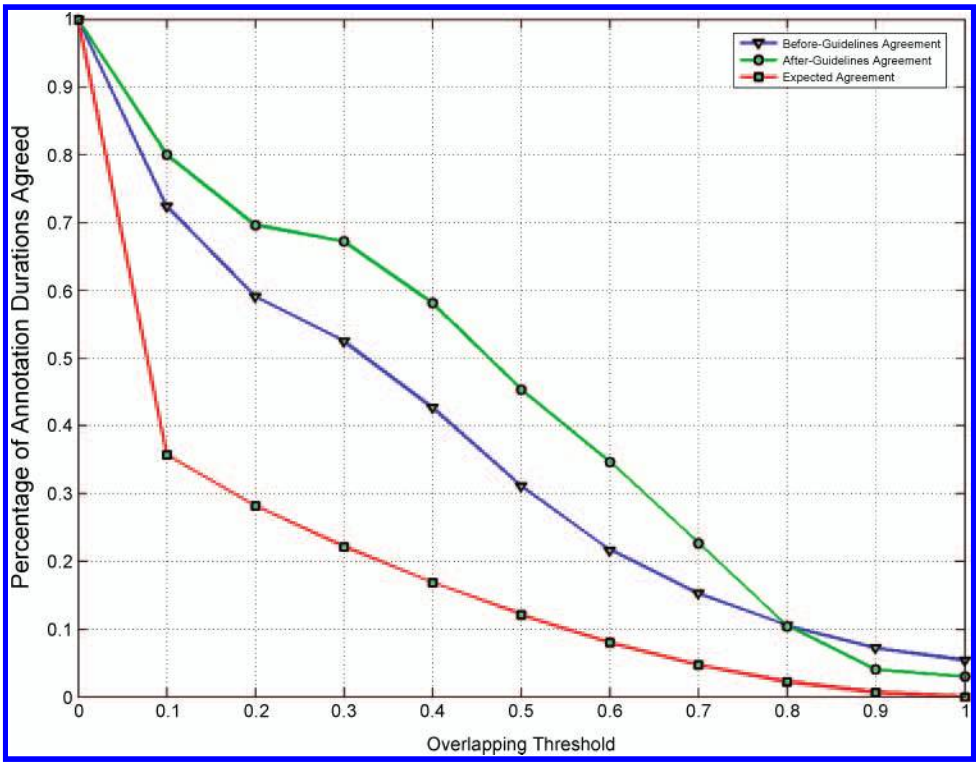


Figure 5
Inter-annotator agreement: Expected, before-guidelines, and after-guidelines.

guidelines agreement ($P(A)$) and the expected or baseline agreement ($P(E)$). The agreement actually gets marginally worse when the agreement criteria is very stringent (i.e., overlap ≥ 0.9), which indicates there really is no consensus at that level of agreement. The overall agreement is relatively low. Thus in this article, we mainly focus on learning coarse-grained event durations with much higher inter-annotator agreement. See Sections 5.2 and 5.3 for more details.

Table 1
Inter-annotator agreement with different overlap thresholds.

Overlap Threshold	Expected Agreement	BeforeG. Agreement	AfterG. Agreement	Kappa (AfterG. A.)
0.1	0.36	0.72	0.80	0.69
0.2	0.28	0.59	0.70	0.58
0.3	0.22	0.52	0.67	0.58
0.4	0.17	0.43	0.58	0.49
0.5	0.12	0.31	0.45	0.38
0.6	0.08	0.22	0.35	0.29
0.7	0.05	0.15	0.23	0.19
0.8	0.02	0.10	0.10	0.08
0.9	0.01	0.07	0.04	0.03
1.0	0.00	0.05	0.03	0.03

4. Extending TimeML with Estimated Event Durations

This section describes the event classes in TimeML and how we can integrate our annotations of estimated event durations with them. This can enrich the expressiveness of TimeML, and provide natural language applications that use TimeML with this additional implicit event duration information for temporal reasoning.

4.1 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, namely, reporting, perception, aspectual, I-action, I-state, state, and occurrence. The TimeML annotation guidelines[3] give detailed descriptions 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, and so forth (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. This class focuses on different facets of event history, that is, 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 intensional action or situation which does not necessarily actually happen but may be only desired or possible (e.g., attempt, try, promise).

I-State. This class of events is similar to the previous class. It includes states that refer to alternative 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).

[3] http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.pdf.

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).

4.2 Integrating Event Duration Annotations

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 to the existing TimeML specifications caused by the integration, we can try to share as much as possible our event classes as described in Section 2.3 with the existing ones in TimeML.

We can see that four event classes are shared with very similar definitions: reporting, aspectual, state, and action/occurrence. For the other three event classes that only belong to TimeML (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.

There are still three classes that only belong to the event duration annotations (i.e., multiple, negation, and positive infinite), however. 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, one for single events and the other for aggregate events. Because 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 earlier example is assigned with the occurrence class in TimeBank.

The events involving negation can be simplified similarly. Because 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.

4.3 Annotation Consistency Evaluation

After the two corpora are integrated, it would be useful to evaluate how consistent the temporal relationship annotations are originally in TimeBank and in the newly integrated event duration annotations. Because not all the events in TimeBank are anchored exactly on a time line, for this consistency evaluation we have only evaluated the "includes" / " is included" TLINK relationship: If event A includes event B, the duration of event A should be no shorter than the duration of event B. Because the event duration annotation is a range, there are three possible relationships between the two duration annotations for event A $[a_1, a_2]$ and event B $[b_1, b_2]$:

1. A strictly includes B if $a_1 \geq b_2$ (strictly compatible)
2. A possibly includes B if $a_1 \leq b_2$, but $a_2 \geq b_1$(softly compatible)
3. A doesn't include B if $a_2 < b_1$(incompatible)

We call the case (i) strictly compatible, (ii) softly compatible, and (iii) incompatible.

For this consistency evaluation, one article was randomly picked from each news source, and for each "include" TLINK relationship in the article, one of the three compatibility labels is assigned based on their definitions. The result shows that out of a total of 116 "include" relationships, 59.5% are strictly compatible, 19.8% are softly compatible, and 20.7% are incompatible. We can merge the first two categories as compatible, which accounts for 79.3%.

Most of the incompatible cases are due to different event interpretations and guidelines for the two corpora. For example, TimeBank bounds most of the I-State events to the article time, whereas our annotations usually give them a much longer duration—for example, the event "appear" in "*Everyone* **appears** *to believe that somehow Cuba is going to change*," and the event "hope" in "*The quarantine* **hopes** *to staunch the flow of Iraqi oil*."

Aspectual events are another class of events that cause many incompatible cases, including those where multiple interpretations are possible, for example, in "*This is quite an extraordinary story* **unfolding** *here*," the event "unfolding" can be interpreted as either the start of the unfolding (TimeBank), or the entire process of the unfolding (duration annotation); Sometimes even when both corpora agree on the interpretation of the event, they may not agree on its duration, for example, in "*But with the task-force investigation just* **getting** *under way, officials have been careful not to draw any firm conclusions*," it is clear that the "getting" event is the start of the investigation, but should it last momentarily (TimeBank) or for a couple of weeks (duration annotation)? Despite the difficulty with this event class, there exists some clear cases, for example, in "*A new Essex County task force* **began delving** *Thursday into the slayings of 14 black women over the last five years in the Newark area*," it is correct to bound "began" to Thursday, whereas the event "delving" should last much longer.

5. Learning Event Durations

It is highly unlikely that the relatively small amount of data we have annotated so far could support an automatic classification task at this fine granularity. But we have identified two classification tasks at a coarser granularity that we can hope to do well on and that have some independent utility. The first exploits the distribution shown in Figure 3, a bimodal distribution of events classifying them into those lasting less than a day and those lasting more than a day. The second coarse-grained task is the approximate identification of the temporal unit most likely to be used to describe the duration of the event.

In Section 5.1 we describe the features that were used in the machine learning experiments. Section 5.2 describes the experiment on classifying events into those lasting more or less than a day. Section 5.3 describes the experiment on identifying the most appropriate temporal unit for the mean durations.

5.1 Features

In this section, we describe the lexical, syntactic, and semantic features that we considered in learning event durations.

5.1.1 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, for $n = \{0, 1, 2, 3\}$. The best n determined via cross validation turned out to be 0, that is, the event itself with no local context. But we also present results for $n = 2$ in Section 5.2.3 to evaluate the utility of local context.

A token can be a word or a punctuation mark. Punctuation marks are not removed, because they can be indicative features for learning event durations. For example, the quotation mark is a good indication of quoted reporting events, and the duration of such events most likely lasts for seconds or minutes, depending on the length of the quoted content. However, there are also cases where quotation marks are used for other purposes, such as for titles of artistic works.

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. The lemma of the token is extracted from parse trees generated by the CONTEX parser (Hermjakob and Mooney 1997), which includes rich context information in parse trees, and the Brill tagger (Brill 1992) is used for POS tagging.

The local context features extracted for the "signed" event in the sentence below is shown in Table 2 (with a window size $n = 2$). The feature vector is [signed, sign, VBD, the, the, DT, plan, plan, NN, Friday, Friday, NNP, on, on, IN].

The two presidents on Friday **signed** *the plan.*

5.1.2 Syntactic Relations. The information in the event's syntactic environment is very important in deciding the durations of events. For example, there is a difference in the durations of the "watch" events in the phrases "**watch** *a movie*" and "**watch** *a bird fly*."

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. 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. When there is no subject or object for an event, "NULL" is used for the feature values.

For the "signed" event in *The two presidents on Friday* **signed** *the plan*, the head of its subject is "presidents" and the head of its object is "plan." The extracted syntactic relation features are shown in Table 3, and the feature vector is [presidents, president, NNS, plan, plan, NN].

5.1.3 WordNet Hypernyms. Events with the same hypernyms may have similar durations. For example, events "ask" and "talk" both have a direct WordNet (Miller et al. 1990) hypernym of "communicate," and most of the time they do have very similar durations in the corpus.

However, closely related events don't always have the same direct hypernyms. For example, "see" has a direct hypernym of "perceive," whereas for "observe" one

Table 2
Local context features for the "signed" event with $n = 2$ in "The two presidents on Friday **signed** *the plan*."

Features	Original	Lemma	POS
Event	signed	sign	VBD
1token-after	the	the	DT
2token-after	plan	plan	NN
1token-before	Friday	Friday	NNP
2token-before	on	on	IN

Table 3
Syntactic relation features for the "signed" event in "The two presidents on Friday **signed** *the plan*."

Features	**Original**	**Lemma**	**POS**
Subject	presidents	president	NNS
Object	plan	plan	NN

needs to go two steps up through the hypernym hierarchy before reaching "perceive." Correlation between events may be lost if only the direct hypernyms of the words are extracted.

It is useful to extract the hypernyms not only for the event itself, but also for the subject and object of the event. For example, 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. The direct hypernyms of nouns are not always general enough for this purpose, but a hypernym at too high a level can be too general to be useful. For our learning experiments, we use the first three levels of hypernyms from WordNet.

Hypernyms are only used for the events and their subjects and objects, not for the local context words. For each level of hypernyms in the hierarchy, it's possible to have more than one hypernym, for example, "see" has two direct hypernyms, "perceive" and "comprehend." For a given word, it may also have more than one sense in WordNet. In such cases, as in Gildea and Jurafsky (2002), we only take the first sense of the word and the first hypernym listed for each level of the hierarchy. A word disambiguation module might improve the learning performance. But because the features we need are the hypernyms, not the word sense itself, even if the first word sense is not the correct one, its hypernyms can still be good enough in many cases. For example, in one news article, the word "controller" refers to an air traffic controller, which corresponds to the second sense in WordNet, but its first sense (business controller) has the same hypernym of "person" (three levels up) as the second sense (direct hypernym). Because we take the first three levels of hypernyms, the correct hypernym is still extracted.

When there are fewer than three levels of hypernyms for a given word, its hypernym on the previous level is used. When there is no hypernym for a given word (e.g., "go"), the word itself will be used as its hypernyms. Because WordNet only provides hypernyms for nouns and verbs, "NULL" is used for the feature values for a word that is not a noun or a verb.

For the "signed" event in *The two presidents on Friday* **signed** *the plan*, the extracted WordNet hypernym features for the event ("signed"), its subject ("presidents"), and its object ("plan") are shown in Table 4, and the feature vector is [write, communicate, interact, corporate_executive, executive, administrator, idea, content, cognition].

Table 4
WordNet hypernym features for the event ("signed"), its subject ("presidents"), and its object ("plan") in *The two presidents on Friday* **signed** *the plan*.

Feature	1-hyper	2-hyper	3-hyper
Event	write	communicate	interact
Subject	corporate executive	executive	administrator
Object	idea	content	cognition

5.2 Learning Coarse-Grained Event Durations

The distribution of the means of the annotated durations in Figure 3 is bimodal, dividing the events into those that take less than a day and those that take a day or more. Thus, in our first machine learning experiment, we have tried to learn this *coarse-grained* event duration information as a binary classification task.

5.2.1 Inter-Annotator Agreement, Baseline, and Upper Bound. Before evaluating the performance of different learning algorithms, we first assess the inter-annotator agreement, the baseline, and the upper bound for the learning task.

Table 5 shows the inter-annotator agreement results among three annotators for binary event durations. The experiments were conducted on the same data sets as in Section 3.3. Two kappa values are reported with different ways of measuring expected agreement ($P(E)$), that is, whether or not the annotators have prior knowledge of the global distribution of the task, as described in Section 3.2.

Human agreement is 0.877 and is a good estimate of the *upper bound* performance for this binary classification task. The *baseline* for the learning task is always taking the most probable class. Because 59.0% of the total data is "long" events, the baseline performance is 59.0%**.**

5.2.2 Data. The original annotated data was translated into a binary classification. For each event annotation, the most likely or mean duration was calculated by averaging the logs of its lower and upper bound durations. If its most likely or mean duration was less than a day (about 11.4 in the natural logarithmic scale), it was assigned to the "short" event class, otherwise it was 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 divided the total annotated non-WSJ data (2,132 event instances) into two data sets: a training data set with 1,705 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) was kept for further test purposes (see Section 5.2.5).

5.2.3 Experimental Results (non-WSJ)

Learning Algorithms*.* Three supervised learning algorithms were evaluated for our binary classification task, namely, Support Vector Machines (SVM) (Vapnik 1995), Naive Bayes (NB) (Duda and Hart 1973), and Decision Trees (C4.5) (Quinlan 1993). The Weka (Witten and Frank 2005) machine learning package was used for the implementation of these learning algorithms. Linear kernel is used for SVM in our experiments.

Table 5
Inter-annotator agreement for binary event durations.

P(A)	P(E)		Kappa
0.877	With global distribution	0.528	0.740
	Without global distribution	0.500	0.755

Table 6
Test performance of three algorithms.

Class	Algor.	Prec.	Recall	F-Score
Short	SVM	0.707	0.606	0.653
	NB	0.567	0.768	0.652
	C4.5	0.571	0.600	0.585
Long	SVM	0.793	0.857	0.823
	NB	0.834	0.665	0.740
	C4.5	0.765	0.743	0.754

Each event instance has a total of 18 feature values, as described in Section 5.1, for the event only condition, and 30 feature values for the local context condition, when $n = 2$. For SVM and C4.5, all features are converted into binary features (6,665 and 12,502 features).

Results. Ten-fold cross validation was used to train the learning models, which were then tested on the unseen held-out test set, and the performance (including the precision, recall, and F-score[4] for each class) of the three learning algorithms is shown in Table 6. The significant measure is overall precision, and this is shown for the three algorithms in Figure 6, together with human agreement (the upper bound of the learning task) and the baseline.

We can see that among the three learning algorithms, SVM achieves the best F-score for each class and also 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.

Feature Evaluation. The best performing learning algorithm, SVM, was then used to examine the utility of combinations of four different feature sets (i.e., event, local context, syntactic, and WordNet hypernym features). The detailed comparison is shown in Table 7.[5]

We can see 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 of interest that the most important information is that from the predicate and arguments describing the event, as our linguistic intuitions would lead us to expect.

5.2.4 Learning Performance by Event Class. It's useful to compare the learning performance between different event classes, and see how they contribute to the overall learning performance. We would intuitively expect that some classes of events (e.g., reporting events) are relatively easier to learn than others.

[4] F-score is computed as the harmonic mean of the precision and recall: F = $(2\times Prec\times Rec)/(Prec+Rec)$.

[5] When all features are used, the results for event and context become the same, which indicates that the syntactic and hypenym features dominate.

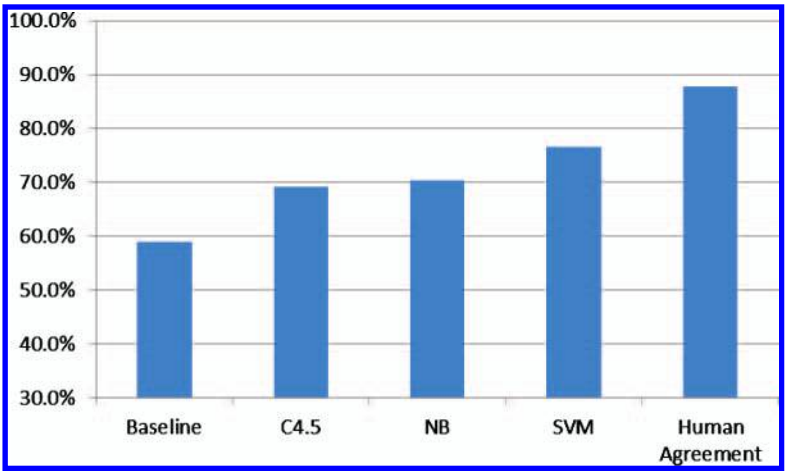


Figure 6
Overall test precision on non-WSJ data.

Table 7
Feature evaluation with different feature sets using SVM.

Class	**Event Only (*n* = 0)** **(*n* = 0)**			**Event Only** **+ Syntactic**			**Event + Syn** **+ Hyper**		
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
Short	0.742	0.465	0.571	0.758	0.587	0.662	0.707	0.606	0.653
Long	0.748	0.908	0.821	0.792	0.893	0.839	0.793	0.857	0.823
Overall Prec.		**74.7%**			**78.2%**			**76.6%**	
	Local Context **(*n* = 2)**			**Context** **+ Syntactic**			**Context + Syn** **+ Hyper**		
Short	0.672	0.568	0.615	0.710	0.600	0.650	0.707	0.606	0.653
Long	0.774	0.842	0.806	0.791	0.860	0.824	0.793	0.857	0.823
Overall Prec.		**74.2%**			**76.6%**			**76.6%**	

Table 8 shows the precision for each TimeML event class for the test set. For each event class, it also includes the total number of event instances, how many of them are as short or long events based on their annotations in the corpus, and the number of error instances.

Although there is too few data to draw firm conclusions, we can see that the event classes with the highest precision are aspectual events (e.g., continue, start) and perception events (e.g., see, look), though they don't contribute much to the overall performance (total instances are only 14 and 9, respectively). As we expected, reporting events perform relatively better than most other event classes, but they don't actually contribute much to the overall performance either (only 39 instances out of a total of 427 instances). The biggest event class is occurrence events (240 instances), and its precision is not much lower than the overall precision (73.3% vs. 76.5%). The most

Table 8
Learning performance for each event class.

Event Class	# Events	# Short	# Long	# Error	Precision
Aspectual	14	3	11	0	100%
Perception	9	0	9	0	100%
Reporting	39	38	1	6	84.6%
I_State	32	8	24	5	84.4%
State	61	2	59	13	78.7%
Occurrence	240	84	156	64	73.3%
I_Action	32	17	15	12	62.5%
Total	427	152	275	100	76.6%

Table 9
Test performance on WSJ data.

Class	Prec.	Rec.	F
Short	0.692	0.610	0.649
Long	0.779	0.835	0.806
Overall Prec.		*75.0%*	

difficult event class for learning their durations seems to be I-action events (e.g., try, insist).

5.2.5 Test on WSJ Data. Section 5.2.3 describes the experimental results with the learned model trained and tested on data from the same genre, that is, non-WSJ articles. 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). The performance (including the precision, recall, and F-score for each class) is shown in Table 9. The precision (75.0%) is very close to the test performance on the non-WSJ data, and indicates the significant generalization capacity of the learned model.

5.3 Learning the Most Likely Temporal Unit

These encouraging results prompted us to try to learn more fine-grained event duration information, namely, the most likely temporal units of event durations (cf. Rieger's [1974] ORDERHOURS, ORDERDAYS).

For each original event annotation, we can obtain the most likely (mean) duration by averaging its lower and upper bound durations, and assigning it to one of seven classes—second, minute, hour, day, week, month, and year—based on the temporal unit of its most likely duration.

However, human agreement on this more fine-grained task is low (44.4%). This is understandable. An annotation of [30 minutes, 1 hour] and [35 minutes, 2 hours] will not match, even though the area of their overlap is 52.72%.[6]

[6] A natural logarithmic scale is used for the overlap/agreement computation, as described in Section 3.1.

Table 10
Inter-annotator agreement for most likely temporal unit.

P(A)	**P(E)**		**Kappa**
0.798	With global distribution	0.151	0.762
	Without global distribution	0.143	0.764

Based on this observation, instead of evaluating the *exact* agreement between annotators, an "*approximate* agreement" is computed for the most likely temporal unit of events. In "approximate agreement," temporal units are considered to match if they are the same temporal unit or an adjacent one. For example, "second" and "minute" match, but "minute" and "day" do not.

We conducted an experiment for learning this multi-classification task. The same data sets as in the binary classification task were used. The only difference was that the class for each instance was now labeled with one of the seven temporal unit classes.

The baseline for this multi-classification task was always taking the temporal unit which with its two neighbors spans the greatest amount of data. Because the "week," "month," and "year" classes together take up the largest portion (51.5%) of the data, the baseline was always taking the "month" class, where both "week" and "year" were also considered a match. Table 10 shows the inter-annotator agreement results for the most likely temporal unit when using "approximate agreement". Human agreement, the upper bound, for this task increases from 44.4% to 79.8%.

Ten-fold cross validation was also used to train the learning models, which were then tested on the unseen held-out test set. The performance of the three algorithms is shown in Figure 7. The best performing learning algorithm is again SVM with 67.9% test precision. Compared with the baseline (51.5%) and human agreement (79.8%), this again is a very promising result, especially for a multi-classification task with such limited training data. It is reasonable to expect that when more annotated data become available, the learning algorithm will achieve higher performance when learning this and more fine-grained event duration information.

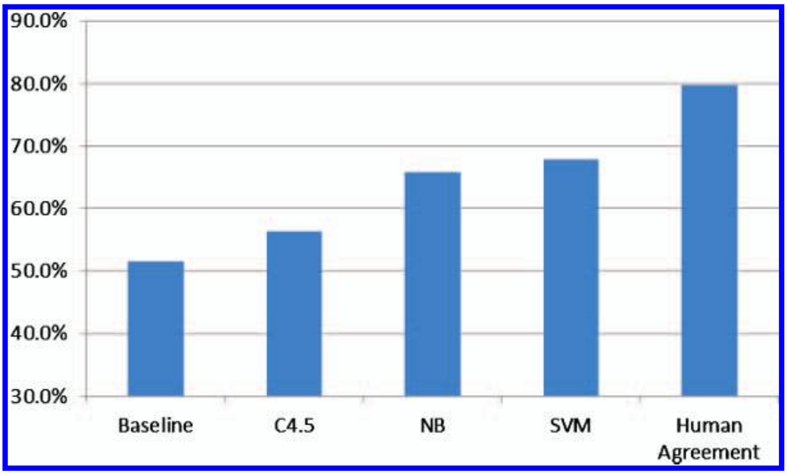


Figure 7
Overall test precisions for learning most likely temporal unit.

Although the coarse-grained duration information may look too coarse to be useful, computers have no idea at all whether a meeting event takes seconds or centuries, so even coarse-grained estimates would give it a useful rough sense of how long each event may take. More fine-grained duration information is definitely more desirable for temporal reasoning tasks. But coarse-grained durations to a level of temporal units can already be very useful. For example, when you want to know how long it takes to learn to use some new software, an answer of "hours" or "months" is often enough.

6. Conclusion

In the research described in this article, we have addressed a problem—extracting information about event durations encoded in event descriptions—that has heretofore received very little attention in the field. It is information that can have a substantial impact on applications where the temporal placement of events is important. Moreover, it is representative of a set of problems—making use of the vague information in text—that has largely eluded empirical approaches in the past. We have explicated the linguistic categories of the phenomena that give rise to grossly discrepant judgments among annotators, and give guidelines for resolving these discrepancies. We have also described a method for measuring inter-annotator agreement when the judgments are intervals on a scale; this should extend from time to other scalar judgments. Inter-annotator agreement is too low on fine-grained judgments. However, for the coarse-grained judgments of more than or less than a day, and of approximate agreement on temporal unit, human agreement is acceptably high. For these cases, we have shown that machine-learning techniques achieve encouraging results.

This article has also provided the necessary foundation for modeling and learning our most fine-grained duration annotations that are intervals on a scale. The duration interval has been modeled using the normal distribution, and the difference between two duration intervals are then the overlap between two distributions. So in order to learn this distribution, only two parameters need to be learned, namely, the mean and the standard deviation of the normal distribution, and the cost function can be defined as a function of distribution overlaps.

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, Inderjeet Mani, Daniel Marcu, James Pustejovsky, Deepak Ravichandran, and Nathan Sobo.

References

Allen, James F. 1984. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154.

Allen, James F. and George Ferguson. 1994. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 4(5):531–579.

Bejan, Cosmin Adrian and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden.

Boguraev, Branimir and Rie Kubota Ando. 2005. Timeml-compliant text analysis for temporal reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 997–1003, Edinburgh, Scotland, UK.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Applied Natural Language Processing*, pages 152–155, Trento, Italy.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Chittaro, Luca and Angelo Montanari. 2000. Temporal representation and reasoning in artificial intelligence: Issues and approaches. *Annals of Mathematics and Artificial Intelligence*, 28:47–106.

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

Duda, R. O. and P. E. Hart. 1973. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Ferro, Lisa. 2001. Instruction manual for the annotation of temporal expressions. MITRE Technical Report MTR 01W0000046, June 2001. The MITRE Corporation, Mc Lean, VA.

Filatova, Elena and Eduard Hovy. 2001. Assigning time-stamps to event-clauses. In *Proceedings of the Workshop on Temporal and Spatial Information Processing - Volume 13*, pages 13:1–13:8, Toulouse, France.

Fortemps, Philippe. 1997. Jobshop scheduling with imprecise durations: A fuzzy approach. *IEEE Transactions on Fuzzy Systems*, 5(1997), pages 557–569.

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Giorgi, Alessandra and Fabio Pianesi. 1997. *Tense and Aspect: From Semantics to Morphosyntax*, *Oxford University Press*, Oxford Studies in Comparative Syntax, Oxford.

Godo, Lluis and Lluís Vila. 1995. Possibilistic temporal reasoning based on fuzzy temporal constraints. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1916–1923, Montreal, Quebec, Canada.

Han, Benjamin and Alon Lavie. 2004. A framework for resolution of time in natural language. *Transactions on Asian Language Information Processing Special Issue on Spatial and Temporal Information Processing*, 3(1):11–32, March.

Harabagiu, Sanda and Cosmin Adrian Bejan. 2005. Question answering based on temporal inference. In *Proceedings of the AAAI-2005 Workshop on Inference for Textual Question Answering*, pages 27–34, Pittsburg, PA.

Hermjakob, Ulf and Raymond J. Mooney. 1997. Learning parse and translation decisions from examples with rich context. In *The 35th Annual Meeting of the Association for Computational Linguistics*, pages 482–489, Madrid, Spain.

Hitzeman, Janet, Marc Moens, and Claire Grover. 1995. Algorithms for analysing the temporal structure of discourse. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 253–260, Dublin, Ireland.

Hobbs, Jerry R. 2000. Half orders of magnitude. In *Proceedings of KR-2000 Workshop on Semantic Approximation, Granularity, and Vagueness*, pages 28–38, Breckenridge, CO.

Hobbs, Jerry R. and Vladik Kreinovich. 2001. Optimal choice of granularity in commonsense estimation: Why half orders of magnitude? In *Proceedings of Joint 9th Information Fuzzy Systems Association World Congress and 20th North American Fuzzy Information Processing Society International Conference*, July 2001, pages 1343–1348, Vancouver.

Hobbs, Jerry R. and Feng Pan. 2004. An ontology of time for the semantic web. *ACM Transactions on Asian Language Information Processing*, 3:66–85.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Lapata, Mirella and Alex Lascarides. 2006. Learning sentence-internal temporal relations. *Journal of AI Research*, 27(1), pages 85–117.

Madden, Carol J. and Rolf A. Zwaan. 2003. How does verb aspect constrain event representations? *Memory & Cognition*, 31:663–672.

Mani, Inderjeet and Barry Schiffman. 2007. Temporally anchoring and ordering events in news. *Event Recognition in Natural Language*, John Benjamins.

Mani, Inderjeet, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Liguistics*, pages 17–18, Sydney, Australia.

Mani, Inderjeet and George Wilson. 2000. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76, Stroudsburg, PA.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Moens, Marc and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14:15–28.

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

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

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Rieger, Charles J. 1974. Conceptual memory: A theory and computer program for processing and meaning content of natural language utterances. *Stanford Artificial Intelligence Laboratory Memo*, Computer Science Department, Stanford University.

Siegel, S. and N. J. Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw–Hill, Inc., 2nd edition, London, UK.

Smith, Carlota. 2005. Aspectual entities and tense in discourse. *Aspectual Inquiries*, pages 223–238, Kluwer, Dordrecht.

Tao, Cui, Harold R. Solbrig, Deepak K. Sharma, Wei-Qi Wei, Guergana K. Savova, and Christopher G. Chute. 2010. Time-oriented question answering from clinical narratives sing semantic-web techniques. In *Proceedings of the 9th International Semantic Web Conference on the Semantic Web - Volume Part II*, pages 241–256, Berlin.

Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY.

Vendler, Zeno. 1967. Linguistics in philosophy. *Aspectual Inquiries*. Cornell University Press, Ithaca, NY.

Witten, I. H. and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2nd edition, San Francisco, CA.

Zhou, Li and George Hripcsak. 2007. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202.

Zhou, Qing and Qing Zhou Richard Fikes. 2002. A reusable time ontology. In *Proceedings of the AAAI Workshop on Ontologies for the Semantic Web, the Eighteenth National Conference on Artificial Intelligence WS-02-11*, Edmonton, Alberta, Canada.