An Annotated Corpus of 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
{pan, rutu, hobbs}@isi.edu

Abstract

In this paper, we present our work on generating an annotated corpus for extracting information about the typical durations of events from texts. We include the annotation guidelines, the event classes we categorized, the way we use normal distributions to model vague and implicit temporal information, and how we evaluate inter-annotator agreement. The experimental results show that our guidelines are effective in improving the inter-annotator agreement.

1. Introduction

Suppose we read the sentence, “George W. Bush *met* with Vladimir Putin in Moscow.” We don’t know exactly how long that meeting lasted, but we do get *some* temporal information from the sentence. 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. 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?

As 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. For example, we know government policies typically last somewhere between one and ten years, while weather conditions fairly reliably persist between three hours and one day.

This research is potentially 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.

Very few events in text have explicit event duration information such as in “a 3-day *visit*” and “he has *worked* here for 5 years”. An immense amount of such event duration information in text is implicitly encoded in event descriptions that do not look at all temporal. Although there has been much work on temporal anchoring and event ordering in text (Mani and Wilson, 2000; Filatova and Hovy, 2001; Boguraev and Ando, 2005), to our knowledge, there has been no serious published empirical effort to model vague and implicit duration information in natural language, such as the typical durations of events, and then to perform reasoning over this information. (Cyc apparently 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 reasoning with imprecise durations (Godo and Vila, 1995; Fortemps, 1997), but these make no attempt to collect human judgments on such durations.)

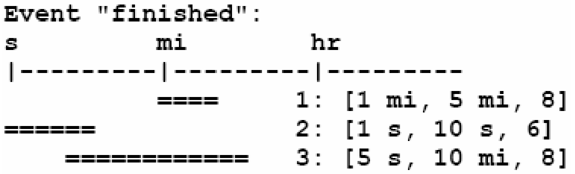
Our goal is to be able to extract such information from texts, and to that end we are annotating the events in news articles with bounds on their durations. For reliability, narrow bounds of duration are needed if we want to infer whether event *e* is happening at time *t*, while wide bounds of duration are needed to infer whether event *e* is *not* happening at time *t*. With a large enough annotated corpus, it will be possible to apply machine learning techniques and investigate which lexical and other features are predictive of this kind of temporal information. A close examination of the annotated corpus may also yield some general principles from which durations can be predicated.

In this paper, we present our work on producing such an annotated corpus, including annotation guidelines, the event classes we categorized, our inter-annotator agreement study, and experimental results. The news articles that we annotated are from the TimeBank corpus (Pustejovky et al., 2003). In Section 2 we first describe our annotation guidelines, including the annotation strategy and assumptions, and the representative event classes with examples. The inter-annotator agreement study and experimental results will be discussed and shown in Section 3. In Section 4 we describe future work.

2. Annotation Guidelines and Events Classes

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 (underlined) in the sentence

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



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

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. The confidence level of the annotators is generally subjective but as all three are 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 are asked to make their judgments as intended readers of the article, using whatever world knowledge is relevant to understand the article. They are 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.
2. Pick the set of probable scenarios, and annotate the bounds of their upper and lower bounds.

We deem the second to be the preferred strategy.

The judgments are to be made in context. First of all, information in the syntactic environment needs 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 need to be annotated in light of the information provided by the entire article. This means annotators should 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 should be used for annotation.

However, they should not 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 are asked to try to put themselves in the shoes of the 1990 readers of that article, and make their judgments accordingly. This is because we want people's estimates of typical durations of events, rather than the exact durations.

Annotation can be made easier and more consistent if *coreferential* and *near-coreferential* descriptions of events are identified initially. Annotators are asked to give the same duration ranges for such cases. For example, in the sentence "*during the demonstration, people chanted antigovernment slogans*", annotators should 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; however, we found that the vast majority of radically different judgments could 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).

These guidelines were then used to annotate 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 Annotation Guidelines: 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:

(1) 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.

(2) 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 <u>said</u> 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 <u>said</u>.

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 <u>destroyed</u> 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 <u>attacked</u>.

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 <u>dead</u>.

Here the time continues for an infinite amount of time, and we allow this as an 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; 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:

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

$P(A)$ is the observed agreement among the annotators, and $P(E)$ is the expected agreement, 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 way 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 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 (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 is range duration 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. 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 distribution (i.e., Gaussian distribution) to model our duration distributions.

In order to determine a normal distribution, we need to know the 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:

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

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, 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 1. The overlap or agreement is 0.508706.

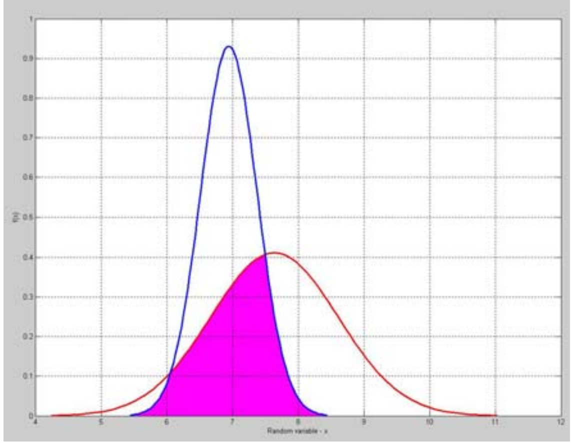


Figure 1. Overlap of Judgments of [10 minutes, 30 minutes] and [10 minutes, 2 hours].

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 1000 years or even positive infinity.

However, not all the durations are equally possible. As in (Krippendorff, 1980; 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 2, 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.

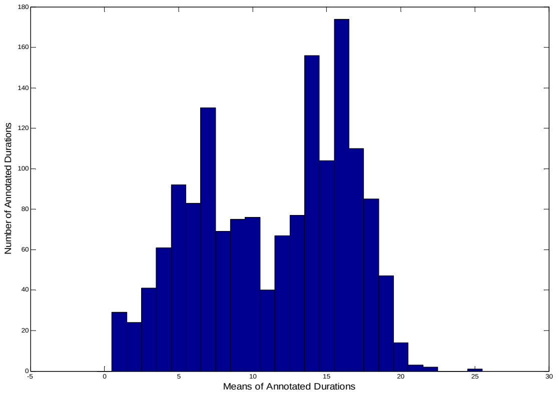


Figure 2. Distribution of Means of Annotated Durations.

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.

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 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.

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 (average) duration is 3.5:

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

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

This is the half orders of magnitude factor that Hobbs and Kreinovich (2001) argue gives the optimal granularity; making something 3 – 4 times bigger changes the way we interact with it.

Since the global distribution is determined by the above mean and width distributions, we can then compute the expected agreement, i.e., the probability that the annotators agree by chance, where the chance is actually 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

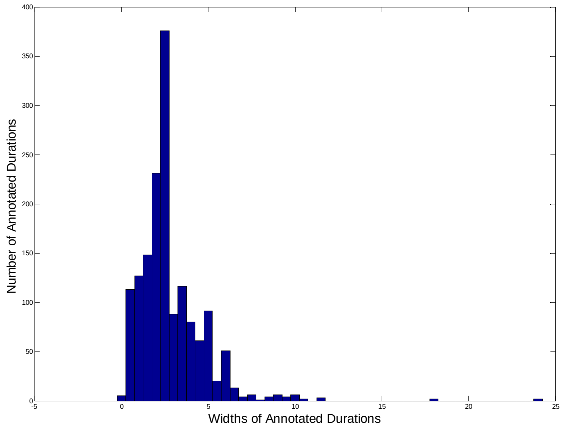


Figure 3. Distribution of Widths of Annotated Durations.

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 1000 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 1000 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 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 set is split into two sets. The first set contains 13 articles (521 events, 1563 annotated durations) which are all political and disaster news stories from ABC, APW, CNN, PRI, and VOA. The annotators annotated independently *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) contains 5 articles (125 events, 375 annotated durations) which are 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 4. 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, i.e., 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 (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.

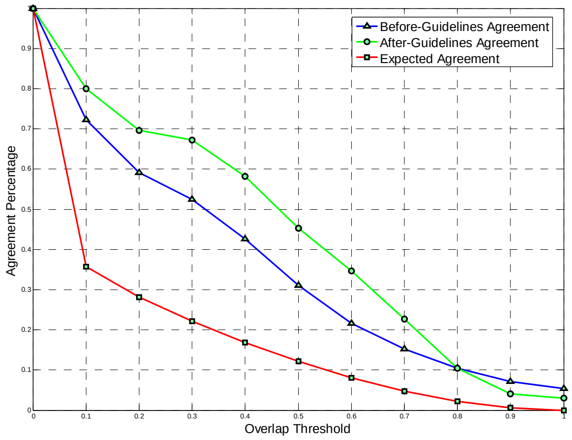


Figure 4. Inter-Annotator Agreement: Expected, Before-Guidelines, and After-Guidelines.

Table 1 shows more detailed experimental results. For each overlap threshold, it shows the expected (baseline) agreement, the before-guidelines agreement, and the after-guidelines agreement, and also the kappa statistic computed from the after-guidelines agreement ($P(A)$) and the expected (baseline) agreement ($P(E)$).

Moreover, Table 1 also shows a factor value that represents how far apart the means of two annotations can be in order to overlap with the given overlap threshold, assuming the width of the two annotations is the mean width, 2.6 on the natural logarithmic scale as computed from the width distribution shown in Figure 3.

For example, when the overlap threshold is 0.1, the factor value is 28.5, which means if one annotator guesses the mean duration for a given event is 1 minute, the other annotator will have 0.8 probability of guessing a mean duration from about 2 seconds (1 minutes / 28.5) to 28.5 minutes and their duration distributions will have at least 10% overlap. This also shows that if someone guesses 1 minute for a given event, it's not very likely (with a 0.2 probability) that the event will last more than 28.5 minutes or less than 2 seconds on average. Obviously, as Table 1 shows, if we want tighter bounds on the duration, our reliability will go down.

This factor is very useful for temporal reasoning tasks where we need to know whether given events have already ended or not.

4. Conclusions and Future Work

In this paper, we have presented our work on generating an annotated corpus for extracting the information about the typical durations of events from texts, including the annotation guidelines, the event classes we categorized, the way we use the normal distribution to model such

Overlap Threshold	Expected Agreement	BeforeGuidelines Agreement	AfterGuidelines Agreement	Kappa (AfterG. A.)	Factor
0.1	0.36	0.72	0.80	0.69	28.50
0.2	0.28	0.59	0.70	0.58	13.46
0.3	0.22	0.52	0.67	0.58	8.17
0.4	0.17	0.43	0.58	0.49	5.47
0.5	0.12	0.31	0.45	0.38	3.86
0.6	0.08	0.22	0.35	0.29	2.86
0.7	0.05	0.15	0.23	0.19	2.23
0.8	0.02	0.10	0.10	0.08	1.65
0.9	0.01	0.07	0.04	0.03	1.28
1.0	0.00	0.05	0.03	0.03	1.00

Table 1. Inter-Annotator Agreement with Different Overlap Thresholds.

vague and implicit temporal information, and how we evaluate inter-annotator agreement. The experimental results also show that our guidelines are effective in improving the inter-annotator agreement.

We have finished annotating all the 48 non-financial (i.e. non-WSJ) articles (2220 events) in TimeBank. We plan to annotate the rest of the articles (i.e. WSJ articles), and incorporate our annotations into TimeBank. The kinds of events that appear in non-financial articles can be quite different from those in financial articles. In TimeBank, for example, the event “kill” appears many times in different non-financial articles (e.g., disaster and crime articles), while it doesn’t appear at all in TimeBank’s WSJ articles. On the other hand, the event “sale” occurs much more often in financial articles. Thus, it is reasonable to learn typical durations of events from the current annotated non-financial articles first.

However, the size of the current annotated data is too small to get good results using machine learning techniques, if we are extracting the *fine-grained* event durations that we currently annotate. Thus we have decided to learn *coarse-grained* duration information from the current corpus first. The distribution of means in Figure 2 is bimodal, dividing the events into those that take less than a day and those that take more than a day. So we will experiment with learning this binary classification task. In subsequent experiments, as the size of our annotated corpus grows, we will move gradually to learning more fine-grained event durations, such as the most likely temporal units for events (e.g., Rieger’s ORDERHOURS, ORDERDAYS). 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 would definitely help more for temporal reasoning tasks, but we believe coarse-grained durations to a level of temporal units can already be very useful.

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 19th International Joint Conference on 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.

B. D. Eugenio and M. Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1).

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

P. Fortemps. 1997. Jobshop Scheduling with Imprecise Durations: A Fuzzy Approach. *IEEE Transactions on Fuzzy Systems* Vol. 5 No. 4.

L. Godo, L. Vila. 1995. Possibilistic Temporal Reasoning based on Fuzzy Temporal Constraints. *Proceedings International Joint Conference on Artificial Intelligence (IJCAI)*.

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. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

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

J. Pustejovsky, P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro and M. Lazo. 2003. The timebank corpus. In *Corpus Linguistics*, Lancaster, U.K.

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

S. Siegel and N. J. Castellan. 1988. Jr. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

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