Modeling and Learning Vague Event Durations for Temporal Reasoning

Feng Pan, Rutu Mulkar-Mehta, and Jerry R. Hobbs

Information Sciences Institute (ISI), University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292, USA
{pan, rutu, hobbs}@isi.edu

Abstract

This paper reports on our recent work on modeling and automatically extracting vague, implicit event durations from text (Pan et al., 2006a, 2006b). It is a kind of commonsense knowledge that can have a substantial impact on temporal reasoning problems. We have also proposed a method of using normal distributions to model judgments that are intervals on a scale and measure their inter-annotator agreement; this should extend from time to other kinds of vague but substantive information in text and commonsense reasoning.

Introduction

Consider the sentence from a news article:

George W. Bush met with Vladimir Putin in Moscow.

How long was the meeting? Our first reaction to this question might be that we have no idea. But in fact we do have an idea. We know the meeting lasted more than ten seconds and less than one year. As we guess narrower and narrower bounds, our chances of being correct go down. How accurately can we make vague 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?

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). For example, 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 (Chittaro and Montanari, 2000).

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. There is much temporal information that has hitherto been largely unexploited, implicitly encoded in the descriptions of events and relying on our knowledge of the range of usual durations of types of events. This paper reviews our recent work of exploring how this information can be captured automatically from text (Pan et al., 2006a, 2006b), including the way we used normal distributions to model the data for measuring inter-annotator agreement and applying machine learning techniques to extract coarse-grained event durations.

This research can be very useful for temporal reasoning applications in which the time course of events is to be extracted from text. For example, whether two events overlap or are in sequence often depends very much on their durations. 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.

Related Work

Although there has been much work on temporal anchoring and event ordering in text (Boguraev and Ando, 2005; Mani et al., 2006), to our knowledge, there has been no serious published empirical effort to model and learn vague and implicit duration information in natural language and to perform reasoning over this information.

There has been work on generalizing Allen's interval-based temporal reasoning (Allen, 1984) with incomplete temporal knowledge (Freksa, 1992), and also on using fuzzy logic for representing and reasoning with imprecise durations (Godo and Vila, 1995; Fortemps, 1997), but these make no attempt to collect human judgments on such durations or learn to extract them automatically from text.

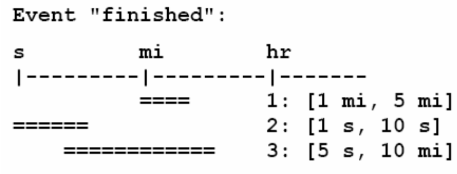
The event calculus (Kowalski and Sergot, 1986; Shanahan, 1999) has been applied to many commonsense reasoning problems (Mueller, 2006). The vague event duration information can be useful commonsense knowledge for reasoning about action and change with non-deterministic (durational) effects.

Annotation Guidelines and Event Classes

Our goal is to be able to extract the vague event duration information from text automatically, and to that end we first annotated the events in news articles with bounds on their durations. For reliability, narrow bounds of duration are needed if we want to infer that event *e* is happening at time *t*, while wide bounds of duration are needed to infer that event *e* is not happening at time *t*.

In the corpus, every event to be annotated was already identified in TimeBank (Pustejovsky et al., 2003). Annotators were instructed to provide lower and upper bounds on the duration of the event, encompassing 80% of the possibilities, excluding anomalous cases, and taking the entire context of the article into account. For example, here is the graphical 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]
======              2: [1 s, 10 s]
    ============    3: [5 s, 10 mi]
```

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

A preliminary exercise in annotation revealed about a dozen classes of systematic discrepancies among annotators’ judgments. We thus developed guidelines to categorize these event classes (e.g., aspectual events, reporting events, multiple events), and to make annotators aware of these cases and to guide them in making the judgments. The use of the annotation guidelines resulted in about 10% improvement in inter-annotator agreement. See (Pan et al., 2006b) for more complete description of the annotation guidelines and the event classes we categorized.

Inter-Annotator Agreement and Data Modeling

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. We used the kappa statistic (Carletta, 1996) for the measurement:

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

What Should Count as Agreement?

Determining what should count as agreement ($P(A)$) 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?

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

In fact, what annotators were instructed to annotate for a given event is not just a range, but a *duration distribution* for the event, where the area between the lower bound and the upper bound covers about 80% of the entire distribution area. Since it’s natural to assume the most likely duration for such distribution is its mean (average) duration, and the distribution flattens out toward the upper and lower bounds, we use the normal or Gaussian distribution to model our duration distributions. If the area between lower and upper bounds covers 80% of the entire distribution area, the bounds are each 1.28 standard deviations from the mean. With this data model, the agreement between two annotations can be defined as the overlapping area between two normal distributions.

Figure 1 shows the overlap in distributions for judgments of [10 minutes, 30 minutes] and [10 minutes, 2 hours], and the overlap or agreement is 0.508706.

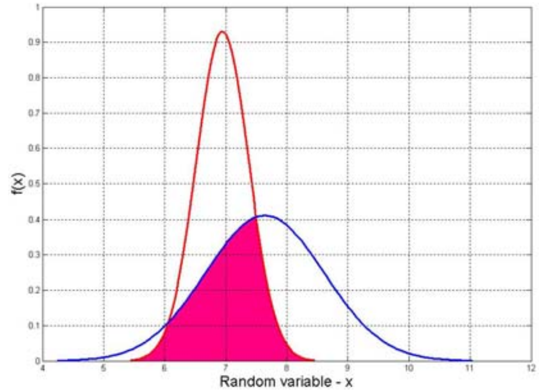


Figure 1: Overlap of Judgments of [10 minutes, 30 minutes] and [10 minutes, 2 hours].

Expected Agreement

What is the probability that the annotators agree by chance for our task (i.e., $P(E)$)? 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 (Carletta, 1996), we assume there exists one global distribution for our task (i.e., the duration ranges for all the events), and the “chance” annotations would be consistent with this distribution. Therefore, we must compute this 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.

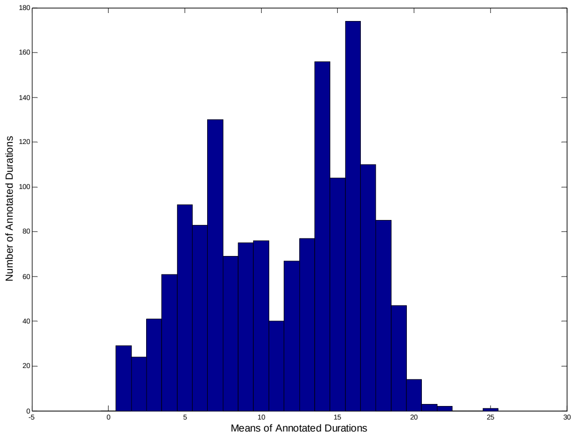


Figure 2: Distribution of Means of Annotated Durations.

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. The lowest point between the two peaks occurs at about 1 day.

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.

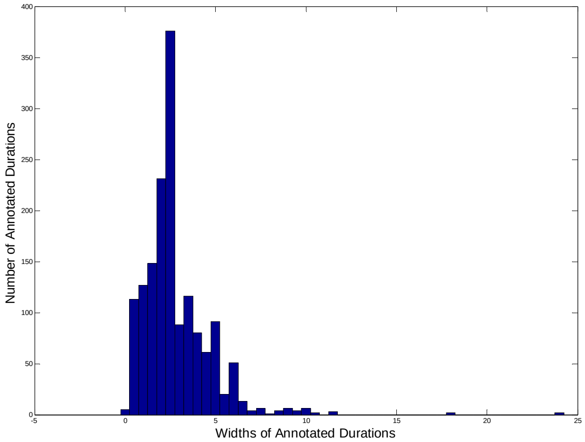


Figure 3: Distribution of Widths of Annotated Durations.

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 commonsense 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. See (Pan et al., 2006b) for the details of the computation.

Learning Coarse-Grained Event Durations

The corpus we have annotated from TimeBank currently contains 58 news articles (a total of 2288 event instances), including both print and broadcast news that are from a variety of news sources, such as ABC, NYT, VOA, and WSJ. The annotated data has already been integrated into the TimeBank corpus.

Because the annotated corpus is still fairly small, we cannot hope to learn to make *fine-grained* judgments of event durations that are currently annotated in the corpus, but we have demonstrated machine learning techniques applied to this data can yield useful *coarse-grained* event duration information, considerably outperforming a baseline and approaching human performance.

Features

We have considered the following lexical, syntactic, and semantic features in learning event durations.

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

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

WordNet Hypernyms. Events that share the same hypernyms may have similar durations. But closely related events don't always have the same direct hypernyms. Thus for our learning experiments, we extract the first 3 levels of hypernyms from WordNet (Miller, 1990). We extract the hypernyms not only for the event itself, but also for the subject and object of the event.

Experimental Results

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

The learning results in (Pan et al., 2006a) show that among all three learning algorithms explored (Naïve Bayes, Decision Trees C4.5, and Support Vector Machines (SVM)), SVM with linear kernel achieved the best overall precision (76.6%). Compared with a 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. Experiments also show very good generalization of the learned model to different news genres. Our feature evaluation study demonstrates 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 and hypernyms do not help and in fact may hurt. 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.

Some preliminary experimental results of learning more *fine-grained* event duration information, i.e., the most likely temporal unit, were also reported in (Pan et al., 2006a), where SVM again achieved the best performance with 67.9% test precision (baseline 51.5% and human agreement 79.8%).

Conclusion

In the research described in this paper, we have addressed a problem – modeling and automatically extracting vague event durations from text -- 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. In this work we also proposed a method of using normal distributions to model judgments that are intervals on a scale and measure their inter-annotator agreement; this should extend from time to other kinds of vague but substantive information in text and commonsense reasoning.

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

J. F. Allen. 1984. Towards a general theory of action and time. *Artificial Intelligence* 23, pp. 123-154.

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

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

J. Carletta. 1996. Assessing agreement on classifica-tion tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

L. Chittaro and A. Montanari. 2000. Temporal Representation and Reasoning in Artificial Intelligence: Issues and Approaches. *Annals of Mathematics and Artificial Intelligence*, vol. 28, no.1-4, pp. 47-106.

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

C. Freksa. 1992. Temporal Reasoning based on Semi-Intervals. *Artificial Intelligence*, Vol. 54:199-227.

L. Godo and L. Vila. 1995. Possibilistic Temporal Reasoning based on Fuzzy Temporal Constraints. In *Proceedings of 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*.

R. A. Kowalski and M. J. Sergot. 1986. A logic-based calculus of events. *New Generation Computing*, 4(1):67–95.

I. Mani, M. Verhagen, B. Wellner, C. M. Lee, J. Pustejovsky. 2006. Machine Learning of Temporal Relations. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL).*

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

E. T. Mueller. 2006. *Commonsense Reasoning*. Morgan Kaufmann, San Francisco.

F. Pan, R. Mulkar, and J. R. Hobbs. 2006a. Learning Event Durations from Event Descriptions. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 393-400, Sydney, Australia.

F. Pan, R. Mulkar, and J. R. Hobbs. 2006b. An Annotated Corpus of Typical Durations of Events. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 77-83, Genoa, Italy.

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.

M. Shanahan. 1999. The event calculus explained. In *Artificial Intelligence Today: Recent Trends and Developments*, Springer, Berlin.