Extracting and modeling durations for habits and events from Twitter

Jennifer Williams
Department of Linguistics
Georgetown University
Washington, D.C., USA
jaw97@georgetown.edu

Graham Katz
Department of Linguistics
Georgetown University
Washington, D.C., USA
egk7@georgetown.edu

Abstract

We seek to automatically estimate typical durations for events and habits described in Twitter tweets. A corpus of more than 14 million tweets containing temporal duration information was collected. These tweets were classified as to their habituality status using a bootstrapped, decision tree. For each verb lemma, associated duration information was collected for episodic and habitual uses of the verb. Summary statistics for 483 verb lemmas and their typical habit and episode durations has been compiled and made available. This automatically generated duration information is broadly comparable to hand-annotation.

1 Introduction

Implicit information about temporal durations is crucial to any natural language processing task involving temporal understanding and reasoning. This information comes in many forms, among them knowledge about typical durations for events and knowledge about typical times at which an event occurs. We know that lunch lasts for half an hour to an hour and takes place around noon, a game of chess lasts from a few minutes to a few hours and can occur any time, and so when we interpret a text such as "After they ate lunch, they played a game of chess and then went to the zoo" we can infer that the zoo visit probably took place in the early afternoon. In this paper we focus on duration. Hand-annotation of event durations is expensive slow (Pan et al., 2011), so it is desirable to automatically determine typical durations. This paper describes a method for automatically extracting information about typical durations for events from tweets posted to the Twitter microblogging site.

Twitter is a rich resource for information about everyday events – people post their tweets to Twitter publicly in real-time as they conduct their activities throughout the day, resulting in a significant amount of mundane information about common events. For example, (1) and (2) were used to provide information about how long a *work* event can last:

(1) *Had **work for an hour and 30 mins** now going to disneyland with my cousins :)*

(2) *I play in a loud rock band, I **worked at a night club for two years**. My ears have never hurt so much @melaniemarnie @giorossi88 @CharlieHi11*

In this paper, we sought to use this kind information to determine likely durations for events and habits of a variety of verbs. This involved two steps: extracting a wide range of tweets such as (1) and (2) and classifying these as to whether they referred to specific event (as in (1)) or a general habit (as in (2)), then summarizing the duration information associated with each kind of use of a given verb.

This paper answers two investigative questions:

- How well can we automatically extract fine-grain duration information for events and habits from Twitter?
- Can we effectively distinguish episode and habit duration distributions ?

The results presented here show that Twitter can be mined for fine-grain event duration information

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 223–227,
Jeju, Republic of Korea, 8-14 July 2012.

with high precision using regular expressions. Additionally, verb uses can be effectively categorized as to their habituality, and duration information plays an important role in this categorization.

2 Prior Work

Past research on typical durations has made use of standard corpora with texts from literature excerpts, news stories, and full-length weblogs (Pan et al, 2006; 2007; 2011; Kozareva & Hovy, 2011; Gusev et al., 2011). For example, Pan et al. (2011) hand-annotated of a portion of the TIMEBANK corpus that consisted of Wall Street Journal articles. For 58 non-financial articles, they annotated over 2,200 events with typical temporal duration, specifying the upper and lower bounds for the duration of each event. In addition they used their corpus to automatically determine event durations with machine learning, predicting features of the duration on the basis of the verb lemma, local textual context. and other information. Their best (SVM) classifier achieved precision of 78.2% on the course-grained task of determining whether an event's duration was longer or shorter than one day (compared with 87.7% human agreement). For determining the fine-grained task of determining the most likely temporal unit–second, minute, hour, day, week, etc.–achieved 67.9% (human agreement: 79.8%). This shows that lexical information can be effectively leveraged for duration prediction.

To compile temporal duration information for a wider range of verbs, Gusev et al. (2011) explored an automatic Web-based query method for harvesting typical durations of events. Their data consisted of search engine "hit-counts" and they analyzed the distribution of durations associated with 1000 frequent verbs in terms of whether the event lasts for more or less than a day (course-grain task) or whether it lasts for seconds, minutes, hours, days, weeks, months, or years (fine-grain task). They note that many verbs have a two-peaked distribution and they suggest that the two-peaked distribution could be a result of the usage referring to a habit or a single episode. (When used with a duration marker, *run*, for example, is used about 15% of the time with hour-scale and 38% with year-scale duration markers). Rather than making a distinction between habits and episodes in their data, they apply a heuristic to focus on episodes only.

Kozareva and Hovy (2011) also collected typical durations of events using Web query patterns. They proposed a six-way classification of ways in which events are related to time, but provided only programmatic analyses of a few verbs using Web-based query patterns. They have proposed a compilation of the 5,000 most common verbs along with their typical temporal durations. In each of these efforts, automatically collecting a large amount of reliable to cover a wide range of verbs has been noted as a difficulty. It is this task that we seek to take up.

3 Corpus Methodology

Our goal was to discover the duration distribution as well as typical habit and typical episode durations for each verb lemma that we found in our collection. A wide range of factors influence typical event durations. Among these are the character of a verb's arguments, the presence of negation and other embedding features. For this preliminary work, we ignored the effects of arguments, and focused only on generating duration information for verb lemmas. Also, tweets that were negated, conditional tweets, and tweets in the future tense were put aside.

3.1 Data Collection

A corpus of tweets was collected from the Twitter web service API using an open-source module called Tweetstream (Halvorsen & Schierkolk, 2010). Tweets were collected that contained reference to a temporal duration. The data collection task began on February 1, 2011 and ended on September 28, 2011. Duplicate tweets were identified by their unique tweet ID provided by Twitter, and were removed from the data set. Also tweets that were marked by Twitter as 'retweets' (tweets that have been reposted to Twitter) were removed. The following query terms (denoting temporal duration measure) were used to filter the Twitter stream for tweets containing temporal duration:

> *second, seconds, minute, minutes, hour, hours, day, days, week, weeks, month, months, year, years, decade, decades, century, centuries, sec, secs, min, mins, hr, hrs, wk, wks, yr, yrs*

The number of tweets in the resulting corpus was 14,801,607 and the total number of words in the

corpus was 224,623,447. Tweets were normalized, tokenized, and then tagged for POS, using the NLTK Treebank Tagger (Bird & Loper, 2004).

3.2 Extraction Frames

To associate each temporal duration with its event, events and durations were identified and extracted using four types of regular expression extraction frames. The patterns applied a heuristic to associate each verb with a temporal expression, similar to the extraction frames used in Gusev et al. (2011). The four types of extraction frames were:

- *verb for duration*
- *verb* in *duration*
- spend *duration verb*ing
- takes *duration* to *verb*

where *verb* is the target verb and *duration* is a duration-measure term. In (3), for example, the verb *work* is associated with the temporal duration term *44 years*.

(3) *Retired watchmaker **worked for 44 years** without a telephone, to avoid unnecessary interruptions, http://t.co/ox3mB6g*

These four extraction frame types were also varied to include different tenses, different grammatical aspects, and optional verb arguments to reach a wide range of event mentions and ordering between the verb and the duration clause. For each matched tweet a feature vector was created with the following features: verb lemma, temporal bucket (seconds, minutes, hours, weeks, days, months or years), tense (past or present), grammatical aspect (simple, progressive, or perfect), duration in seconds, and the extraction frame type (for, in, spend, or take). For example, the features extracted from (3) were:

[work, years, past, simple, 1387584000, FOR]

Tweets with verbal lemmas that occur fewer than 100 times in the extracted corpus were filtered out. The resulting data set contained 390,562 feature vectors covering 483 verb lemmas.

3.3 Extraction Precision

Extraction frame performance was estimated using precision on a random sample of 400 hand-labeled tweets. Each instance in the sample was labeled as correct if the extracted feature vector was correct in its entirety. The overall precision for extraction frames was estimated as 90.25%, calculated using a two-tailed t-test for sample size of proportions with 95% confidence (p=0.05, n=400).

3.4 Duration Results

In order to summarize information about duration for each of the 483 verb lemmas, we calculated the frequency distribution of tweets by duration in seconds. This distribution can be represented in histogram form, as in Figure 1 for the verb lemma *search*, with with bins corresponding to temporal units of measure (seconds, minutes, etc.).

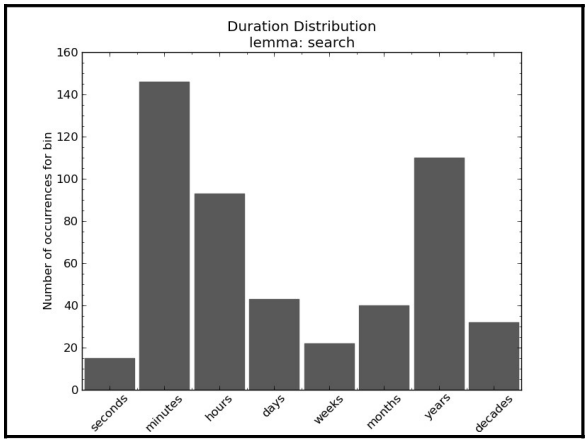

Figure 1: Frequency distribution for *search*

This histogram shows the characteristic bimodal-distributions noted by Pan et al., (2011) and Gusev et. al., (2011), an issue taken up in the next section.

4 Episodic/Habitual Classification

Most verbs have both episodic and habitual uses, which clearly correspond to different typical durations. In order to draw this distinction we built a system to automatically classify our tweets as to their habituality. The extracted feature vectors were used in a machine learning task to label each tweet in the collection as denoting a habit or an episode, broadly following Mathew & Katz (2009). This classification was done with bootstrapping, in a partially supervised manner.

4.1 Bootstrapping Classifier

First, a random sample of 1000 tweets from the extracted corpus was hand-labeled as being either

habit or episode (236 habits; 764 episodes). The extracted feature vectors for these tweets were used to train a C4.5 decision tree classifier (Hall et al., 2009). This classifier achieved an accuracy of 83.6% during training. We used this classifier and the hand-labeled set to seed the generic Yarowsky Algorithm (Abney, 2004), iteratively inducing a habit or episode label for all the tweets in the collection, using the WEKA output for confidence scoring and a confidence threshold of 0.96.

The extracted corpus was classified into 94,643 habitual tweets and 295,918 episodic tweets. To estimate the accuracy of the classifier, 400 randomly chosen tweets from the extracted corpus were hand-labeled, giving an estimated accuracy of 85% accuracy with 95% confidence, using the two-tailed t-test for sample size of proportions (p=0.05, n=400).

4.2 Results

Clearly the data in Figure 1 represents two combined distributions: one for episodes and one for habits, as we illustrate in Figure 2. We see that the verb *search* describes episodes that most often last minutes or hours, while it describes habits that go on for years.

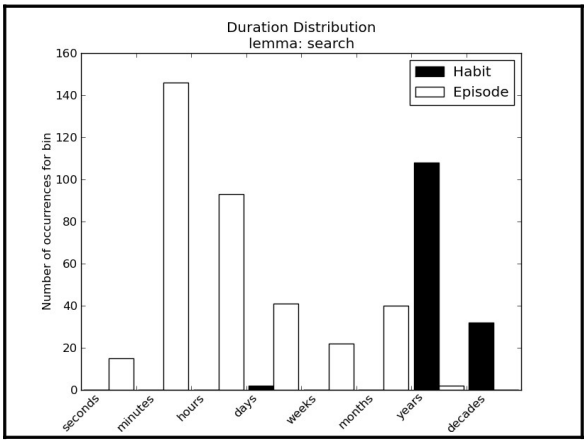

Figure 2: Duration distribution for *search*

These two different uses are illustrated in (4) and (5).

(4) *Obviously I'm the one who found the tiny lost black Lego in 30 seconds after the 3 of them searched for 5 minutes.*

(5) *@jaynecheeseman they've been searching for you for 11 years now. I'd look out if I were you.*

In Table 1 we provide summary information for several verb lemmas, indicating the average duration for each verb and the temporal unit corresponding to the largest bin for each verb.

Verb	Episodic Use		Habitual Use	
	Modal bin	**Mean**	**Modal bin**	**Mean**
snooze	minutes	1.6 hrs	decades	7.5 yrs
coach	hours	10 days	years	8.5 yrs
approve	minutes	1.7 mon.	years	1.4 yrs
eat	minutes	5.3 wks	days	5.7 yrs
kiss	seconds	4.5 days	weeks	1.8 yrs
visit	weeks	7.2 wks.	years	4.9 yrs

Table 1. Mean duration and mode for 6 of the verbs

It is clear that the methodology overestimates the duration of episodes somewhat – our estimates of typical durations are 2-3 times as long as those that come from the annotation in Pan, et. al. (2009). Nevertheless, the modal bin corresponds approximately to that the hand annotation in Pan, et. al., (2011) for nearly half (45%) of the verbs lemmas.

5 Conclusion

We have presented a hybrid approach for extracting typical durations of habits and episodes. We are able to extract high-quality information about temporal durations and to effectively classify tweets as to their habituality. It is clear that Twitter tweets contain a lot of unique data about different kinds of events and habits, and mining this data for temporal duration information has turned out to be a fruitful avenue for collecting the kind of world-knowledge that we need for robust temporal language processing. Our verb lexicon is available at: https://sites.google.com/site/relinguistics/.

References

Steven Abney. 2004. “Understanding the Yarowsky Algorithm”. *Computational Linguistics* 30(3): 365-395.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.

Andrey Gusev, Nathaniel Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. “Using query patterns to learn the durations of events”. *IEEE IWCS-2011, 9th International Conference on Web Services.* Oxford, UK 2011.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Rune Halvorsen, and Christopher Schierkolk. 2010. Tweetstream: Simple Twitter Streaming API (Version 0.3.5) [Software]. Available from https://bitbucket.org/runeh/tweetstream/src/.

Jerry Hobbs and James Pustejovsky. 2003. “Annotating and reasoning about time and events”. In *Proceedings of the AAAI Spring Symposium on Logical Formulation of Commonsense Reasoning*. Stanford University, CA 2003.

Zornitsa Kozareva and Eduard Hovy. 2011. “Learning Temporal Information for States and Events”. In *Proceedings of the Workshop on Semantic Annotation for Computational Linguistic Resources (ICSC 2011)*, Stanford.

Thomas Mathew and Graham Katz. 2009. “Supervised Categorization of Habitual and Episodic Sentences”. *Sixth Midwest Computational Linguistics Colloquium.* Bloomington, Indiana: Indiana University.

Marc Moens and Mark Steedman. 1988. “Temporal Ontology and Temporal Reference”. *Computational Linguistics* 14(2):15-28.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R. Hobbs. 2006. “An Annotated Corpus of Typical Durations of Events”. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 77-82. Genoa, Italy.

Feng Pan, Rutu Mulkar-Mehta, and Jerry R. Hobbs. 2011. "Annotating and Learning Event Durations in Text." *Computational Linguistics* 37(4):727-752.