A New Twitter Verb Lexicon for Natural Language Processing

Jennifer Williams, Graham Katz
Department of Linguistics
Georgetown University, Washington, DC., USA
E-mail: jaw97@georgetown.edu, egk7@georgetown.edu

Abstract

We describe in-progress work on the creation of a new lexical resource that contains a list of 486 verbs annotated with quantified temporal durations for the events that they describe. This resource is being compiled from more than 14 million tweets from the Twitter microblogging site. We are creating this lexicon of verbs and typical durations to address a gap in the available information that is represented in existing research. The data that is contained in this lexicon is unlike any existing resources, which have been traditionally comprised of literature excerpts, news stories, and full-length weblogs. This kind of knowledge about how long an event lasts is crucial for natural language processing and is especially useful when the temporal duration of an event is implied. We are using data from Twitter because Twitter is a rich resource since people are publicly posting real events and real durations of those events throughout the day.

Keywords: NLP, natural language understanding, temporal reasoning

1. Introduction

We describe in-progress work for the creation of a new lexical resource that contains a list of verbs that are annotated with quantified temporal durations for the events that they describe. This lexical resource is being compiled from more than 14 million tweets from the Twitter microblogging site. We are creating this lexicon of verbs and typical durations to address a gap in existing research (Pan et al., 2011; Kozareva & Hovy, 2011; Gusev et al., 2011).

The data that is contained in this lexicon is unlike any existing resources, which have been traditionally comprised of literature excerpts, news stories, and full-length weblogs. One of the advantages of using data from Twitter is that tweets are very short. Twitter limits the length of each tweet to be less than 140 characters. Users of the Twitter web service are updating their public status regarding their personal or business affairs at all times throughout the day with a simple and short message. In this way, real people are reporting real events and their durations.

In this lexical resource, each verb is annotated with typical duration concerning two kinds of usage descriptions: verbs describing particular events which constitute a single episiode, as in (1), and verbs that describe characterizing events as habits such as in (2):

(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

Our data contains events that are commonly mentioned on Twitter and we are especially interested in the tweets that specify how long an event lasts. For example, we used (1) to find out that a *work* event can last for an hour and a half, and we used (2) to find out that a *work* habit can last for years. Making this distinction allows us to report typical durations for events in our collection in terms of habit-describing durations and episode-describing durations.

The kind of knowledge about how long an event lasts is crucial for natural language processing and is especially useful when the temporal duration of an event is implied. Implicit information comes in many forms, among them knowledge about typical durations for events, as well as knowledge about typical times at which an event occurs – we know that lunch lasts for perhaps 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.

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 example, *eating a snack* is different from *eating a meal* since these events have different durations. Tweets that describe a negated event, tweets that describe an event as being conditional, and tweets in the future tense were put aside.

In this paper we show the duration distributions for different events that we have added to the Twitter Verb Lexicon. We describe some of the characterizations of these distributions. The typical duration for a particular event and a habit are not the same and this information is especially useful for natural language understanding and temporal reasoning.

2. Prior Work

Past research on typical durations have extracted information from literature excerpts, news stories, and full-length weblogs (Pan et al, 2011; Kozareva & Hovy, 2011; Gusev et al., 2011). In this work, we have been able to collect reliable data for over 400 verb lemmas. The data that is contained in this lexicon is unlike any existing resources in both breadth and variety.

Our research builds on existing works. Pan et al. (2006;2011) were the first to annotate events in a corpus with typical temporal durations. They annotated a portion of the TIMEBANK corpus that consisted of Wall Street Journal articles. For 48 non-financial articles, they annotated 2220 events with typical temporal duration. Pan et al. (2006) defined their annotation task in terms of granularity. The coarse-grain annotation task is to determine if an event lasts for more than a day or less than a day. The fine-grain annotation task is to determine if an event lasts for seconds, minutes, hours, days, weeks, months, or years. For example, a *war* may last months or years, but it will never last for seconds. Human annotation is a time-consuming way of acquiring typical duration this information. We have found a way to extract this information automatically at a very fine-grain scale.

In order to expand the temporal duration information to a wider range of verbs, Gusev et al. (2011) explored a Web-query-based method for harvesting typical durations of events. Their data consisted of search engine "hit-counts" instead of a corpus, and they compiled a database of typical durations for 1000 frequent verbs. 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 call for a compilation of the 5,000 most common verbs along with their typical temporal durations. In each of these efforts, the distinction between a single episode – say smoking a cigarette – and a habit – say being a cigarette smoker – is noted as a difficulty.

3. Lexicon Creation

We used data that we collected from Twitter to compile our lexicon. This involves both parsing the Twitter feeds to extract temporal information, and classifying the verb use as to whether it is describing an event or a habit (Williams & Katz, 2012). Our lexicon of verbs and the typical durations for the events they describe was built from our collected Twitter tweets that were filtered and normalized, tagged with part-of-speech tags based on the Penn Treebank tagset, and each tweet has unique identification number.

3.1 Data Collection

All of the data was collected from the Twitter web service API using a module called Tweetstream (Halvorsen & Schierkolk, 2010). The online data collection task began on February 1, 2011 and ended on September 28, 2011. The corpus contains 14,801,607 unique tweets, making a total of 224,623,434 words.

Our data collection method first filtered text for each tweet to ensure that each tweet in the collection contained a quantified temporal duration. Text filtering was done with a set of 28 initial query words that we used with the Tweetstream software module:

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 initial query words that we used are the enumerated variations of temporal duration units found in the work done by Pan and Hobbs (2006). The variations were enumerated here because unlike news stories, the language that is used in tweets can vary significantly among speaker styles. For example, one favorite variation is "mins" for "minutes" and we tried to account for that variation in our query words. For every tweet containing any of our query words, that tweet was then matched to a set of regular expressions to determine if the temporal interval was given a numerical measure.

Tweets are streamed as a data structure that contains useful meta information. We used the unique tweet ID that was assigned by Twitter to remove duplicate tweets from the data. It is not the case that all data from Twitter is in English. In order to determine that each tweet in the collection was in English, we excluded tweets that specified any language other than English in the Twitter user language identification field.

3.2 Data Processing

Tweets that contained temporal duration specification underwent text cleaning. The goal of text cleaning was to normalize the text. Each normalized tweet contained only word or digit tokens, so we removed URIs, "@" mentions, and "#" hashtags. We also standardized each of the duration units in our dataset. For example, we translated "mins" and "minutes" to "minute" to make the temporal duration units consistent. Tweets were tokenized on whitespace, and then tagged for POS using the NLTK treebank tagger (Bird & Loper, 2004).

We associated a temporal duration with each event in our corpus and extracted events and their durations using regular expression pattern matching. Our patterns were designed to match text exclusively based on part of speech, and part of speech tags have therefore played an important role. We created four main types of patterns that correspond to the four main types of extractors that we used. Our four extractors were the following: *for*, *spend*, *take*, and *in*. However, in the case of *take* and *spend*, we accounted for different tenses and aspects at the sentence level such as: *have taken*, *has taken*, *took*, *taking*, *takes*, etc. We used these four types of extraction frames

as a starting point as described in the work of Gusev et al. (2011).

Our patterns can be characterized in terms of these four types, and the regular expressions that are associated with each type will match: a verb, one of the extractors (*for*, *spend*, *take*, *in*), and a duration. The duration must include one of our standardized temporal duration units: *second*, *minute*, *hour*, *day*, *week*, *month*, *year*, *decade*. We show examples of our extraction frame types in (a) through (e):

(a) [had worked, working...] for [6, six...] [hours, days, weeks...]

(b) [finished, had finished...] [laundry, a book...] in [2, two...] [hours, weeks...]

(c) [was spending, had spent...] [25, twenty-five...] [seconds, minutes...] [talking on the phone, cooking in the kitchen...]

(d) [had taken, has took...] [me, someone...] [10, ten] [seconds, hours..] to [upload something, download something...]

(e) [uploading something, downloading something...] [has taken, takes, took...] [two seconds, two hours...]

The above examples of our four pattern types in (a) through (e) show some different ways that a verb and duration can be extracted. Example (a) demonstrates our pattern type that uses the extractor *for* to say that we can extract any tense or aspect, for any verb, a duration, and an optional auxiliary. We can match both digits or spelled numbers. Example (b) uses the extractor *in*, which we found the be particularly interesting. Use of *in* can sometimes denote a future event, depending on the use. To avoid extracting future events, we restricted patterns with the *in* extractor to match only past tense except where perfect aspect is used. So we did allow for the present perfect, as in: "Sally has finished her homework in 30 minutes". For the pattern type with extractor *spend*, shown in example (c), we did not vary the word order because the pattern overgeneralized which could be due to errors from part of speech tagging. Examples (c) and (d) show how we extract tense and aspect at the sentence level. The word ordering is varied with patterns that use extractors *for*, *in*, and *take* and we show an example of this in (d) and (e).

Since we are using part of speech tags in all of the regular expressions to do matching, we are at liberty to allow for some variation in what we match. Based on the Penn Treebank Tagset we can extract verbs that are tagged with any one of the following parts of speech: *VB, VBZ, VBN, VBG, VBD*. The pattern to match duration units is always expressed by a disjunction of *second, minute, hour, day, week, month, year, decade*.

Tense and aspect is determined by the part of speech of the verb and the nature of the auxiliary. Auxiliary verbs are optional. The tense and aspect is always extracted at the sentence level. By matching with an optional auxiliary, we can account for a lot of variation in the kinds of tenses and aspects that we collect, such as: *have worked*, *is working*, *are working*, *was working*, *worked*, etc.

We also allow for some variation in our patterns for matching quantified durations as we showed in (a). We match a duration that is in the form of a digit, or a spelled out number, and we can match any of the following: *25*, *twenty-five*, *twentyfive*, or *twenty five*.

In addition to the optional tenses and aspects, our patterns will match optional adverbials such that a duration could be preceded by an adverbial phrase as in: *nearly*, *almost*, etc. The four types of patterns were also varied to match for optional verb arguments as in: *planning a party*, *planning a party for Sally*, etc. Examples (b), (c), (d), and (e) demonstrate optional verb arguments.

All of the duration mentions were converted into seconds using regular expressions. So if a tweet contained the phrase "twentyfive minutes" then we consider the duration to be 1500 seconds. If a tweet contained the quantifiers *a* or *an*, then we treated these with a value of 1 so that "an hour" is converted into 3600 seconds. Not all of the durations that we extracted were quantified by digits and some of the tweets contained durations such as "some hours" or "a few years" in which case we treated the duration to be a single hour in the former case, or a single year in the latter. Some of the durations are expressed figuratively, in which case the duration that we extracted is not necessarily reliable. In the case where an event was mentioned as having a very long duration, such a billion seconds, we omitted it from our data. We dropped 6,389 tweets wherein the event was said to have lasted for more than one billion seconds.

4. Analysis

The extracted corpus contains 396,951 tweets that cover 486 verb lemmas. Extraction frame precision was measured on a randomly selected sample of 400 tweets and these were hand-labeled for correct or incorrect. Each instance in our random sample was labeled as correct only if we were able to correctly extract the verb, the tense, the aspect, and the duration. The overall precision for our extraction frames was 90.25%. We calculated this using a two-tailed t-test for sample size of proportions with 95% confidence (p=0.05, n=400).

Typical event durations can be examined in two ways. Since all of the durations have been converted to seconds, we can describe the duration distribution of an event using mean and standard deviation, or we can bin the data and examine the duration distribution by bins. We discuss both in this section.

We applied the method described in Williams and Katz (2012) to divide our collection into a group that consists of habit events and another group that consists of episode

events. There are 99,918 tweets identified as habits and 297,032 tweets identified as episodes. We are analysing the duration distributions based on the durations for events and habits that are described by the verbs that we extracted.

Duration distributions that are disaggregated by habit-describing use and episode-describing use are very informative for typical durations. After we had converted all of the durations into seconds, we binned the durations into 10 bins using a log10 scale. We found that a log10 scale is appropriate for the histograms since duration units nicely correspond to each of our bins. The x-axis is labeled such that a duration that is in the range of 100 to 1000 seconds is *minutes*, a duration in the range of 1001 to 10000 seconds is *hours*, etc.

Consider Figure 1 and Figure 2, below. In Figure 1, the distribution does not include any information about which durations are associated with habits and which durations are associated with single episodes. In addition, there is no previous work that makes this distinction when reporting typical durations of events.

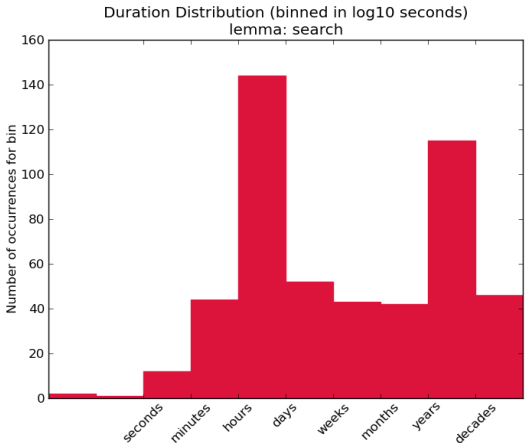

Figure 1: distribution for *search* without any habit/episode distinction

In Figure 1, we see that a *search* event can typically last for hours or years. This kind of double-peak distribution is common in our data as well as previous work. Since our data has been disaggregated into characterizing events (habits) and particular events (episodes), we are able to show that the bimodal distributions exist because there are two kinds of events represented in our data: events that are particular and events that are characterizing. We can see from Figure 2 that a particular *search* event will typically last for some hours and a characterizing *search* habit can go on for for years.

We also found some interesting groups in our collection. In Figure 3, we see that an *answer* event is most often reported as an episode and less often reported as a habit This is also the case for a *camp* event, shown in Figure 4. However an *achieve* event is most frequently reported as a habit, which we show in Figure 5.

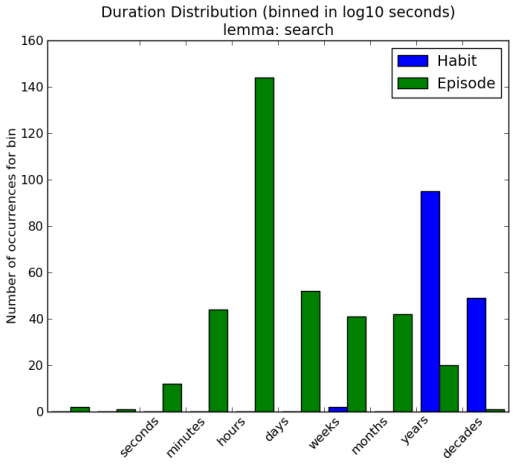

Figure 2: duration distribution for *search* (a single episode can typically lasts for hours but a habit goes on for years)

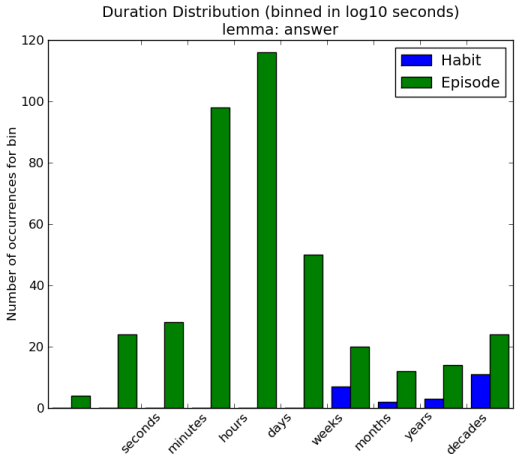

Figure 3: duration distribution for *answer*

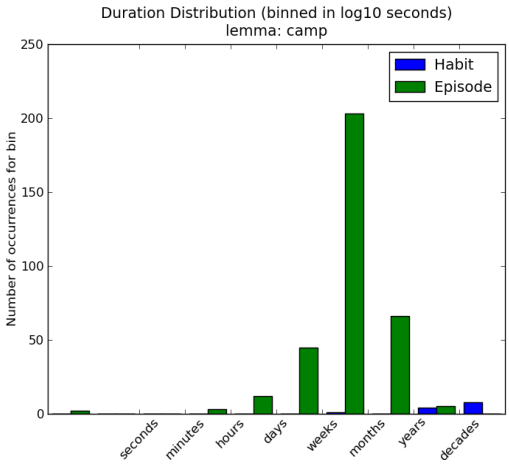

Figure 4: duration distribution for *camp*

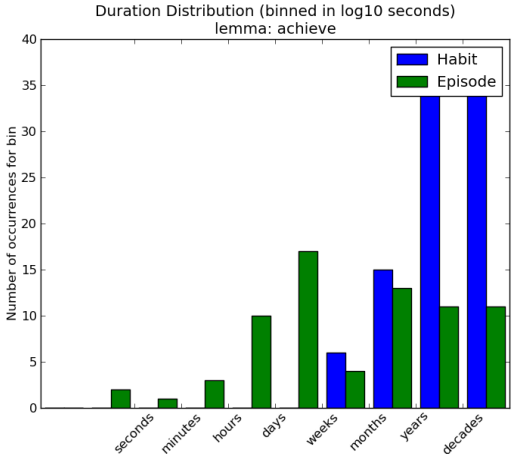

Figure 5: duration distribution for *achieve*

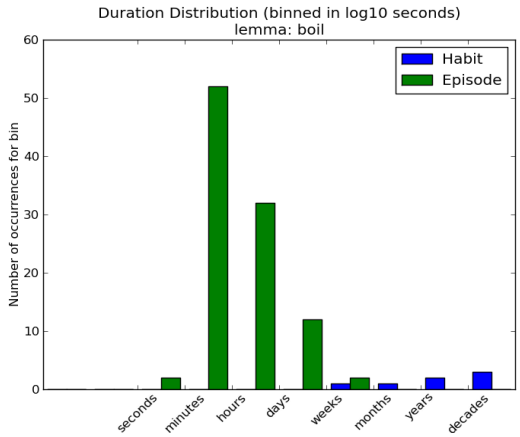

Figure 6: duration distribution for *boil*

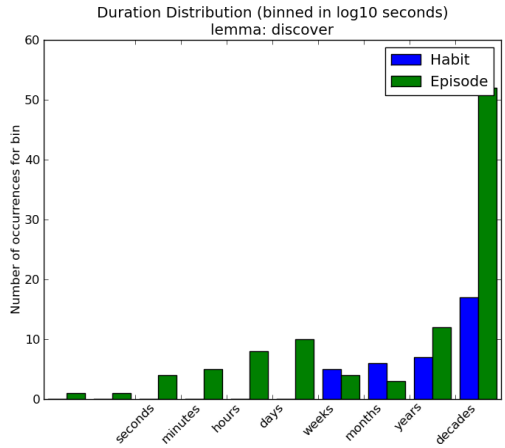

Figure 7: duration distribution for *discover*

When reasoning about events, we can see that a *camp* event can take weeks, but a single episode of camping does not take seconds and if it does go on for years or decades then it is a habit. Consider the event *boil*. As seen in Figure 6, the event lasts for some short minutes and if a *boil* event lasts much longer then it is probably describing a habit. An event such as *discover* shown in Figure 7, can take decades for a single episode. *Discover* events are interesting because they typically have long episode durations and long habit durations. We know that *click* lasts for seconds and can go on for minutes or hours during a particular event. A *click* event can have a very brief episode duration as well as a very long habit duration, shown in Figure 8.

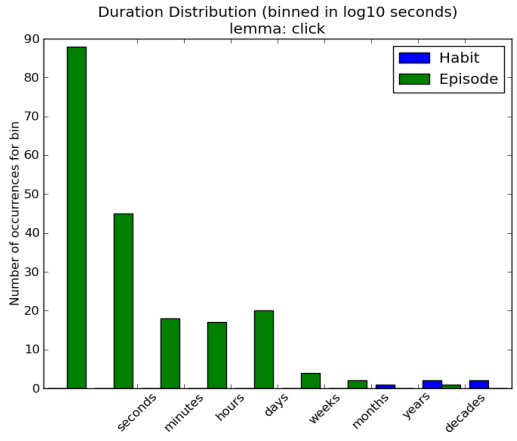

Figure 8: duration distribution for *click*

We also analyzed duration distributions by calculating the mean and standard deviation. We report the modes for episode and habit durations in Table 1 for several lemmas, as well as the collection overall.

Verb lemma	Episode Duration	Habit Duration
snooze	minutes	decades
lie	hours	years
say	seconds	years
approve	minutes	years
Overall Collection	minutes	years

Table 1. Typical episode and habit durations for events described by verbs

5. Discussion

Our lexicon consists of 486 verbs for which we have collected at least 30 typical durations. For many verbs there is a significant difference between the typical episode length and typical habit length – and our data shows that Twitter users tweet about both.

The Twitter Verb Lexicon of typical durations of events is a resource for many interesting NLP tasks as well as

theoretical research. Information such as what we have assembled for the Twitter Verb Lexicon will allow researchers to begin to address the problems in automatic temporal interpretation that only information about typical durations can solve, such as how to distinguish the interval specified by "shortly after" in "Shortly after I went running, I took a shower" from that in "Shortly after I lost my job, I moved to Colorado."

6. References

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

Gusev, A., Chambers, N., Khaitan, P., Khilnani, D., Bethard, D., & Jurafsky, D. 2011 "Using query patterns to learn the durations of events". *IEEE IWCS-2011, 9th International Conference on Web Services.* Oxford, UK 2011.

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

Hobbs, J., & Pustejovsky, J. 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.

Kozareva, Z., & Hovy, E. 2011. "Learning Temporal Information for States and Events". In *Proceedings of the Workshop on Semantic Annotation for Computational Linguistic Resources (ICSC 2011)*, Stanford.

Mathew, T. and Katz, G. 2009. "Supervised Categorization of Habitual and Episodic Sentences". *Sixth Midwest Computational Linguistics Colloquium*. Bloomington, Indiana: Indiana University, 2009.

Pan, F., Mulkar, R., and Hobbs, J.R. 2006. "An Annotated Corpus of Typical Durations of Events". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 77-82, Genoa, Italy.

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

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

Williams, J. and Katz, G. 2012. "Extracting and Modeling typical durations of events and habits from Twitter". In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju, Republic of Korea.