Extracting fine-grained durations for verbs from Twitter

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

Abstract

This paper presents recent work on a new method to automatically extract fine-grained duration information for common verbs using a large corpus of Twitter tweets. Regular expressions were used to extract verbs and durations from each tweet in a corpus of more than 14 million tweets with 90.38% precision covering 486 verb lemmas. Descriptive statistics for each verb lemma were found as well as the most typical fine-grained duration measure. Mean durations were compared with previous work by Gusev et al. (2011) and it was found that there is a small positive correlation.

1 Introduction

Implicit information about events 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 perhaps an hour and takes place around noon, 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 chess game probably lasted for a few hours and not for several months.

This paper describes a new method for extracting information about typical durations for verbs 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 information about common events. Data from Twitter is more diverse than the data found in news articles that has typically been used for looking at event durations (Pan et al., 2011). For example, consider that (1) was used find out that working can last for an hour and a half:

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

I extracted and analyzed a large number of such tweets containing temporal duration information. This involved identifying relevant tweets, extracting the temporal phrases, and associating these with the verb they modified. The processes are described below. Two objectives were investigated in this paper: (1) how to automatically extract duration information for common verbs from Twitter, and (2) to discover the duration distributions for common verbs. A wide range of factors influence typical 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. To simplify the task, I set aside tweets wherein the sentence-level verb was negated, or in the conditional or future tenses. Examining the effect of verb arguments was also set aside in this work.

Proceedings of the 2012 Student Research Workshop, pages 49–54,

Jeju, Republic of Korea, 8-14 July 2012.

The problem of finding typical duration for events can be viewed as a coarse-grained task or a fine-grained task. At the coarse-grained level it could be determined whether or not a chess game lasts for more or less than one day, whereas a fine-grained analysis would indicate that a chess game lasts for minutes or hours.

The results of this work show that Twitter can be mined for duration information with high accuracy using regular expressions. Likewise, the typical durations for verbs can be summarized in terms of the most frequent duration-measure (seconds, minutes, hours, days, weeks, months, years, decades) as well as by descriptive statistics.

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., 2011; Kozareva & Hovy, 2011; Gusev et al., 2011). However, data from Twitter has been useful for other NLP tasks such as detecting sarcasm (González-Ibáñez et al., 2011), as well as sentiment for Twitter events (Thelwall et al., 2011). The present work used data from Twitter because it is readily available and diverse in its linguistic nature.

2.1 Hand-Annotation

The first to examine typical durations of events was Pan et al. (2011). They describe a method to annotate events with duration information. They hand-annotated a portion of the TIMEBANK corpus that consisted of news articles and non-financial articles from the Wall Street Journal. They did this for 48 news articles (for 2,132 events) and 10 Wall Street Journal articles (for 156 events). For each event, three annotators indicated a lower-bound duration and an upper-bound duration that would cover 80% of the possible cases provided that durations are normally distributed. They converted the upper and lower bounds into distributions. They defined annotator agreement to be the average overlap of all the pairwise overlapping areas, calculated using the kappa statistic.

In their experiments, Pan et al. (2011) examined their annotation guidelines and found that annotator agreement was significantly improved after annotators were instructed to use their guidelines. These guidelines took into consideration information about event classes. The final guidelines addressed the following kinds of classes: actions vs. states, aspectual events, reporting events (quoted and unquoted reporting), multiple events, events involving negation, appearance events, and positive infinitive duration[1]. Human agreement for coarse-grained analysis was reported to be 87.7% whereas agreement for fine-grained analysis was 79.8%.

Hand-annotation is an expensive way of acquiring typical duration and human annotators do not always agree on how long events last. This paper presents a way to extract duration information automatically and at a fine-grained scale to discover the kinds of distributions of durations for different verbs as well as their typical durations.

2.2 Web Extraction

To compile temporal duration information for a wider range of verbs, Gusev et al. (2011) explored a Web-based query method for harvesting typical durations of events. They used five different kinds of query frames to extract events and their durations from the web at a coarse-grained level and at a fine-grained level. They compiled a lexicon of 10,000 events and their duration distributions.

In the work of Gusev et al. (2011), they calculated the most likely duration for events at a fine-grained scale. To obtain each of the fine-grained duration distributions, they first binned durations into their temporal unit measures (seconds, minutes, hours, etc.). Next, they discarded data that was extracted using patterns that had very low "hit-counts" in their effort to judge the reliability of their extraction frames. Finally, they normalized the distributions based on how often each pattern occurs in general. They note that many verbs have a two-peaked distribution. 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. In the case of the event *say*, Gusev et al. (2011) chose to normalize their duration distributions with a heuristic to account for the possibility that all of the year-scale durations could

[1] Positive infinitive durations describe states that will last forever once they begin, such as being dead.

be attributed to the common phrase "... for years".

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 asked for 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 data which covers a wide range of verbs has been noted as a difficulty.

3 Methodology

3.1 Data Collection

For the present study, tweets were collected from the Twitter web service API using an open-source Python module called Tweetstream (Halvorsen & Schierkolk, 2010)[2]. Specifically, 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. The total number of tweets in the collected corpus was 14,801,607 and the total number of words in the corpus was 224,623,447.

The following query terms (denoting temporal duration measure) were used to extract tweets containing temporal duration from the Twitter stream:

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

Tweets were normalized, tokenized, and then tagged for POS, using the NLTK Treebank Tagger (Bird & Loper, 2004). Each tweet came with a unique tweet ID number provided by Twitter and this ID was used to inform whether or not there were duplicate entries in the dataset, and all duplicate entries were removed. The twitter stream was also filtered so that it did not include re-tweets (tweets that have been reposted to Twitter).

3.2 Extraction Frames

To associate a temporal duration with each verb, the verbs and durations were matched 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 by Gusev et al. (2011). Unlike Gusev et al. (2011) four different extraction frames were used (*for*, *in*, *spend,* and *take*) with varied tense and aspect on each frame, in an effort to widen the coverage of extractions compared with that of Gusev et al. (2011). Each of the four frames were associated with a set of regular expressions to match and extract verbs for two tenses (past and present), and three different aspects (simple, perfect, and progressive). Durations could match spelled out numbers (one hour), hyphenated numbers (twenty-one minutes), or digits (30 minutes).

FOR: The for-adverbial extraction frame was designed to match two tenses and three aspects. The regular expressions accounted for variation in the word ordering. Consider some simplified pattern examples below, which show varied word order and tense-aspect combinations:

- *John ran for 10 minutes*
- *for ten minutes Sally was running*

IN: The in-adverbial extraction frame is tricky for extracting durations because the in-adverbial is sometimes used to describe pending events or things that are about to happen, such as, "Sally is going to the store in 5 minutes". However, I wanted to avoid collecting durations for future events. Therefore any verbs that matched the in-adverbial extraction frame were restricted to match the perfect aspect with any tense or the past tense and with any aspect, to indicate that a given event has been completed.

SPEND/TAKE: The tense and aspect were not restricted and the tweets were matched for tense and aspect on *spend* and *take*. In these cases the durations were syntactically associated with spend and take whereas semantically, the durations were associated with the verb in the complement clause (*read*, *work*, etc.). Variations in word order, like that found in examples of the *for* extraction frame, were not allowed for tweets matching the *spend* extraction frame. We see in the examples below that the verb is *read* and the tense and aspect in each of the examples were found to be past progressive:

- *Susie was spending 30 minutes reading*

[2] This Python module is available open-source at: https://bitbucket.org/runeh/tweetstream/src/

- *Susie was taking 5 minutes to read it*

3.3 Post-Processing Extracted Tweets

There were several steps to the post-processing of tweets. First, I identified the verb lemmas using NLTK WordNet (Bird and Loper, 2004). Verb lemmas that occurred less than 100 times were removed.

Next, all of the durations-measures were converted into seconds using a separate set of regular expressions. Instances where the duration lasted for longer than 1 billion seconds were removed. There were 6,389 tweets that met this condition. These tweets were removed in an attempt to avoid figurative speech that can occur on Twitter. So tweets such as the ones shown in (2) and (3) were removed:

(2) *I hate when I order food and it takes 2009779732 years to come*

(3) *I think my iTunes library is too big, it takes 7987694564 years to open*

Not all of the temporal durations that were extracted were numerically measured. Tweets that contained indefinite determiners *a* or *an* were treated as having a value of 1 temporal unit so that the noun phrase "an hour" could be converted to *3600 seconds*. There were 51,806 such tweets. Some of the tweets contained expressions such as: "some hours", "many hours", and "several hours". In cases like these, the duration was treated as having a value of based on its temporal unit so that durations like "many hours" were treated as *one hour*. This was applied to all of the temporal durations that were not numerically measured[3].

In addition, tweets that matched more than one extraction frame were removed. After the post-processing stage 390,562 tweets were extracted that covered 486 verb lemmas.

3.4 Extraction Frame Evaluation

Extraction frame precision was estimated for each frame by hand-annotating a randomly selected sample and labeling each extracted tweet as relevant if the duration, tense, aspect and verb were identified. The extraction frames performed overall with 90.38% precision, estimated from a sample size determined by the two-tailed t-test for proportions with 95% confidence (n=400, p=0.05). The extraction frame precision is reported below in Table 1.

Extraction Frame Type	**Estimated Precision**	**# Tweets**
for	91.25%	270,624
in	72.25%	83,061
spend	99.75%	2,593
take	98.25%	34,284
Overall	90.38%	390,562

Table 1. Number of extracted tweets

4 Analysis of Durations

4.1 Duration Distributions

Twitter is a lucrative resource for gathering typical durations associated with common verbs at a fine-grained level. Some verbs were found to have a very short mean duration (consider *rain* and *snooze*) while some had a longer mean duration (consider *live* and *work*), shown in Table 2.

Short durations		**Long durations**	
doze	32,721	*grow*	197,921,586
jog	405,550	*smoke*	246,557,468
cough	4,756,427	*live*	247,274,960
rain	4,994,776	*marry*	312,000,000
meet	40,503,127	*exist*	341,174,881

Table 2. Mean durations (in seconds) for a sample of verb lemmas

The following plots (Figures 1-3) show the frequency distribution for three different lemmas: *wrestle*, *say*, and *boil*. Similar to the work done by Pan et al. (2011) and Gusev et al. (2011), this research also shows that some of the duration distributions are bimodal. Gusev et al. (2011), Pan et al. (2011), and recent work by Williams and Katz (2012) show that some bimodal distributions could be associated with iterative events or habituality.

[3]There were 35,553 tweets matching this criteria.

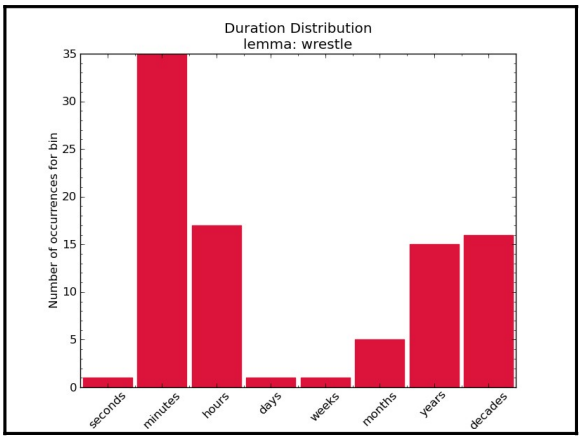

Figure 1. Distribution for *wrestle*, typically takes minutes or years

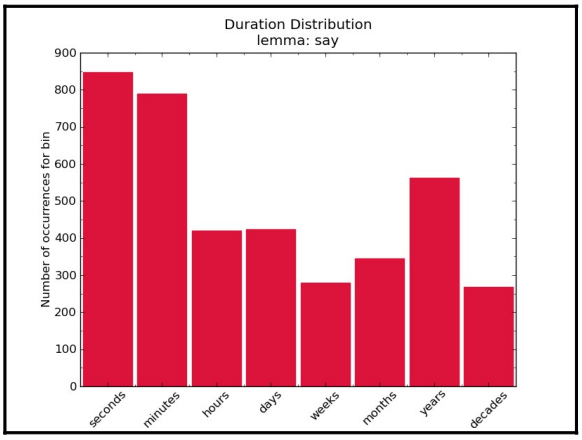

Figure 2. Distribution for *say*, typically takes seconds or years

The bimodal distributions for *wrestle* and *say* could possibly indicate that there are two phenomena present in the distributions: durations for events, and durations for habits. Consider that the sentence "John wrestled for half and hour with his kids" describes an event whereas the sentence "John wrestled for 30 years as a pro" describes a habit. An analysis of the relationship between bimodal distributions and habituality would provide more information in future work.

Not all of the distributions are bimodal, in fact we can see that is the case with the distribution for *boil*. Users of Twitter are not usually reporting long durations for that verb, but they do in several rare cases. This could be due to the effects of figurative speech, as in "John has been making by blood boil for decades".

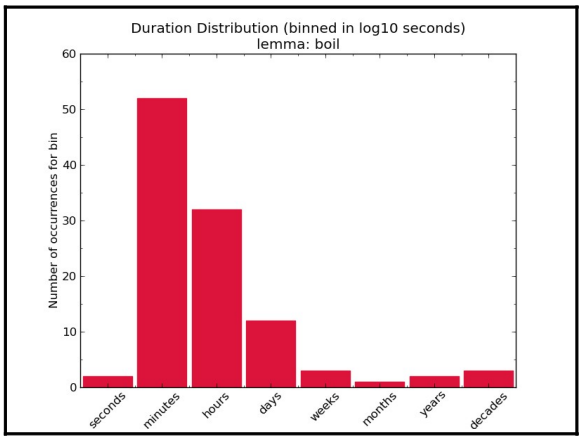

Figure 3. Distribution for boil, typically takes minutes

4.2 Comparison of Previous Work

To compare my work with Gusev et al., (2011), I found the overlap of verb lemmas. There were 356 verb lemmas in common. I calculated the log10 of each mean duration associated with each verb lemma, for my data and theirs. I plotted my means versus their means and I used linear regression to find a best fit line. The Pearson correlation value was 0.46 ($p < 0.01$), which suggests a weak positive correlation. Some of the outliers that we see in Figure 4 correspond to the following verb lemmas: *freeze*, *judge*, *age*, *double*, *load*, *lock*, *revise*, *score*, *heat*, *remove*, *lose*, *meet*, *head*, *ring*, *skate*, *yell*, and *fall.*

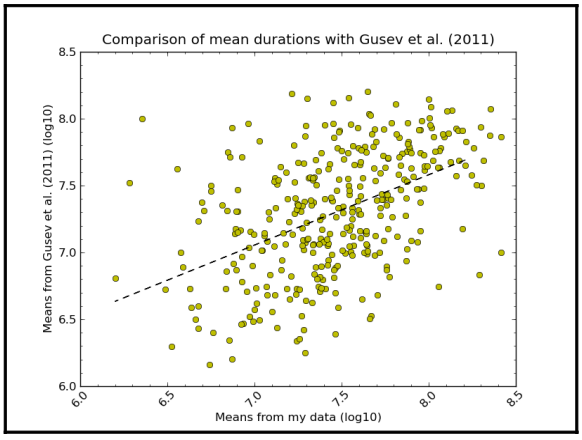

Figure 4. Mean durations vs. Gusev et al. (2011) in log10 seconds

5 Discussion

This paper has presented a new method to automatically extract duration information for verbs using data from Twitter. The four extraction frames used here were 90.25% accurate. This indicates that regular expressions can be applied to tweets to associate an event with its duration. Comparison with previous work shows that there is a positive correlation, and this indicates that the method presented here is nearly comparable. Corpora, extracted tweets, durations, and other materials used in this study will be made publicly available at the following website:

https://sites.google.com/site/relinguistics/

6 Future Work

There were several aspects of natural language that were put aside in this research. Future work should compare how the duration distributions are affected by modality, negation, and the future tense/aspect combinations. And, although I briefly addressed the presence of figurative language, this work could benefit from knowing which tweets were figurative, since this may affect how we examine typical durations.

Only four types of extraction frames were used in this study. More work is needed to find out if there are other extraction frames that can be used for this same task, and exactly which extraction frames should be used under various circumstances. Future work could also address the combinatorial effects of modality, negation, future tenses, and verb arguments with typical duration. Events like "John might finish writing his email soon" and "John might finish writing his memoir soon" will have different kinds of durations associated with them.

Looking at the distributions presented here, it is not clear where the boundary might be between single episodes, iterative events or habits. This kind of distinction between habits and events could prove to be important because an event such as *exist* can go on for years, decades or centuries, and in some cases *exist* might only last for a few seconds – but we would not say that *exist* is a habit. At the same time, the frequency distribution for *wrestle* in Figure 1 indicates that the event *wrestle* lasts for hours, but the fact that it is reported to last for years suggests that there are some habits in the collected data.

References

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)*.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. "Identifying sarcasm in Twitter: a closer look". In P*roceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 581–586), Portland, Oregon, June 19-24.

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.

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.

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. 2011. "Annotating and Learning Event Durations in Text." *Computational Linguistics*. 37(4):727-752.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. "Sentiment in Twitter events." *Journal of the American Society of Information Science and Technology*, 62: 406–418. doi: 10.1002/asi.21462

Jennifer Williams and Graham Katz. 2012. "Extracting and modeling durations for habits and events from Twitter". In *Proceedings of Association for Computational Linguistics, ACL 2012*. Jeju, Republic of Korea.