Using Textual Patterns to Learn Expected Event Frequencies

Jonathan Gordon
Department of Computer Science
University of Rochester
Rochester, NY, USA
jgordon@cs.rochester.edu

Lenhart K. Schubert
Department of Computer Science
University of Rochester
Rochester, NY, USA
schubert@cs.rochester.edu

Abstract

Commonsense reasoning requires knowledge about the frequency with which ordinary events and activities occur: How often do people eat a sandwich, go to sleep, write a book, or get married? This paper introduces work to acquire a knowledge base pairing factoids about such events with frequency categories learned from simple textual patterns. We are releasing a collection of the resulting event frequencies, which are evaluated for accuracy, and we demonstrate an initial application of the results to the problem of knowledge refinement.

1 Introduction

A general problem in artificial intelligence is knowledge acquisition: AI applications require both a background of general, commonsense knowledge about the world and the specific knowledge pertaining to the application's domain. This knowledge needs to be available in a form that facilitates reasoning, and it needs to be of high quality. While early work tended to hand-code knowledge – and this continues to be the preferred method for projects like Cyc (Lenat, 1995) – this is labor-intensive and neglects the systematic connection that can be made between natural language and representations suitable for inference. However, most efforts to acquire knowledge from text, such as KNEXT (Schubert, 2002), TEXTRUNNER (Banko et al., 2007), or DART (Clark and Harrison, 2009), are underspecified in a number of important respects, including word sense, quantificational structure, and the likelihood of their conclusions.

In this paper, we address the lack of information about the expected temporal frequency of ordinary events. While Gordon and Schubert (2010) addressed the problems of quantificational structure and strength for refining knowledge learned with KNEXT, they distinguish only three kinds of temporal predications:

- those that hold for the existence of the subject (*individual-level*), e.g., a house being big;
- those that hold at a specific moment in time (*non-repeatable stage-level*), e.g., a person dying; and
- those that hold at multiple moments in time (*repeatable stage-level*), which they quantify as "occasional" events, e.g., a person drinking a cup of coffee.

Repeatable stage-level predications vary from those done with great frequency, such as a person saying something, to those done quite infrequently, such as a woman giving birth. We will describe a simple method to learn rough frequencies of such events from text.

Our focus is on the commonsense knowledge needed for many AI applications, rather than more specific domain knowledge. This work looks for the frequency of everyday events – such as going to work – that might be mentioned in ordinary text like newspaper articles, rather than big events – like earthquakes devastating a city, which tend to be rare and unpredictable – or small events – like atoms decaying, which would typically escape our notice.

We are unaware of any previous work aimed at systematically learning the expected or normal frequency of events in the world. However, our basic approach to this problem aligns with a long-running line of work using textual references to learn specific kinds of world knowledge. This approach has been popular at least since Hearst (1992) used lexico-syntactic patterns like 'NP_0 such as $\{NP_1, NP_2, \ldots,$ (and|or)$\}$ NP_n' to learn hyponym relations, such as 'Bambara ndang is a bow lute' from large text corpora.

In addressing the problem of quantificational disambiguation, Srinivasan and Yates (2009) learn the expected sizes of sets of entities that participate in a relation; e.g., how many capitals a country has or how many cities a person tends to live in. They do this by using buckets of numeric phrases in hand-crafted extraction patterns like '(I|he|she) ⟨word⟩+ ⟨numeric⟩ ⟨noun⟩', which would match 'she visited four countries'. They apply these patterns to Google's Web1Tgram Corpus of n-grams.

Gusev et al. (2011) presented a similar approach to learning event durations using query patterns sent to a Web search engine, e.g., '$\langle \text{event}_{\text{past}}$ for * ⟨bucket⟩', where the bucket is a category in [seconds, minutes, hours, ..., decades] for classifying the event's expected duration. Both of these papers are notable for gaining wide coverage by indirectly using Web-scale text. However, they are limited by the brevity of patterns in n-grams and by the coarse matching abilities of Web queries, respectively. We will discuss these trade-offs and our approach, focusing on large offline corpora, in Section 2.

The contribution of this paper is the application of a traditional technique to a new problem. Temporal frequencies are of key importance to improving the quality of automatically learned knowledge for commonsense reasoning. Additionally, we hope that providing a knowledge base of expected frequencies for factoids about everyday events will serve as a new resource for other work in knowledge extraction and reasoning.

2 Textual Patterns of Frequency

The most direct linguistic expression of temporal frequency comes from frequency adverbs: words like *usually* and *always*, distinct in their meaning from other adverbs of quantification like *twice*. Sentences that contain a frequency adverb are referred to as *frequency statements*, e.g., 'John sometimes jogs in the park.' Frequency statements are interesting because their truth depends not just on the existence of some past events that support them but on a regular distribution of events in time. That is, saying that John 'sometimes jogs' means that it is a habitual rather than incidental activity.

As Cohen (1999) observes, much of our knowledge about the world is expressed through frequency statements, but it's not entirely clear what these sentences mean. From the perspective of knowledge extraction, they can seem quite opaque as their meaning seems to rely on our pre-existing ideas of what a normal temporal frequency for the event would be. For instance, to say that 'Mary snacks constantly' (or 'frequently' or 'occasionally') only makes sense if you already have in mind some range of frequencies that would be normal or unremarkable.

More absolute frequency adverbials, such as *daily*, *weekly*, or *every other week* avoid the problem of depending on a person's expectations for their meaning. However, these tend to occur with extraordinary rather than ordinary claims. For instance, in the British National Corpus we see

> 'Clashes between security forces and students had occurred almost daily.'
> 'New [viruses] are discovered every week'

Both of these are expressing surprising, unexpected information.

Following the example of Gordon and Schubert (2011) in considering "defied expectations", we look for textual expressions that indicate a person's frequency expectation has not been met and, looking at these in aggregate, we conclude what the original, implicit expectation is likely to have been. An example of such a defied expectation is

> 'Bob hasn't slept in two days.'

The production of sentences like this suggests that this is an unusually long gap between sleep periods for most people. We are unlikely to find many sentences saying, e.g., 'Bob hasn't slept in 2 hours' as this would not defy our expectation. (And while we will find exaggerations, such as 'I hadn't slept in weeks', the classification technique we describe

will favor the smaller interval unless the counts for a longer interval are quite high.)

In this initial approach, we make use of two other patterns indicating temporal frequency. An additional indication of an upper-bound on how infrequent an event tends to be is a reference to the last time it was completed, or the next time it's anticipated, e.g., 'He walked the dog yesterday' or 'She'll go to the dentist next month'.

The other pattern is the use of *hourly*, *daily*, *every week*, etc. While frequency statements with such adverbs can be communicating a frequency that's much higher or lower than expected, they serve as an important source of information when we don't find matches for the defied expectations. They also occur as prenominal modifiers: For a factoid like 'A person may eat bread', we want to match references to 'his daily bread'. This use is presumptive and, as such, indicates a usual or expected frequency, as in 'our weekly meeting' or 'the annual conference'.

Method

Rather than relying on query-based retrieval from the Web, or on the use of n-gram databases, we have chosen to process a selection of large text corpora including the Brown Corpus (Kučera and Francis, 1967), the British National Corpus (BNC Consortium, 2001), the Penn TreeBank (Marcus et al., 1994), Gigaword (Graff et al., 2007), a snapshot of English Wikipedia (Wikipedia, 2009), a collection of weblog entries (Burton et al., 2009), and Project Gutenberg e-books (Hart and volunteers, 2006).

The motivation for doing so is the larger context offered and the flexibility of matching. Search engine queries for patterns are limited to quoted strings, possibly containing wildcards: There's no reasonable mechanism to prevent matching patterns nested in a sentence in an unintended way. For instance, searching for 'I hadn't eaten for months' can easily match not just the expected hyperbole but also sentences like 'I felt like I hadn't eaten for months'. Sets of n-grams pose the problem of limiting pattern length. While it's possible to chain n-grams for longer matches, this forfeits the guarantee of any actual sentence containing the match.

As a set of appropriate, everyday events abstracted away from specific instances, we used a corpus of factoids learned most frequently by the KNEXT knowledge-extraction system.[1] We heavily filtered the knowledge base both for quality (e.g., by limiting predicate names to known words) and to focus on those factoids describing the sort of action to which we want to assign a frequency. This included removing passives ('A person may be attacked') and subjects that aren't causal agents (according to WordNet). We abstracted multiple subjects to low common hypernyms for compactness and to focus on classes of related individuals, such as 'a parent', 'an executive', or 'a scholar'.

A good indication that a factoid can be annotated with a frequency is telicity: Telic verb phrases describe events rather than continuous actions or states. To check if the predication in a factoid is possibly telic, we look in the Google n-gram data set for short patterns. For each factoid of form (X Y Z*) and each set of indicators S,

(quickly|immediately|promptly)
(suddenly|abruptly|unexpectedly)
(inadvertently|unintentionally|deliberately|
unwittingly|purposely|accidentally)
(repeatedly|frequently)

we look for: 'S X Yed Z*', 'X Yed Z* S', and 'X S Yed Z*' where X is the subject, Yed is the past tense of the verb, and Z consists of any arguments. Any factoid with non-zero counts for more than one set of indicators was considered "possibly telic" and included for frequency extraction.

For each possibly telic factoid, we first determine whether it describes a regular event or not. A regular event doesn't need to be a rigid, scheduled appointment, just something done fairly consistently. 'Brush your teeth' is regular, while 'Overcome adversity' is not, depending instead on some scenario arising. Regularity can be indicated explicitly:

Ys/Yed regularly/habitually
Ys/Yed invariably/inveterately/unvaryingly
Ys/Yed like clockwork
Ys/Yed at regular intervals

It can also be suggested by a stated interval:

[1] Collections of KNEXT factoids can be browsed and are available for download at http://cs.rochester.edu/research/knext. Larger collections are available from the authors on request.

Ys/Yed hourly/daily/weekly/monthly/yearly/annually
Ys/Yed every hour/day/week/month/year
every hour/day/week/month/year X Ys/Yed

If we do not match enough of these patterns, we don't consider the factoid to be regular: It may be an occasional or existence-level predication, or we may just lack sufficient data to determine that it's regular.

For each regular-frequency factoid, we then check the corpora for matches in our three categories of patterns:

Explicit Frequency Matches These indicate the exact frequency but may be hyperbolic. The 'hourly' and 'every hour' style patterns used for checking regularity are explicit frequency indicators. In addition, if the factoid contains 'may have a Z', we search for the prenominal modifiers:

's/his/her/my/your/our
hourly/daily/weekly/monthly/yearly/annual Z

Defied Expectation Matches These indicate that people expect the activity to be done "at least *bucket* often". These include many small variations along these lines:

Hourly/multiple times a day:
Has X Yed this morning/afternoon/evening?
Didn't X Y this/last/yesterday morning/afternoon/evening?
Hasn't Yed for/in over an hour
Has not Yed for the whole/entire day

Daily/multiple times a week:
Have X not Yed today?
Did X not Y today/yesterday?
Had not Yed for/in more than N days
Haven't Yed for the whole/entire week

Weekly/multiple times a month:
Haven't X Yed this week?
Didn't X Y this/last week?
Hadn't Yed for more than a week
Had not Yed for the whole/entire month

Monthly/multiple times a year:
Hasn't X Yed this month?
Did X Y this/last month?
Hadn't Yed for over N months
Hadn't Yed for the whole/entire year

Yearly/multiple times a decade:
Have X Yed this year?
Didn't X Y this/last year?
Haven't Yed for/in over a year
Hadn't Yed for an entire decade

Last Reported Matches These are statements of the last time the predication is reported as being done or when it's expected to happen next. These are useful, as you wouldn't say 'I took a shower last year' if you take one daily. They indicate that the event happens "at most *bucket* often".

Hourly/multiple times a day:
Yed an hour ago
Yed earlier today
'll/will Y later today

Daily/multiple times a week:
Yed today/yesterday
Yed on Sunday/.../Saturday
'll/will Y tomorrow/Sunday/.../Saturday
'll/will Y on Sunday/.../Saturday

Weekly/multiple times a month:
Yed this/last week(end)
'll/will Y next week(end)

Monthly/multiple times a year:
Yed this/last month
'll/will X next month

Yearly/multiple times a decade:
Yed this/last year/season/spring/.../winter/January/.../December
'll/will Y next year/season/spring/.../winter/January/.../December

Decision For each of the three categories of patterns, we select the frequency bucket that it most strongly supports: We iterate through them from *hourly* to *yearly*, moving to the next bucket if its count is at least 2/3 that of the current one. For the 'last reported' matches, we go in the opposite direction: *yearly* to *hourly*.

From the three choices, the two buckets with the highest supporting counts are selected. If the range of these buckets is wide (that is, there is more than one intervening bucket), the bucket for a more frequent reading is chosen; otherwise, the less frequent one is chosen. This choice compensates for some hyperbole: If people claim they haven't slept for *days* and for *years*, we choose *days*. However, if we find that people haven't showered for *hours* or *days*, we choose days as a reasonable lower bound.

3 Evaluation

To evaluate how accurately this method assigns an expected frequency to a factoid, we sample 200 fac-

toids that were classified as describing a regular occurrence. Each of these is verbalized as a conditional, e.g.,

> If a person drives taxis regularly, he or she is apt to do so daily or multiple times a week.
>
> If a male plays (video games) regularly, he is apt to do so daily or multiple times a week.

Note that we do not take the factoid to apply to all possible subjects, but for those it applies to, we're indicating our expected frequency. Arguments are taken to be narrow-scope, e.g., for 'a person may greet a friend', it can be a different friend for each greeting event rather than the same friend every time.

For each of the sampled factoids, two judges evaluated the statement "This is a reasonable and appropriately strong frequency claim (at least on some plausible understanding of it, if ambiguous)."

1. Agree
2. Lean towards agreement.
3. Unsure.
4. Lean towards disagreement.
5. Disagree.

The average rating for Judge 1 was 2.45, the average rating for Judge 2 was 2.46, and the Pearson correlation was 0.59.

A simple baseline for comparison is to assign the most common frequency ('daily') to every factoid. However, for this to be a fair baseline, this needs to be done at least for the entire possibly-telic KB, not just the factoids identified as being regular, as that classification part of the method being evaluated.This baseline was evaluated for 100 factoids, with an average ratings of 3.06 and 3.51 (correl. 0.66) – worse than 'unsure'. This result would be even lower if we applied this frequency to all factoids rather than just the telic ones: We would claim, for instance, that a person has a head daily.

The authors also judged a random sample of 100 of the factoids that were marked as not being regular actions. These were verbalized as denials of regularity:

> Even if a person files lawsuits at all, he or she doesn't do so regularly.

Of these, on average the judges indicated that 30 could reasonably be thought to be regular events that we would like to assign a frequency to.

Based on these encouraging preliminary results, we are releasing a corpus of the annotations for 10,000 factoids. This collection is available for download at http://cs.rochester.edu/research/knext.

One anticipated application of these annotations is as a guide in the sharpening (Gordon and Schubert, 2010) of KNEXT factoids into full Episodic Logic forms. For instance, from the factoid 'A person may eat lunch', we can select the correct episodic quantifier *daily*:

```
(all-or-most x: [x person.n]
   (daily e
      (some y: [y lunch.n]
         [[x eat.v y] ** e])))
```

That is, for all or most persons, there is a daily episode that is characterized by the person eating some lunch.

4 Future Work

There is room to improve the frequency labeling, for instance, using machine-learning techniques to combat sparsity issues by discovering new textual patterns for event frequencies. It would also be interesting to see how performance could be improved by automatically weighting the different patterns we've discussed as classification features.

5 Conclusions

The acquisition of temporal frequency information for everyday actions and events is a key problem for improving automatically extracted commonsense knowledge for use in reasoning. We argue that this information is readily available in text by looking at patterns expressing that a specific instance is at odds with the expected frequency, those that report frequencies explicitly, and those stating the last time such an event occurred. We find that a simple approach assigns event frequencies with good accuracy, motivating the release of an initial knowledge base of factoids with their frequencies.

Acknowledgements

This work was supported by NSF grants IIS-1016735 and IIS-0916599 and a subcontract to ONR STTR contract N00014-10-M-0297.

References

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the Web. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

BNC Consortium. 2001. The British National Corpus, v.2. Dist. by Oxford University Computing Services.

Kevin Burton, Ashkay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Peter Clark and Phil Harrison. 2009. Large-scale extraction and use of knowledge from text. In *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP 2009)*, pages 153–60.

Ariel Cohen. 1999. Generics, frequency adverbs, and probability. *Linguistics and Philosophy*, 22(3):221–253.

Jonathan Gordon and Lenhart Schubert. 2010. Quantificational sharpening of commonsense knowledge. In *Proceedings of the AAAI 2010 Fall Symposium on Commonsense Knowledge*.

Jonathan Gordon and Lenhart Schubert. 2011. Discovering commonsense entailment rules implicit in sentences. In *Proceedings of the EMNLP 2011 Workshop on Textual Entailment*.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. *English Gigaword*. Linguistic Data Consortium.

Andrey Gusev, Nathanael Chambers, Pranav Khaitan, Divye Khilnani, Steven Bethard, and Dan Jurafsky. 2011. Using query patterns to learn the duration of events. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS '11)*, pages 145–154, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael S. Hart and volunteers. 2006. Project Gutenberg. www.gutenberg.org.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545.

Henry Kučera and W. N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.

Douglas B. Lenat. 1995. Cyc: A Large-scale Investment in Knowledge Infrastructure. *Communications of the Association for Computing Machinery*, 38(11):33–48.

Mitchell P. Marcus, Beatrice Santorini, and Mary A. Marcinkiewicz. 1994. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Lenhart K. Schubert. 2002. Can we derive general world knowledge from texts? In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT02)*.

Prakash Srinivasan and Alexander Yates. 2009. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wikipedia. 2009. English Wikipedia snapshot, 2009-07-09. en.wikipedia.org.