Unsupervised Biographical Event Extraction Using Wikipedia Traffic

Alexander Hogue **Joel Nothman** **James R. Curran**
ə-lab, School of Information Technologies
University of Sydney
NSW 2006, Australia
{ahog5691@uni.,joel.nothman@,james.r.curran@}sydney.edu.au

Abstract

Biographical summarisation can provide succinct and meaningful answers to the question "Who is X?". Current supervised summarisation approaches extract sentences from documents using features from textual context.

In this paper, we explore a novel approach to biographical summarisation, by extracting important sentences from an entity's Wikipedia page based on internet traffic to the page over time. Using a pilot data set, we found that it is feasible to extract key sentences about people's notability without the need for a large annotated corpus.

1 Introduction

"What is Julian Assange known for?" is a question which can be answered in many ways. Previous computational approaches to answering questions like these have focused on summarisation: selecting a subset of sentences from a group of documents relating to a person and ordering them (Biadsy et al., 2008; Zhou et al., 2004). Full text summaries do provide some insight into the notability of their subject, but can also contain superfluous information. To pinpoint the notoriety of individuals, we aim to extract the sentences from a document which show how the document's subject is notable.

We provide an alternate, unsupervised approach to the broader task of biography abstraction, which exploits external information about text, rather then extracting textual features directly.

In this paper, we respond to "What is Julian Assange known for?" with sentences mentioning important events which have occurred in his life. It is the breadth of possible reasons for which one could be notable that make supervised approaches to this task difficult — the creation of a corpus large enough to cover the range of notable events would require a prohibitive amount of annotation. Furthermore, this task would be tedious for annotators, since sentences expressing notability are sparse among documents, and some high-level understanding of the notability of the page subject is required to judge each sentence's notability.

We hypothesise that when a notable event happens to a person, traffic to their Wikipedia page peaks abruptly, and an edit is made to their page describing the event.

To explore this hypothesis, a simple outlier-based method is applied to extract peaks (short periods of sudden activity) from Wikipedia page traffic data, which are used to locate page edits which align to sentences contributing to the notability of the page subject. Event reference identification is a difficult task (Nothman, 2014), and errors in event extraction may mask the performance of our system, so in our initial approach we choose the sentence as our unit of event description.

We evaluate by creating a corpus of Wikipedia pages about people. Each sentence annotated with its human-judged significance to the person's notability. We then measure how reliably page traffic data can be used to identify these most notable events. Our initial investigation into extracting key sentences has shown that it is feasible to approach the task in this unsupervised manner.

Exploring the relationship between Wikipedia traffic, page edits, and the occurrence of notable events can provide us with a deeper understanding of how the public respond to events, and an extrinsic source of information on the importance of sentences in Wikipedia articles.

2 Background

The goal of many approaches to biography abstraction is to provide some distilled knowledge on the notability of a person. A simple approach to biographical abstraction is to summarise exist-

Alexander Hogue, Joel Nothman and James R. Curran. 2014. Unsupervised Biographical Event Extraction Using Wikipedia Traffic. In *Proceedings of Australasian Language Technology Association Workshop*, pages 41−49.

ing documents about an entity, selecting the most representative content of the text while adhering to length constraints.

Early approaches to the task train a sentence classifier (Teufel and Moens, 1997) on a corpus of sentences which are in some way biographical. This corpus is typically existing biographies, or manually selected sentences from a larger corpus. Previous work has used Wikipedia as large, alternate source of biographical sentences (Biadsy et al., 2008), hypothesising that most sentences in Wikipedia's articles about people are biographical.

Zhou et al. (2004) experiment with non-binary sentence classification, requiring a summary to have at least one sentence of each category in a "biographical checklist", with categories such as work, scandal, and nationality. Training a classifier to categorise sentences into these classes requires costly manual annotation. A similar effort would be required for a supervised learning approach to extract important biographical sentences.

Biographical abstraction has also been approached as a relation extraction task. DIPRE (Brin, 1999) has been an influential pattern extraction system which bootstraps using a small set of seed facts to extract not only the patterns they represent, (e.g.@ [person] WORKS_FOR [organisation]) but also to extract additional patterns. Liu et al. (2010) presented BIOSNOWBALL for the biographical fact extraction domain, which extracts biographical key-value pairs. It is the wide range of reasons for notoriety (which would require a large number of potential patterns to fill) motivating our novel source of measures of importance — Wikipedia page traffic over time.

Rather than the traditional approach of classifying sentences via textual features (Schiffman et al., 2001) or locating events (Filatova and Hatzivassiloglou, 2004), we explore the use of an extrinsic source of information indicating what is interesting. Motivating this approach is our hypothesis that many people are most well-known for the events they were involved in. These events have previously been ordered temporally by supervised learning from textual features, (Filatova and Hovy, 2001), and our extrinsic information may assist with the temporal location of events with little temporal information mentioned in text.

Various features of Wikipedia have been previously exploited in NLP, since they provide a massive source of human-written semi-structured information. Plain text has been used to assist named entity recognition (Nothman et al., 2013), page categories have been used to create an ontology (Suchanek et al., 2007) and infoboxes (key-value pairs of facts) have been used to provide additional context to information in text (Wu and Weld, 2010). Wikipedia's revision history is exploited less frequently, but has proven useful to train a model of sentence compression (Yamangil and Nelken, 2008). We know of no prior work that aligns page traffic to text in Wikipedia.

2.1 Timeseries Analysis

To exploit the Wikipedia page traffic data, we need to extract peaks from timeseries data. There are many definitions of *peaks* in the literature on timeseries peak extraction, and many approaches to detecting them. Motivating much of this research is the need to automatically detect spikes in Electroencephalography results (EEG) (Wilson and Emerson, 2002). EEG peaks are typically moderate in amplitude, whereas spikes in page traffic are often several standard deviations above the mean, so our peaks are easier to detect.

A simple approach is to keep a moving average over some window of previous points, comparing each point to the average of the window of previous points. Vlachos et al. (2004) employ this approach using only two window sizes (short term and long term) to detect high traffic periods for the MSN search engine. Their results show instances where a peak in search traffic appears at the time notable events occur to some entities (for instance, the death of a famous British actor), which has also been observed in both page traffic and edits by Nunes et al. (2008). This approach suits our task since the peaks we wish to detect are so prominent.

Through peak extraction techniques, we extract the date on which important events happened to people. By extracting edits to Wikipedia articles near the time of these peaks, we can find the single sentence in the current version of the article which is most similar, and associate it with the important event which happened at the time of the peak.

By providing a sentence-level summary of key events which occur to a person, we pinpoint the notoriety of individuals without the need for a hand-annotated training corpus.

3 Edits and Events

Before considering page traffic data, we performed a preliminary manual analysis of the relationship between the occurrence of real-world events and views and edits to the relevant Wikipedia pages.

Wikipedia includes yearly summaries of key events, their date of occurrence and main participants. We randomly sampled people from these pages for the years 2008–2013,[1] and inspected edits made to those people's Wikipedia entries around the event's date. Findings from our analysis follow:

Editors are quick to respond We manually investigated the typical delay between an event happening and a person's Wikipedia page being updated to reflect the event. The time difference between the recorded date of the event and the date of the page edit mentioning the event was manually recorded, and we found that the typical delay was less that 1 day in 19 of 20 cases. Note that this experiment only considered events notable enough to appear in a short summary of the year, so this result may not generalise to less notable people.

Edits occur in bursts We observed a pattern in the distribution of page edits in response to popular events. Before the day of the event, edits are sparse and mostly minor. On the day of the event, the earliest edit tends to briefly describe the event (e.g. On May 31, he was shot). This edit is followed by a burst of edits soon after, with the volume and frequency of edits correlated with to the notability of the event.

Edits are iteratively mutated Within a burst of edits, consecutive edits consist of modifications to the original edit, new information as the story unfolds, vandalism, reversion of vandalism, updates to outdated sections of the article, and elaboration on sections of the article unrelated to the event. Nunes et al. (2008) have also observed this phenomenon, noting that when a burst of edits occurs, editors tend to contribute "updates on the specific event and generic revisions to the whole topic".

The text introduced at the time of the event may have been heavily modified, or even removed completely as the page is updated over time. Often the original edit is lengthened, and this property has been previously used to create a training corpus for sentence compression (Yamangil and Nelken, 2008). For instance, for a candidate edit of George Tiller was shot on May 31, 2009, we might extract from the current-day article On May 31, 2009, Tiller was shot through the eye and killed by anti-abortion activist Scott Roeder.

Our overall method in the following section builds on these findings, but additionally relies on mapping page view data to edit data.

4 Method

Our task is to extract sentences corresponding to biographical events from Wikipedia articles. We do this by exploring the relationship between Wikipedia page view traffic timeseries and Wikipedia page edits. Specifically, we detect peaks in the page traffic timeseries data for each page, and search the edit history at the time of those peaks for an edit mentioning the event. Since the important sentences we identify are in a snapshot of Wikipedia substantially later than the edit, an alignment step is required, as the originally-inserted text may well be edited further.

4.1 Our Hypothesis

There are three related timeseries we explore in this paper: real-world events occurring to people, visits to their Wikipedia pages, and edits to those pages. Figure 1 shows the relationship between page view spikes and sentences in the article text mentioning the real-world events that caused them. We hypothesise that these timeseries relate such that when a notable event happens to a person, it is reported in the media, traffic to their Wikipedia page increases, and an edit is made to the page adding the occurrence of the event.

4.2 Wikipedia traffic

Wikipedia[2] provides hourly pageview counts[3] (the number of times each article was visited in each hour) for every article on the `wikipedia.org` domain[4] since December 2007. We import each year's worth of data into an instance of WiredTiger[5], a space-efficient key-value store.

Motivated by our experiment measuring the delay between an event occurring and an edit reflecting it being made (see Section 6), we combined

[1] 2008–2013 have coverage in Wikipedia traffic data.

[2] Also https://stats.grok.se/

[3] As well as the URL suffix from each HTTP request, regardless of status.

[4] https://dumps.wikimedia.org/pagecounts-raw/

[5] https://wiredtiger.org/

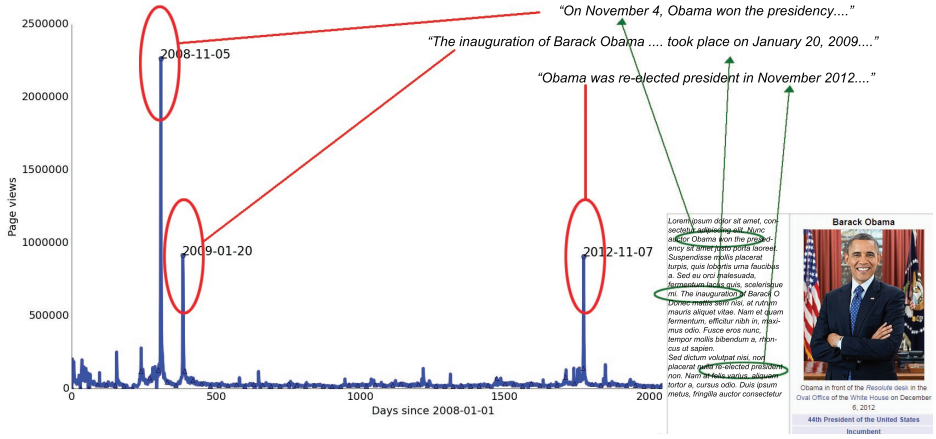

Figure 1: System overview. Peaks in the timeseries of page traffic data are used to find sentences in Wikipedia articles which express the notability of the page subject.

the hourly page view counts into daily counts. This also smooths our timeseries, averaging out the sinusoidal patterns at the hourly level which reflect the day/night cycle of the timezone which has the most Wikipedia readers. We also filter out any namespace modifiers (e.g. Category:People), and no longer existing articles as of 2014. There are some caveats — for instance, we cannot detect events which happened to an individual before their Wikipedia page was created. Most critically, since Wikipedia's page traffic statistics were first recorded in December 2007, much of the (timestamped) edit history for some pages does not have corresponding page view data.

4.3 Peak Detection

Once we have extracted the timeseries for a particular Wikipedia article, we then locate the dates on which the article received a spike in traffic. We use a simple standard deviation-based method to locate peaks in the timeseries data, computing the weighted average of previous points. For each point, the weights for each previous point decay exponentially.

Specifically, μ_i is defined for the ith point of the timeseries p_i by:

$$\mu_0 = p_0$$

$$\mu_i = dp_i + (1-d)\mu_{i-1}$$

Where $0 < d < 1$ is a dampening constant. For a timeseries with standard deviation σ, the ith point p_i is a peak if:

1. p_i is a local maximum
2. $p_i > \mu_i + \theta\sigma$

for a constant $\theta > 0$ determining the extent to which detected peaks differ from the mean, and where local maxima are defined simply as points larger than their immediate neighbours. So our peaks are maxima which are also outliers. Figure 2 shows the effects of varying θ on spike detection. Increasing θ linearly increases the magnitude a maximum must have to be considered a spike. Since our data forms a timeseries, each peak corresponds to a date. We can search Wikipedia's edit history at that time for edits potentially mentioning an event that may have caused the peak.

4.4 Edit Extraction and Selection

Given the date at which a spike in page traffic occurred, we next search for an edit to the page potentially mentioning this event. All edits to a particular page are stored in the page's revision history, and each edit is represented as additions and removals from the previous version of the page. Since we are searching for new information, we consider an edit as one or more additions to the page text (removal of text is ignored). There is a wide range of potential ways a page can be edited. Edits typically contain (a) new information (b) corrections to false information (c) vandalism (d) reversal of vandalism (e) spelling and grammar corrections. We observed that both long additions and elaborations as well as spelling/grammar edits tend to appear in the days after the announcement of a notable event as the increased page traffic prompts editors to update the article.

Motivated by this, we extract all edits within a

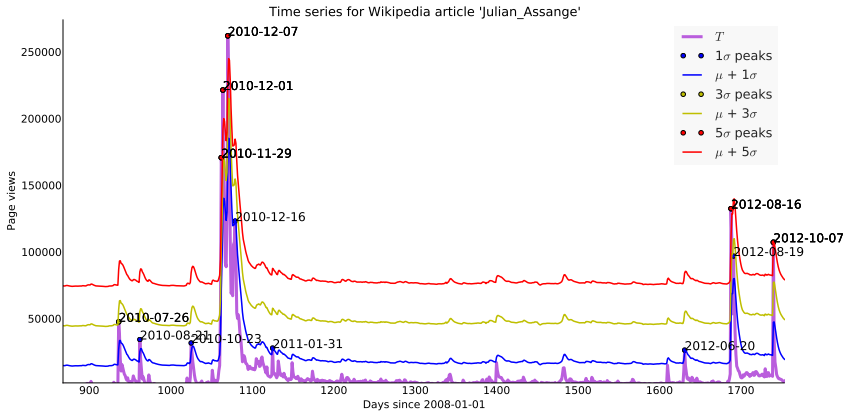

Figure 2: The effect of variation in σ on peak detection. In this image, T is the timeseries, σ is the T's standard deviation and μ is T's weighted moving average. ($d = 0.25$)

window spanning W days on either side from the reported date of the peak, and filter out additions within them which do not have between 5 and 100 words to create our set of candidate edits. The minimum size restriction is to account for small spelling/grammar edits or categorisation edits (e.g. the addition of `Category:Footballer`), and the maximum size restriction helps ignore vandalism (which often deletes the entire article) and its reversion, as well as rewrites which are much broader than the statement of a particular fact.

From this list of candidate edits, we associate the earliest edit within the window to our detected peak. Motivating this approach is the distribution of Wikipedia edits over time which we have observed when a notable event happens to someone, as discussed in Section 3. We observed that this edit is most likely to contain new information on the recent occurrence of a notable event.

Once we have obtained a candidate edit for each spike, we attempt to find the sentence in the current-day Wikipedia article which corresponds to the edit. Aware of the iterative mutation which occurs to edits from our analysis in Section 3, we associate with the edits around the time of a peak the most similar sentence in the current-day Wikipedia article. For each traffic peak we associate at most one important sentence. We measure the cosine similarity of bag-of-word representations of the edit and candidate sentence. To vectorise both the article sentences and our edit, we first remove stop words[6], and convert all text to lower-case. Non-alphanumeric tokens are then removed, and tokens are stemmed by the Porter stemming algorithm (Porter, 1980). Frequency-weighted cosine similarity scores $\in [0, 1]$ are computed between our candidate edit and each sentence in the current-day Wikipedia article, and the most similar sentence is returned.

5 Annotated Corpus

To evaluate our system, we created a manually-annotated test set comprising of a random sample of Wikipedia articles that (a) had less than 100 sentences (b) had the most frequently mentioned year[7] in the range $[2007, 2014]$ (c) had the most frequently mentioned year occuring at least twice (d) was categorised within Wikipedia's Category:1950 births to Category:2001 births.

We chose these restrictions to find Wikipedia articles about people who have had notable events occur within the time period for which we have page traffic data (2008–2013). The sentence restriction was made in order to control the amount of annotation work to be done, but has the side effect of choosing at most moderately notable people, since Wikipedia articles on popular people are substantially longer than 100 sentences. We note that key events relating to people of great fame are well documented, and it is extracting events for the long tail of less notable people which is more difficult.

[6]We use the English stop word list provided in NLTK (Bird, 2006).

[7]Any token that is a number from 1900 to 2020 is considered a year.

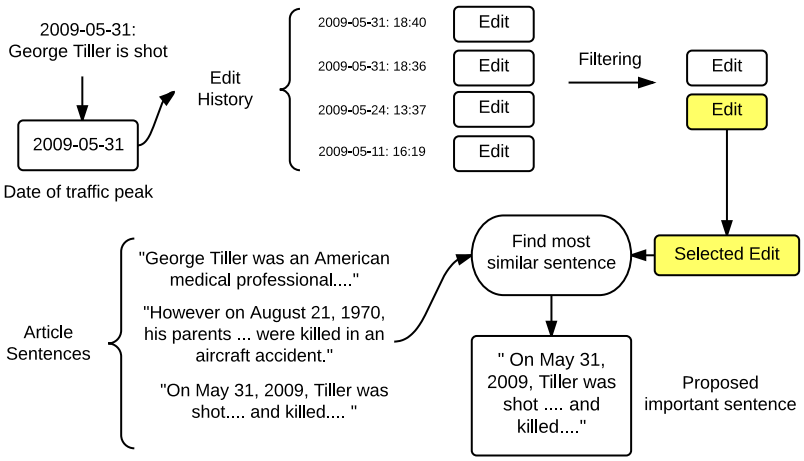

Figure 3: Process taken to align edits to sentences in the current-day Wikipedia article.

5.1 Annotation Procedure

To obtain only the sentences in Wikipedia articles, we parsed the Wikipedia[8] article markup to extract body paragraph text, and used Punkt (Kiss and Strunk, 2006) unsupervised sentence boundary detection trained on a collection of Wikipedia articles, as in Nothman et al. (2013). The sentences were imported into a web-based annotation tool, where 10 native English speakers were tasked with annotating sentences all from 10 Wikipedia articles with scores from 1 to 5 (or X) based on the contribution of the sentence to the notability of the page entity, with the X category for sentences which do not mention the page entity (e.g. facts about their family). We chose this numerical scheme since articles may differ in the number of important sentences, and a multi-category annotation (rather than binary) task allows annotators to calibrate and distinguish between major and minor notable events. We also asked annotators not to distinguish between the page subject, and groups of people the page subject is part of (for instance, Kurt Cobain and his band Nirvana). Example annotations are presented in Table 3.

An initial experiment on a single page suggested that annotators had difficulty distinguishing between facts and events. For instance, the sentence She joined the Labour party in 2006 could be interpreted as containing a fact (being *in* the Labour Party) or an event (*joining* the Labour Party). To allow annotators to focus on interpreting the contributions sentences had to notability, we did not ask annotators to differentiate between facts and events. Due to the redundancy in our annotation (10 annotators annotated each sentence), and the difficulty of the annotation task, we use a consensus-based method to interpret which sentences are important to the notability of the page entity. A sentence is important if 80% or more of its annotations are 4 or 5.

5.2 Corpus Analysis

In total, 261 sentences were annotated over 10 articles, once by all annotators[9]. This is a difficult task, and in this section we explore some of the difficulties annotators experience.

Table 1 lists statistics about our created corpus. We are most interested in the sentences annotators rate as important, rather than the distribution of low scores. We see that annotators had difficulty reaching consensus, with about two thirds of sentences having entropy greater than one bit. On average once per article, annotators also had difficulty determining if a sentence was about the page subject, with there being a mid-range number of Xs, rather than agreement on whether the sentence merits an X or not (the X category marks sentences not about the page subject). An example of a difficult sentence to annotate as such is The album debuted at #2 on the Swedish albums chart and stayed at this position for a second week. Annotators had difficulty reaching consensus on whether this sentence pertained to the page subject. 13% of sentences were considered important according to our criteria (at least 80% 4s and 5s). We see in Table 2 the distribution of all annotations for our task. The most frequently assigned scores were 2 and 3, suggesting that the most frequent variety of

[8]The current version as of 2014-04-01.

[9]With a small number of exceptions

Criteria on sentence s	Sentences
s is important	35
$H(s) > 1$	176
$H(s) > 2$	23
$33\% < X(s) < 66\%$	15

Table 1: Annotated corpus statistics, where H is entropy, and $X(s)$ is the percentage of Xs assigned to s. Sentences to which 4 or 5 were assigned by at least 80% of annotators are considered important.

Score	Annotations	Sentences
1	356	100
2	541	166
3	536	181
4	416	154
5	278	78
Total	2404	261

Table 2: For each score, the total number of times it was assigned, and the number of sentences which received the score at least once.

sentence annotated was of minor notability. Annotators were reserved in assigning 5s to sentences (with 11% of annotations assigned being 5s), but did not necessarily agree on which sentences to annotate as 5 — less than half of sentences with at least one 5 also had 80% or more 4s and 5s.

6 Results and Analysis

6.1 Peak detection

We set $\theta = 5$ in our experiments in order to detect the maximum number of spikes in our timeseries which were sufficiently many standard deviations above the (weighted) mean. We saw that as θ increased, the number of peaks detected dropped off rapidly. So, our approach is robust to parameter variation, and peaks are easy enough to detect that we can set θ to be large. We also set d to 0.25.

6.2 Important Sentence Extraction

From our result in Section 3 measuring the typical delay between events and edits reflecting them, (1 day) we chose a window size W of 5 (2 days either side of the peak) to account for additional delays which may occur for less notable people.

Dev	P	R	$F_{\beta=1}$
First sentence baseline	27%	80%	40%
Peaks only	7%	20%	10%
Combined	33%	50%	40%

Table 4: Baseline set-based comparison of our preliminary system on our development data

Test	P	R	$F_{\beta=1}$
First sentence baseline	13%	50%	21%
Peaks only	7%	33%	11%
Combined	13%	33%	19%

Table 5: Baseline comparison of our preliminary system on test data)

	Important	Unimportant
Important	1	13
Unimportant	4	100

Table 6: Baseline Confusion Matrix. Rows show the gold standard sentence classifications and columns show our system's classifications

Tables 4 and 5 lists our set-based precision, recall, and f-score for our corpus of 10 articles, (5 development, 5 test) comparing the sentences marked as important by our system and by annotators. By convention, the first sentence of each Wikipedia article tends to be the most informative. For instance, Caroline Lind (born October 11, 1982) is an American rower, and is a two-time Olympic gold medalist.. The information contained in this sentence is often repeated later in the article. Since it is so informative, our annotators ranked this sentence highly for all articles in our corpus. This inspired the development of a simple baseline: A system which returns only the first sentence for every article. This is similar to the typical (hard to beat) baseline for summarisation of news articles — the first 2-3 sentences of the article.

We configured our system to additionally return the first sentence of each article, and saw an increase in f-score from 11% to 19% on our development set, but decrease in overall f-score compared to the first sentence only baseline. Table 6 shows our baseline system's confusion matrix. We see that the majority of errors are in recall — our system does not extract 13 of the 14 gold-standard sentences. We see in Table 7 that returning the first sentence of each article helps with these recall errors.

There are many stages in our pipeline where errors can occur. Annotators can tag a sentence as important which has no spike associated with it, due to lack of timeseries data coverage (before 2008), a lack of spike associated with the sentence, or due to the page traffic increase being too small to be detected as a spike. The most frequent cause of these recall errors is during the edit detection phase, when there are co-incidental edits

Person	Sentence	Score (1 - 5 or X)
Caroline Lind	In her Olympic debut at the 2008 Summer Olympics in Beijing, Lind won a gold medal as a member of the women's eight team.	5
Ed Stoppard	In 2007, he played the title role in the BBC's drama-documentary Tchaikovsky: Fortune and Tragedy.	4
Charlie Webster	This lasted for just a few months and she moved on to present the Red Bull Air Race worldwide for ITV4.	3
Charlie Webster	In April 2009, Webster ran in the London Marathon raising money for the Bobby Moore Fund for Cancer Research UK.	2
Caroline Lind	Lind pursued an M.B. A. with an Accounting Concentration at Rider University, in Lawrenceville, New Jersey.	1
Charlie Day	His father, Dr. Thomas Charles Day, is retired and was a professor of Music History and Music Theory at Salve Regina University in Newport, Rhode Island.	X

Table 3: Sample annotations of sentences from several Wikipedia articles. Each sentence is scored from 1 to 5 or with X, with 5 being a sentence critical to the fame of the page subject, 1 being a sentence which is about the page subject, but does not contain an event or fact, and X being a sentence which is not about the page subject.

	Important	Unimportant
Important	5	9
Unimportant	5	99

Table 7: Confusion Matrix — Baseline system + first sentence of each article always returned. Rows show the gold standard classifications and columns show our system's classifications

to the page in the days leading to a notable event, and when the first edit to a page on the day of a notable event does not mention the event (for instance, vandalism inspired by the notable event).

Errors in the alignment phase can occur because some edits correspond to multiple sentences in the final document. For instance, a revision listing several films in which an actor appeared is later split into several sentences, one for each film. This results in several similar candidate sentences for the edit, our system can choose only one. Furthermore, in a number of cases the iterative updating of the page as a story unfolds causes the text from the original edit to be missing entirely from the final version of the article.

A peak can also be detected for which there is no corresponding page edit, nor a corresponding sentence in the current article. For instance, a spike appears in singer Jimmy Barnes' page view timeseries in 2012, at the same time his daughter first appears on a popular television program.

7 Conclusion

In this paper, we have proposed a novel approach for summarising the notability of a person, and explored the relationship between Wikipedia traffic, page edits, and the occurrence of notable events.

Our experiments have been limited by our small sample of annotated data. A critical next step will be developing a larger sample of annotated data, which will help us understand the task better, and allow exploration into ordering sentences by the amplitude of their associated spike. Two stages that require further exploration are the selection of edits, and their alignment to the current page.

Extracting edits corresponding to page traffic peaks need not limit the source of corresponding sentences to the Wikipedia article text itself. The extracted edits may also bear comparison to other biographies of the same entity, or to sentences in the corpus of news articles about them. It may be interesting to use a similar approach in an entity-centred long-term news retrieval query.

The page traffic data need not be the only source of information indicating interestingness. Other work could use the appearence of the entity in media or in query logs to identify key edits to their Wikipedia page.

We have provided insight into which parts of this task are easy and which are difficult. Our initial exploration into exploiting this timeseries data to detect any of the wide variety of reasons one might be famous has set the stage for further exploration of this new, free, extrinsic source of information about what is interesting.

8 Acknowledgements

This work was supported by ARC Discovery grant DP1097291. The authors thank the anonymous reviewers and the ə-lab researchers for their helpful feedback.

References

Fadi Biadsy, Julia Hirschberg, Elena Filatova, and LLC InforSense. 2008. An Unsupervised Approach to Biography Production Using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Lingusitics: Human Language Technologies*, pages 807–815.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*, pages 172–183. Springer.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization.

Elena Filatova and Eduard Hovy. 2001. Assigning time-stamps to event-clauses. In *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, page 13. Association for Computational Linguistics.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Xiaojiang Liu, Zaiqing Nie, Nenghai Yu, and Ji-Rong Wen. 2010. BioSnowball: automated population of Wikis. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 969–978.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Joel Nothman. 2014. *Grounding event references in news*. Ph.D. thesis, School of Information Technologies, University of Sydney.

Sérgio Nunes, Cristina Ribeiro, and Gabriel David. 2008. Wikichanges: exposing wikipedia revision activity. In *Proceedings of the 4th International Symposium on Wikis*, page 25. ACM.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137.

Barry Schiffman, Inderjeet Mani, and Kristian J Concepcion. 2001. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 458–465. Association for Computational Linguistics.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, pages 697–706. ACM.

Simone Teufel and Marc Moens. 1997. Sentence extraction as a classification task. In *Proceedings of the ACL*, volume 97, pages 58–65.

Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 131–142. ACM.

Scott B Wilson and Ronald Emerson. 2002. Spike detection: a review and comparison of algorithms. *Clinical Neurophysiology*, 113(12):1873–1881.

Fei Wu and Daniel S Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127. Association for Computational Linguistics.

Elif Yamangil and Rani Nelken. 2008. Mining wikipedia revision histories for improving sentence compression. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 137–140. Association for Computational Linguistics.

Liang Zhou, Miruna Ticrea, and Eduard H Hovy. 2004. Multi-Document Biography Summarization. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 434–441.