Eliciting Disease Data from Wikipedia Articles

Geoffrey Fairchild
Sara Y. Del Valle
Los Alamos National Laboratory
Defense Systems & Analysis Division
Los Alamos, New Mexico, USA

Lalindra De Silva
The University of Utah
School of Computing
Salt Lake City, Utah, USA

Alberto M. Segre
The University of Iowa
Department of Computer Science
Iowa City, Iowa, USA

Abstract

Traditional disease surveillance systems suffer from several disadvantages, including reporting lags and antiquated technology, that have caused a movement towards internet-based disease surveillance systems. Internet systems are particularly attractive for disease outbreaks because they can provide data in near real-time and can be verified by individuals around the globe. However, most existing systems have focused on disease monitoring and do not provide a data repository for policy makers or researchers. In order to fill this gap, we analyzed Wikipedia article content.

We demonstrate how a named-entity recognizer can be trained to tag case counts, death counts, and hospitalization counts in the article narrative that achieves an F1 score of 0.753. We also show, using the the 2014 West African Ebola virus disease epidemic article as a case study, that there are detailed time series data that are consistently updated that closely align with ground truth data.

We argue that Wikipedia can be used to create the first community-driven open-source emerging disease detection, monitoring, and repository system.

Introduction

Most traditional disease surveillance systems rely on data from patient visits or lab records (Losos 1996; Burkhead and Maylahn 2000; Adams et al. 2013). These systems, while generally recognized to contain accurate information, rely on a hierarchy of public health systems that causes reporting lags of up to 1–2 weeks in many cases (Burkhead and Maylahn 2000). Additionally, many regions of the world lack the infrastructure necessary for these systems to produce reliable and trustworthy data. Recently, in an effort to overcome these issues, timely global approaches to disease surveillance have been devised using internet-based data. Data sources such as search engine queries (e.g., (Polgreen et al. 2008; Ginsberg et al. 2009)), Twitter (e.g., (Culotta 2010; Aramaki, Maskawa, and Morita 2011; Paul and Dredze 2011; Signorini, Segre, and Polgreen 2011)), and Wikipedia access logs (e.g., (McIver and Brownstein 2014; Generous et al. 2014)) have been shown to be effective in this arena.

A notably different internet-based disease surveillance tool is HealthMap (Freifeld et al. 2008). HealthMap analyzes, in real-time, data from a variety of sources (e.g., ProMED-mail (Madoff 2004), Google News, the World Health Organization) in order to allow simple querying, filtering, and visualization of outbreaks past and present. During emerging outbreaks, HealthMap is often used to understand the current state (e.g., incidence and death counts, outbreak locations). For example, HealthMap was able to detect the 2014 Ebola epidemic nine days before the World Health Organization (WHO) officially announced it (Greenemeier 2014).

While HealthMap has certainly been influential in the digital disease detection sphere, it has some drawbacks. First and foremost, it runs on source code that is not open and relies on certain data sources that are not freely available in their entirety (e.g., Moreover Newsdesk[1]). Some argue that there is a genuine need for open source code and open data in order to validate, replicate, and improve existing systems (Generous et al. 2014). They argue that while certain closed source services, such as HealthMap and Google Flu Trends (Ginsberg et al. 2009), are popular and useful to the public, there is no way for the public to contribute to the service or continue the service, should the owners decide to shut it down. For example, Google offers a companion site to Google Flu Trends, Google Dengue Trends[2]. However, since Google's source code and data are closed, it is not possible for anyone outside of Google to create similar systems for other diseases, e.g., Google Ebola Trends. Additionally, it is not possible for anyone outside of the HealthMap development team to add new features or data sources to HealthMap. For these reasons, Generous et al. argue for the use of Wikipedia access logs coupled with open source code for digital disease surveillance.

Much richer Wikipedia data are available, however, than just access logs. The entire Wikipedia article content and edit histories are available, complete with edit history metadata (e.g., timestamps of edits and IP addresses of anonymous editors). A plethora of open media—audio, images, and video—are also available.

Wikipedia has a history of being edited and used, in

[1] http://www.moreover.com/

[2] http://www.google.org/denguetrends/

many cases, in near real-time during unfolding news events. Keegan et al. have been particularly instrumental in understanding Wikipedia's dynamics during unfolding breaking news events, such as natural disasters and political conflicts and scandals (Keegan, Gergle, and Contractor 2011; 2013; Keegan 2013). They have provided insight into editor networks as well as editing activity during news events. Recognizing that Wikipedia might offer useful disease data during unfolding epidemiological events, this study presents a novel use of Wikipedia article content and edit history in which disease data (i.e., case, death, and hospitalization counts) are elicited in a timely fashion.

We study two different aspects of Wikipedia content as it relates to unfolding disease events:

1. Using standard natural language processing (NLP) techniques, we demonstrate how to capture case counts, death counts, and hospitalization counts from the article text.
2. Using the 2014 West African Ebola virus epidemic article as a case study, we show there are valuable time series data present in the tables found in certain articles.

We argue that Wikipedia data can not only be used for disease surveillance but also as a centralized repository system for collecting disease-related data in near real-time.

Methods

Disease-related information can be found in a number of places on Wikipedia. We demonstrate how two aspects of Wikipedia article content (historical changes to article text and tabular content) can be harvested for disease surveillance purposes. We first show how a named-entity recognizer can be trained to elicit "important" phrases from outbreak articles, and we then study the accuracy of tabular time series data found in certain articles using the 2014 West African Ebola epidemic as a case study.

Wikipedia data

Wikipedia is an open collaborative encyclopedia consisting of approximately 30 million articles across 287 languages (Wikimedia Foundation 2014f; 2014g). The English edition of Wikipedia is by far the largest and most active edition; it alone contains approximately 4.7 million articles, while the next largest Wikipedia edition (Swedish) contains only 1.9 million articles (Wikimedia Foundation 2014g). The textual content of the current revision of each English Wikipedia article totals approximately 10 gigabytes (Wikimedia Foundation 2014d).

One of Wikipedia's primary attractions to researchers is its openness. All of the historical article content, dating back to Wikipedia's inception in 2001, is available to anyone free of charge. Wikipedia content can be acquired through two means: *a*) Wikipedia's official web API[3] or *b*) downloadable database dumps[4]. Although the analysis in this study could have been done offline using the downloadable database dumps, this option is in practice difficult, as the database dumps containing all historical English article revisions are very large (multiple terabytes when uncompressed) (Wikimedia Foundation 2014h). We therefore decided to use Wikipedia's web API, caching content when appropriate.

Wikipedia contains many articles on specific disease outbreaks and epidemics (e.g., the 2014 West Africa Ebola epidemic[5] and the 2012 Middle Eastern Respiratory Syndrome Coronavirus (MERS-CoV) outbreak[6]). We identified two key aspects of Wikipedia disease outbreak articles that can aid disease surveillance efforts: *a*) key phrases in the article text and *b*) tabular content. Most outbreak articles we surveyed contained: dates, locations, case counts, death counts, case fatality rates, demographics, and hospitalization counts in the text. These data are, in general, swiftly updated as new data become available. Perhaps most importantly, sources are often provided so that external review can occur. The following two excerpts came from the articles on the 2012 MERS-CoV outbreak and 2014 Ebola epidemic, respectively:

> On 16 April 2014, Malaysia reported its first MERS-COV related death.[34] The person was a 54 year-old man who had traveled to Jeddah, Saudi Arabia, together with pilgrimage group composed of 18 people, from 15–28 March 2014. He became ill by 4 April, and sought remedy at a clinic in Johor on 7 April. He was hospitalized by 9 April and died on 13 April.[35] (Wikimedia Foundation 2014a)

> On 31 March, the U.S. Centers for Disease Control and Prevention (CDC) sent a five-person team to assist Guinea's Ministry of Health and the WHO to lead an international response to the Ebola outbreak. On that date, the WHO reported 112 suspected and confirmed cases including 70 deaths. Two cases were reported from Liberia of people who had recently traveled to Guinea, and suspected cases in Liberia and Sierra Leone were being investigated.[24] On 30 April, Guinea's Ministry of Health reported 221 suspected and confirmed cases including 146 deaths. The cases included 25 health care workers with 16 deaths. By late May, the outbreak had spread to Conakry, Guinea's capital, a city of about two million inhabitants.[24] On 28 May, the total cases reported had reached 281 with 186 deaths.[24] (Wikimedia Foundation 2014b)

Although most outbreak articles contain content similar to the above examples, not all outbreak articles on Wikipedia contain tabular data. The tabular data that do exist, though, are often consistently updated. For example, Figure 1 presents a screenshot of a table taken from the 2014 Ebola epidemic article. This table contains case counts and death counts for all regions of the world affected by the epidemic, complete with references for the source data. The time granularity is irregular, but updated counts are consistently provided every 2–5 days.

[3] http://www.mediawiki.org/wiki/API:Main_page

[4] http://dumps.wikimedia.org/enwiki/latest/

[5] http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa

[6] http://en.wikipedia.org/wiki/2012_Middle_East_respiratory_syndrome_coronavirus_outbreak

Ebola cases and deaths by country and by date – 1 August to most recent WHO update

Date	Total		Guinea		Liberia		Sierra Leone		Nigeria		Senegal		United States		Spain		Refs
	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths	
5 Oct 2014	8,033	3,865	1,298	768	≥3,924	≥2,210	2,789	879	20	8	1	0	1	0			[note 1][235]
1 Oct 2014	7,492	3,439	1,199	739	≥3,834	≥2,069	2,437	623	20	8	1	0	1	0			[note 2][146]
28 Sep 2014	7,192	3,286	1,157	710	≥3,696	≥1,998	2,317	570	20	8	1	0	1	0			[note 3] [158][240][241]
25 Sep 2014	6,808	3,159	1,103	668	≥3,564	≥1,922	2,120	561	20	8	1	0					[note 4] [242][243][244]
23 Sep 2014	6,574	3,043	1,074	648	≥3,458	≥1,830	2,021	557	20	8	1	0					[note 5][115][136]
21 Sep 2014	6,263	2,900	1,022	635	≥3,280	≥1,707	1,940	550	20	8	1	0					[note 6][42][245]
17 Sep 2014	5,762	2,746	965	623	≥3,022	≥1,578	1,753	537	21	8	1	0					[note 7][246][247][248]
14 Sep 2014	5,339	2,586	942	601	≥2,720	≥1,461	1,655	516	21	8	1	0					[note 8][249][250][251]
10 Sep 2014	4,846	2,375	899	568	2,415	1,307	1,509	493	22	8	3	0					[note 9][252][253]
7 Sep 2014	4,366	2,177	861	557	2,081	1,137	1,424	476	22	7	3	0					[note 10][254][255]
3 Sep 2014	4,001	2,089	823	522	1,863	1,078	1,292	452	22	7	1	0					[256]
31 Aug 2014	3,707	1,808	771	494	1,698	871	1,216	436	21	7	1	0					[note 11] [39][206]
25 Aug 2014	3,071	1,553	648	430	1,378	694	1,026	422	19	7							[257]
20 Aug 2014	2,615	1,427	607	406	1,082	624	910	392	16	5							[6]
18 Aug 2014	2,473	1,350	579	396	972	576	907	374	15	4							[258]
16 Aug 2014	2,240	1,229	543	394	834	466	848	365	15	4							[259]
13 Aug 2014	2,127	1,145	519	380	786	413	810	348	12	4							[260]
11 Aug 2014	1,975	1,069	510	377	670	355	783	334	12	3							[261]
9 Aug 2014	1,848	1,013	506	373	599	323	730	315	13	2							[262]
6 Aug 2014	1,779	961	495	367	554	294	717	298	13	2							[263]
4 Aug 2014	1,711	932	495	363	516	282	691	286	9	1							[264]
1 Aug 2014	1,603	887	485	358	468	255	646	273	4	1							[265]

Figure 1: Table containing updated worldwide Ebola case counts and death counts. This is a screenshot taken directly from the 2014 Ebola epidemic Wikipedia article (Wikimedia Foundation 2014b). Time granularity is irregular but is in general every 2–5 days. References are also provided for all data points.

While there are certainly other aspects of Wikipedia article content that can be leveraged for disease surveillance purposes, these are the two we focus on in this study. The following sections detail the data extraction methods we use.

Named-entity recognition

In order to recognize certain key phrases in the Wikipedia article narrative, we trained a *named-entity recognizer* (NER). Named-entity recognition is a task commonly used in natural language processing (NLP) to identify and categorize certain key phrases in text (e.g., names, locations, dates, organizations). NERs are *sequence labelers*; that is, they label sequences of words. Consider the following example (Wikimedia Foundation 2014e):

Jim bought 300 shares of Acme Corp. in 2006.

Entities in this example could be named as follows:

[Jim]$_{\texttt{PERSON}}$ bought 300 shares of [Acme Corp.]$_{\texttt{ORGANIZATION}}$ in [2006]$_{\texttt{TIME}}$.

This study specifically uses Stanford's NER (Finkel, Grenager, and Manning 2005)[7]. The Stanford NER is an implementation of a conditional random field (CRF) model (Sutton 2011). CRFs are probabilistic statistical models that are the discriminative analog of hidden Markov models (HMMs). Generative models, such as HMMs, learn the joint probability $p(x, y)$, while discriminative models, such as CRFs, learn the conditional probability $p(y \mid x)$. In practice, this means that generative models like HMMs classify by modeling the actual distribution of each class, while discriminative models like CRFs classify by modeling the boundaries between classes. In most cases, discriminative models outperform generative models (Ng and Jordan 2002).

[7] http://nlp.stanford.edu/software/CRF-NER.shtml

While Stanford's NER includes models capable of recognizing common named entities, such as `PERSON`, `ORGANIZATION`, and `LOCATION`, it also provides the capability for us to train our own model so that we can capture new types of named entities we are interested in. For this specific task, we were interested in automatically identifying three entity types: *a*) `DEATHS` *b*) `INFECTIONS`, and *c*) `HOSPITALIZATIONS`. Our trained model should therefore be able to automatically tag phrases that correspond to these three entities in the text documents it receives as input.

NERs possess the ability to learn and generalize in order to identify unseen phrase patterns. Since the classifier is dependent on the features we provide to it (e.g., words, part of speech tags), it should hopefully generalize well for the unseen instances. A more simplistic pattern-matching approach, such as regular expressions, is not practical due to inherent variation. For example, the following phrases from our dataset all contain `INFECTIONS` entities:

1. . . . a total of 17 patients with confirmed H7N9 virus infection . . .
2. . . . there were only sixty-five cases and four deaths . . .
3. . . . more than 16,000 cases were being treated . . .

Example 1 has the pattern `[number] patients`, while examples 2 and 3 follow the pattern `[number] cases`. However, example 2 spells out the number, while example 3 provides the numeral. A simple regular expression cannot capture the variability found in our dataset; we would need to define dozens of regular expressions for each entity type, and rigidity of regular expressions would limit the likelihood that we would be able to identify entities in new unseen patterns.

A number of steps were required to prepare the data for annotation so that the NER could be trained:

1. We first queried Wikipedia's API in order to get the complete revision history for the articles used in our training set.
2. We cleaned each revision by stripping all MediaWiki markup from the text, as well as removing tables.
3. We computed the diff (i.e., textual changes) between successive pairs of articles. This provided lines deleted and added between the two article revisions. We retained a list of all the line additions across all article revisions.
4. Many lines in this resulting list were similar to one another (e.g., "There are 45 new cases." → "There are 56 new cases."). For the purposes of training the NER, it is not necessary to retain highly similar or identical lines. We therefore split each line into sentences and removed similar sentences by computing the Jaccard similarity between each sentence using trigrams as the constituent parts in the Jaccard equation. The Jaccard similarity equation for measuring the similarity between two sets A and B, defined as $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, is commonly used for near-duplicate detection (Manning, Raghavan, and Schütze 2009). We only kept sentences for which the similarity with all the distinct sentences retained so far was no greater than 0.75.

5. We split each line into tokens in order to create a tab-separated value file that is compatible with Stanford's NER.

6. Finally, we used Stanford's part-of-speech (POS) tagger (Toutanova et al. 2003)[8] to add a POS feature to each token.

In order to train the NER, we annotated a dataset derived from the following 14 Wikipedia articles generated according to the above methodology: *a*) Ebola virus epidemic in West Africa[9], *b*) Haiti cholera outbreak[10], *c*) 2012 Middle East respiratory syndrome coronavirus outbreak[11], *d*) New England Compounding Center meningitis outbreak[12], *e*) Influenza A virus subtype H7N9[13], *f*) 2013–14 chikungunya outbreak[14], *g*) Chikungunya outbreaks[15], *h*) Dengue fever outbreaks[16], *i*) 2013 dengue outbreak in Singapore[17], *j*) 2011 dengue outbreak in Pakistan[18], *k*) 2009–10 West African meningitis outbreak[19], *l*) Mumps outbreaks in the 21st century[20], *m*) Zimbabwean cholera outbreak[21], and *n*) 2006 dengue outbreak in India[22]. The entire cleaned and annotated dataset contained approximately 55,000 tokens. The inside-outside-beginning (IOB) scheme, popularized in part by the CoNLL-2003 shared task on language-independent named-entity recognition (Tjong Kim Sang and De Meulder 2003), was used to tag each token. The IOB scheme offers the ability to tie together sequences of tokens that make up an entity.

The annotation task was split between two annotators (the first and second authors). In order to tune inter-annotator agreement, the annotators each annotated three sets of 5,000 tokens. After each set of annotations, differences were identified, and clarifications to the annotation rules were made. The third set resulted in a Cohen's kappa coefficient of 0.937, indicating high agreement between the annotators.

[8]http://nlp.stanford.edu/software/tagger.shtml

[9]http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa

[10]http://en.wikipedia.org/wiki/Haiti_cholera_outbreak

[11]http://en.wikipedia.org/wiki/2012_Middle_East_respiratory_syndrome_coronavirus_outbreak

[12]http://en.wikipedia.org/wiki/New_England_Compounding_Center_meningitis_outbreak

[13]http://en.wikipedia.org/wiki/Influenza_A_virus_subtype_H7N9

[14]http://en.wikipedia.org/wiki/2013%E2%80%9314_chikungunya_outbreak

[15]http://en.wikipedia.org/wiki/Chikungunya_outbreaks

[16]http://en.wikipedia.org/wiki/Dengue_fever_outbreaks

[17]http://en.wikipedia.org/wiki/2013_dengue_outbreak_in_Singapore

[18]http://en.wikipedia.org/wiki/2011_dengue_outbreak_in_Pakistan

[19]http://en.wikipedia.org/wiki/2009%E2%80%9310_West_African_meningitis_outbreak

[20]http://en.wikipedia.org/wiki/Mumps_outbreaks_in_the_21st_century

[21]http://en.wikipedia.org/wiki/Zimbabwean_cholera_outbreak

[22]http://en.wikipedia.org/wiki/2006_dengue_outbreak_in_India

Tabular data

To understand the viability of tabular data in Wikipedia, we concentrate on the Ebola virus epidemic in West Africa article[23]. We chose this article for two reasons. First, the epidemic is still unfolding, which makes it a concern for epidemiologists worldwide. Second, the epidemiological community has consistently updated the article as new developments are publicized. Ideally, we would analyze *all* disease articles that contain tabular data, but the technical challenges surrounding parsing the constantly changing data leave this as future work.

Ebola is a rare but deadly virus that first appeared in 1976 simultaneously in two different remote villages in Africa. Outbreaks of Ebola virus disease (EVD), previously known as Ebola hemorrhagic fever (EHF), are sporadic and generally short-lived. The average case fatality rate is 50%, but it has varied between 25% and 90% in previous outbreaks. EVD is transmitted to humans from animals (most commonly, bats, apes, and monkeys) and also from other humans through direct contact with blood and body fluids. Signs and symptoms appear within 2–21 days of exposure (average 8–10 days) and include fever, severe headache, muscle pain, weakness, diarrhea, vomiting, abdominal pain, and unexplained bleeding or bruising. Although there is currently no known cure, treatment in the form of aggressive rehydration seems to improve survival rates (World Health Organization 2014a; Centers for Disease Control and Prevention 2014).

The West African EVD epidemic was officially announced by the WHO on March 25, 2014 (World Health Organization 2014b). The disease spread rapidly and has proven difficult to contain in several regions of Africa. At the time of this writing, it has spread to 7 different countries (including two outside of Africa): Guinea, Liberia, Sierra Leone, Nigeria, Senegal, United States, and Spain.

The Wikipedia article was created on March 29, 2014, four days after the WHO announced the epidemic (Wikimedia Foundation 2014c). As seen in Figure 1, this article contains detailed tables of case counts and death counts by country. The article is regularly updated by the Wikipedia community (see Figure 2); over the 165-day period analyzed, the article averaged approximately 31 revisions per day.

We parsed the Ebola article's tables in several steps:

1. We first queried Wikipedia's API to get the complete revision history for the West African EVD epidemic article. Our initial dataset contained 5,137 revisions from March 29, 2014 to October 14, 2014.

2. We then parsed each revision to pull out case count and death count time series for each revision. To parse the tables, we first used pandoc[24] to convert the MediaWiki markup to consistently formatted HTML and then used BeautifulSoup[25] to parse the HTML. Because the Wikipedia time series contain a number of missing data points

[23]http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa

[24]http://johnmacfarlane.net/pandoc/

[25]http://www.crummy.com/software/BeautifulSoup/

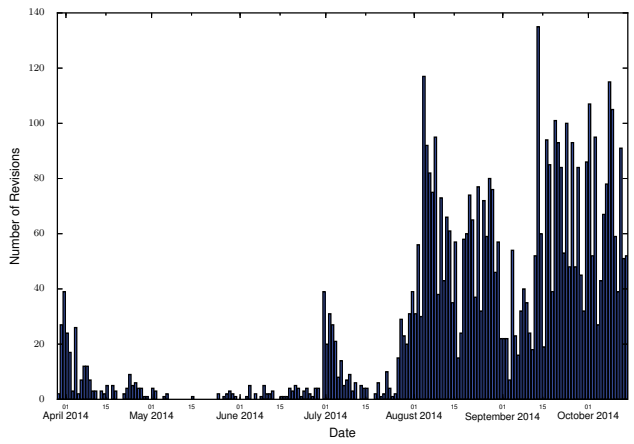


Figure 2: The number of revisions made each day to the 2014 Ebola virus epidemic in West Africa Wikipedia article (http://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa). A total of 5,137 revisions were made over the 165-day period analyzed.

prior to June 30, 2014, we use this date for the beginning of our analysis; time series data prior to June 30, 2014 are not used in this study. This resulting dataset contained 3,803 time series.

3. As Figure 1 shows, there are non-regular gaps in the Wikipedia time series; these gaps range from 2–5 days. We used linear interpolation to fill in missing data points where necessary so that we have daily time series. Daily time series data simplify comparisons with ground truth data (described later).

4. Recognizing that the tables will not necessarily change between article revisions (i.e., an article revision might contain edits to only the text of the article, not to a table in the article), we then removed identical time series. This final dataset contained 39 time series.

Results

Named-entity recognition

To test the classifier's performance, we averaged precision, recall, and F1 score results from 10-fold cross-validation. Table 1 demonstrates a typical confusion matrix used to bin cross-validation results, which are then used to compute precision, recall, and the F1 score. Precision asks, "Out of all the examples the classifier labeled, what fraction were correct?" and is computed as $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$. Recall asks, "Out of all labeled examples, what fraction did the classifier recognize?" and is computed as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$. The F1 score is the harmonic mean of precision and recall: $2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision+recall}}$. All three scores range from 0 to 1, where 0 is the worst score possible and 1 is the best score possible.

Table 2 shows these results as we varied the `maxNGramLeng` option (Stanford's default value is 6). The `maxNGramLeng` option determines sequence length when training. We were somewhat surprised to discover that larger `maxNGramLeng` values did not improve the performance of the classifier, indicating that more training data are likely necessary to further improve the classifier. Furthermore, roughly maximal performance is achieved with `maxNGramLeng` $= 4$; there is no tangible benefit to larger sequences (despite this, we concentrate on the `maxNGramLeng` $= 6$ case since it is the default). Our 14-article training set achieved precision of 0.812 and recall of 0.710, giving us an F1 score of 0.753 for `maxNGramLeng` $= 6$.

For `maxNGramLeng` $= 6$, Table 3 shows the average precision, recall, and F1 scores for each of the named entities we annotated (`DEATHS`, `INFECTIONS`, and `HOSPITALIZATIONS`). There were a total of 264 `DEATHS`, 633 `INFECTIONS`, and 16 `HOSPITALIZATIONS` entities annotated across the entire training dataset. Recall that we used the IOB scheme for annotating sequences; this is reflected in Table 3, with `B-`$\star$ indicating the beginning of a sequence and `I-`$\star$ indicating the inside of a sequence. It is generally the case that identifying the beginning of a sequence is easier than identifying all of the inside words of a sequence; the only exception to this is `HOSPITALIZATIONS`, but we speculate that the identical beginning and inside results for this entity are due to the relatively small sample size.

Table 1: Typical classifier confusion matrix.

	Ground truth positive	**Ground truth negative**
Test positive	True positive (TP)	False positive (FP)
Test negative	False negative (FN)	True negative (TN)

Table 2: Classifier performance determined from 10-fold cross-validation.

`maxNGramLeng`	**Precision**	**Recall**	**F1 score**
1	0.820	0.693	0.747
2	0.810	0.690	0.740
3	0.815	0.702	0.750
4	0.814	0.709	0.753
5	0.813	0.709	0.753
6	0.812	0.710	0.753
7	0.812	0.706	0.751
8	0.814	0.708	0.753
9	0.815	0.707	0.753
10	0.815	0.708	0.753
11	0.813	0.708	0.753
12	0.811	0.709	0.752

Tabular data

To compute the accuracy of the Wikipedia West African EVD epidemic time series, we used Caitlin Rivers' crowdsourced Ebola data[26]. Her country-level data come from official WHO data and reports. As with the Wikipedia time series, we used linear interpolation to fill in missing data where necessary so that the ground truth data are specified

[26]https://github.com/cmrivers/ebola

Table 3: Classifier performance for each of the entities we used in our annotations.

Named entity	Precision	Recall	F1 score
B-Deaths	0.888	0.744	0.802
I-Deaths	0.821	0.730	0.764
B-Infections	0.812	0.719	0.756
I-Infections	0.762	0.714	0.730
B-Hospitalizations	0.933	0.833	0.853
I-Hospitalizations	0.933	0.833	0.853

daily; this ensured that the Wikipedia and ground truth time series were specified at the same granularity. Note that time granularity of the WHO-based ground truth dataset is generally finer than the Wikipedia data; the gaps in the ground truth time series were not the same as those in the Wikipedia time series. In many cases, the ground truth data were updated every 1–2 days.

We compared the 39 Wikipedia epidemic time series to the ground truth data by computing the root-mean-square error (RMSE). We use the RMSE rather than the mean-square error (MSE) because the testing and ground truth time series both have the same units (cases or deaths); when they have the same units, the computed RMSE also has the same unit, which makes it easily interpretable. The RMSE,

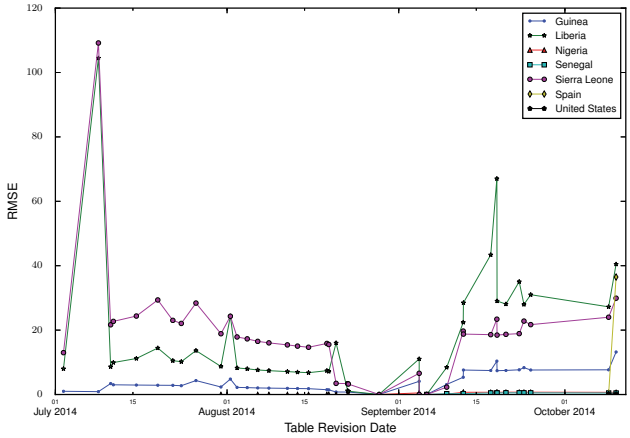
$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}, \qquad (1)$$

computes the average number of cases or deaths difference between a Wikipedia epidemic time series ($\hat{Y}$) and the ground truth time series (Y). Figure 3 shows how the case time series and death time series RMSE changes with each table revision for each country. Of particular interest is the large spike in Figure 3a on July 8, 2014 in Liberia and Sierra Leone. Shortly after the 6:27pm spike, an edit from a different user at 8:16pm the same day with edit summary "correct numbers in wrong country columns" corrected the error.

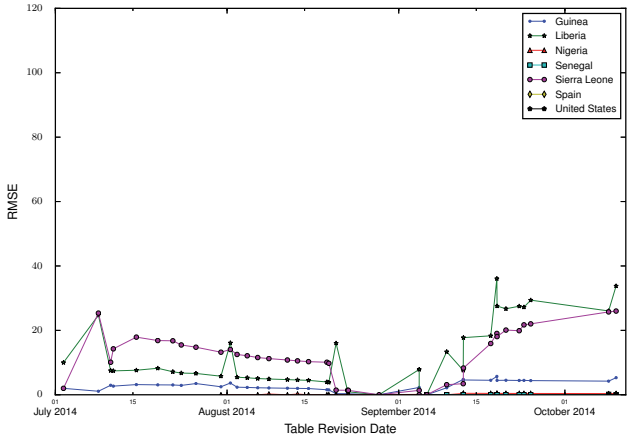
The average RMSE values for each country's time series are listed in Table 4. Even in the worst case, the average deviation between the Wikipedia time series and the ground truth is approximately 19 cases and 12 deaths. Considering the magnitude of the number of cases (e.g., approximately 1,500 in Liberia and 3,500 in Sierra Leone during the time period considered) and deaths (e.g., approximately 850 in Liberia and 1,200 in Sierra Leone), the Wikipedia time series are generally within 1–2% of the ground truth data.

Conclusions

Internet data are becoming increasingly important for disease surveillance because they address some of the existing challenges, such as the reporting lags inherent in traditional disease surveillance data, and they can also be used to detect and monitor emerging diseases. Additionally, internet data can simplify global disease data collection. Collecting disease data is a formidable task that often requires browsing websites written in an unfamiliar language, and data are specified in a number of formats ranging from



(a) Cases



(b) Deaths

Figure 3: Root-mean-square error (RMSE) values for the cases and deaths time series are shown for each revision where the tables changed. The RMSE spikes on July 8, 2014 (Liberia and Sierra Leone) and August 20, 2014 (Liberia) in 3a were due to Wikipedia contributor errors and were fixed shortly after they were made. Most RMSE spikes are quickly followed by a decrease; this is due to updated WHO data or contributor error detection.

Table 4: Average cases and deaths RMSE across all table revisions.

Country	Mean Cases RMSE	Mean Deaths RMSE
Guinea	3.790	2.701
Liberia	18.168	11.983
Nigeria	0.310	0.189
Senegal	0.403	0.008
Sierra Leone	18.847	12.015
Spain	18.243	0.050
United States	0.174	0.000

well-formed spreadsheets to unparseable PDF files containing low resolution images of tables. Although several popular internet-based systems exist to help overcome some of these traditional disease surveillance system weaknesses, most notably HealthMap (Freifeld et al. 2008) and Google Flu Trends (Ginsberg et al. 2009), no such system exists that relies solely on open data and runs using 100% open source code.

Previous work explored Wikipedia access logs to tackle some of the disadvantages traditional disease surveillance systems face (McIver and Brownstein 2014; Generous et al. 2014). This study explores a new facet of Wikipedia: the content of disease-related articles. We present methods on how to elicit data that can potentially be used for near-real-time disease surveillance purposes. We argue that in some instances, Wikipedia may be viewed as a centralized crowd-sourced data repository.

First, we demonstrate using a named-entity recognizer (NER) how case counts, death counts, and hospitalization counts can be tagged in the article narrative. Our NER, trained on a dataset derived from 14 Wikipedia articles on disease outbreaks/epidemics, achieved an F1 score of 0.753, evidence that this method is fully capable of recognizing these entities in text. Second, we analyzed the quality of tabular data available in the 2014 West Africa Ebola virus disease article. By computing the root-mean-square error (RMSE), we show that the Wikipedia time series very closely align with WHO-based ground truth data.

There are many future directions for this work. First and foremost, more training data are necessary for an operational system in order to improve precision and recall. There are many more disease- and outbreak-related Wikipedia articles that can be annotated. Additionally, other open data sources, such as ProMED-mail, might be used to enhance the model. Second, a thorough analysis of the quality and correctness of the entities tagged by the NER is needed. This study presents the methods by which disease-related named entities can be recognized, but we have not throughly studied the correctness and timeliness of the data. Third, our analysis of tabular data consisted of a single article. A more rigorous study looking at the quality of tabular data in more articles is necessary. Finally, the work presented here considers only the English Wikipedia. NERs are capable of tagging entities in a variety of other languages; more work is needed to understand the quality of data available in the 286 non-English Wikipedias.

There are several limitations to this work. First, the ground truth time series we used to compute RMSEs is static, while the Wikipedia time series vary over time. Because the relatively recent static ground truth time series may contain corrections for reporting errors made earlier in the epidemic, the RMSE values may be artificially inflated in some instances. Second, we are ignoring the user-provided edit summary. This edit summary provides information about why the edit was made. The edit summary identifies article vandalism (and subsequent vandalism reversion) as well as content corrections and updates. Taking these edit summaries into account can further improve model performance (e.g., processing edit summaries would allow us to disregard the erroneous edit that caused the July 8, 2014 spike in Figure 3a).

Ultimately, we envision this work being incorporated into a community-driven open-source emerging disease detection and monitoring system. Wikipedia access log time series gauge public interest and, in many cases, correlate very well with disease incidence. A community-driven effort to improve global disease surveillance data is imminent, and Wikipedia can play a crucial role in realizing this need.

Acknowledgments

This work is supported in part by NIH/NIGMS/MIDAS under grant U01-GM097658-01 and the DTRA Joint Science and Technology Office for Chemical and Biological Defense under project numbers CB3656 and CB10007. LANL is operated by Los Alamos National Security, LLC for the Department of Energy under contract DE-AC52-06NA25396.

References

Adams, D. A.; Gallagher, K. M.; Jajosky, R. A.; Kriseman, J.; Sharp, P.; Anderson, W. J.; Aranas, A. E.; Mayes, M.; Wodajo, M. S.; Onweh, D. H.; and Abellera, J. P. 2013. Summary of Notifiable Diseases, United States, 2011. Technical Report 53, Centers for Disease Control and Prevention, Atlanta, Georgia.

Aramaki, E.; Maskawa, S.; and Morita, M. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proceedings of the 2011 Conference on Emperical Methods in Natural Language Processing*, 1568–1576. Edinburgh, United Kingdom: Association for Computational Linguistics.

Burkhead, G. S., and Maylahn, C. M. 2000. State and Local Public Health Surveillance. In Teutsch, S. M., and Churchill, R. E., eds., *Principles and Practice of Public Health Surveillance*. New York: Oxford University Press, 2nd edition. chapter 12, 253–286.

Centers for Disease Control and Prevention. 2014. Ebola (Ebola Virus Disease). http://www.cdc.gov/vhf/ebola/. Accessed: 2014-10-27.

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 115–122. Washington, DC: ACM Press.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, number June, 363–370. Morristown, NJ, USA: Association for Computational Linguistics.

Freifeld, C. C.; Mandl, K. D.; Reis, B. Y.; and Brownstein, J. S. 2008. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association* 15(2):150–157.

Generous, N.; Fairchild, G.; Deshpande, A.; Del Valle, S. Y.; and Priedhorsky, R. 2014. Global Disease Monitoring and

Forecasting with Wikipedia. *PLOS Computational Biology* 10(11):e1003892.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Greenemeier, L. 2014. Smart Machines Join Humans in Tracking Africa Ebola Outbreak. *Scientific American*.

Keegan, B.; Gergle, D.; and Contractor, N. 2011. Hot off the wiki: dynamics, practices, and structures in Wikipedia's coverage of the Thoku catastrophes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 105–113. Mountain View, California: ACM.

Keegan, B.; Gergle, D.; and Contractor, N. 2013. Hot Off the Wiki: Structures and Dynamics of Wikipedia's Coverage of Breaking News Events. *American Behavioral Scientist* 57(5):595–622.

Keegan, B. C. 2013. A history of newswork on Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, 7:1–7:10. Hong Kong, China: ACM.

Losos, J. Z. 1996. Routine and sentinel surveillance methods. *Eastern Mediterranean Health Journal* 2(1):46–50.

Madoff, L. C. 2004. ProMED-mail: An Early Warning System for Emerging Diseases. *Clinical Infectious Diseases* 39(2):227–232.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2009. *Introduction to Information Retrieval*. Number c. Cambridge, England: Cambridge University Press.

McIver, D. J., and Brownstein, J. S. 2014. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLOS Computational Biology* 10(4):e1003581.

Ng, A., and Jordan, M. 2002. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In Dietterich, T. G.; Becker, S.; and Ghahramani, Z., eds., *Proceedings of the 2001 Neural Information Processing Systems Conference*, 841–848. British Columbia, Canada: MIT Press.

Paul, M. J., and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 265–272.

Polgreen, P. M.; Chen, Y.; Pennock, D. M.; and Nelson, F. D. 2008. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases* 47(11):1443–1448.

Signorini, A.; Segre, A. M.; and Polgreen, P. M. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS ONE* 6(5):e19467.

Sutton, C. 2011. An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning* 4(4):267–373.

Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, volume 4, 142–147. Morristown, NJ, USA: Association for Computational Linguistics.

Toutanova, K.; Klein, D.; Manning, C. D.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 173–180. Morristown, NJ, USA: Association for Computational Linguistics.

Wikimedia Foundation. 2014a. 2012 Middle East respiratory syndrome coronavirus outbreak. http://en.wikipedia.org/w/index.php?title=2012_Middle_East_respiratory_syndrome_coronavirus_outbreak&oldid=628796140. Accessed: 2014-10-10.

Wikimedia Foundation. 2014b. Ebola virus epidemic in West Africa. http://en.wikipedia.org/w/index.php?title=Ebola_virus_epidemic_in_West_Africa&oldid=629094432. Accessed: 2014-10-10.

Wikimedia Foundation. 2014c. Ebola virus epidemic in West Africa. https://en.wikipedia.org/w/index.php?title=Ebola_virus_epidemic_in_West_Africa&oldid=601868739. Accessed: 2014-03-24.

Wikimedia Foundation. 2014d. English Wikipedia. http://en.wikipedia.org/w/index.php?title=English_Wikipedia&oldid=627512912. Accessed: 2014-10-07.

Wikimedia Foundation. 2014e. Named-entity recognition. http://en.wikipedia.org/w/index.php?title=Named-entity_recognition&oldid=627138157. Accessed: 2014-10-11.

Wikimedia Foundation. 2014f. Wikipedia. https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=636552708. Accessed: 2014-12-04.

Wikimedia Foundation. 2014g. Wikipedia Statistics. http://stats.wikimedia.org/EN/Sitemap.htm. Accessed: 2014-10-07.

Wikimedia Foundation. 2014h. Wikipedia:Database download. http://en.wikipedia.org/w/index.php?title=Wikipedia:Database_download&oldid=627253774. Accessed: 2014-10-08.

World Health Organization. 2014a. Ebola virus disease. http://www.who.int/mediacentre/factsheets/fs103/en/. Accessed: 2014-10-27.

World Health Organization. 2014b. Ebola virus disease in Guinea (Situation as of 25 March 2014). http://www.afro.who.int/en/clusters-a-programmes/dpc/epidemic-a-pandemic-alert-and-response/outbreak-news/4065-ebola-virus-disease-in-guinea-25-march-2014.html. Accessed: 2014-12-01.