Syndromic Surveillance through Measuring Lexical Shift in Emergency Department Chief Complaint Texts

Hafsah Aamer, Bahadorreza Ofoghi, Karin Verspoor
Department of Computing and Information Systems
The University of Melbourne
Parkville Victoria 3010, Australia
haamer@student.unimelb.edu.au
bahadorreza.ofoghi@unimelb.edu.au
karin.verspoor@dunimelb.edu.au

Abstract

Syndromic Surveillance has been performed using machine learning and other statistical methods to detect disease outbreaks. These methods are largely dependent on the availability of historical data to train the machine learning-based surveillance system. However, relevant training data may differ from region to region due to geographical and seasonal trends, meaning that the syndromic surveillance designed for one area may not be effective for another. We proposed and analyse a semi-supervised method for syndromic surveillance from emergency department chief complaint textual notes that avoids the need for large training data. Our new method is based on identification of lexical shifts in the language of Chief Complaints of patients, as recorded by triage nurses, that we believe can be used to monitor disease distributions and possible outbreaks over time. The results we obtained demonstrate that effective lexical syndromic surveillance can be approached when distinctive lexical items are available to describe specific syndromes.

1 Introduction

The increase in new emerging pathogenic diseases like SAARS, Ebola, and the Zika virus requires an ongoing effective syndromic surveillance system. A syndromic surveillance system keeps track of the frequency of patients experiencing specific syndromes over time. Any abnormality in the normal trend of syndromes with respect to time may imply a proximal disease outbreak.

There are many data sources that can be used to perform syndromic surveillance, such as data from hospital emergency departments. Chief complaints provide us with one such rich data source to perform syndromic surveillance. A chief complaint is the set of signs and symptoms that the triage nurse registers upon a patient's arrival at the emergency department of a hospital. The symptoms are usually described in a terse, ungrammatical, and abbreviated language. Once the chief complaint is registered at the initial assessment, the patient sees the doctor and receives a detailed check up. This may include sending the patient's specimen for testing in the laboratory especially when dealing with a potentially dangerous infectious disease. The results from the laboratory tests may take days before the end result is available and by then, there is a large risk of spreading the disease to other people who may come in contact with the infected person.

Chief complaints are readily available in a digital format and can therefore be easily processed using natural language processing algorithms. In the past, various work has been done to perform syndromic surveillance using supervised machine learning and statistical algorithms (Tsui et al., 2003; Espino et al., 2007; Chapman et al., 2005; Bradley et al., 2005). The major drawback of machine learning methods is the requirement of historical data that can be used to train the system, and a sensitivity to the characteristics of specific text types (Baldwin et al., 2013). In the case of syndromic surveillance, there is evidence of a need for localized training data; as Ofoghi and Verspoor (2015) found, a machine learning classifier trained on an American data set may not be effective on an Australian data set. The authors found that the American off-the-shelf syndromic classifier (CoCo) achieved a lower F-score on the Australian data set compared with another classifier (SyCo) that was trained with the Australian data set. Moreover, there may be a lack of resources to collect ongoing data for chief complaints, especially in remote areas. Therefore, there is a need for a surveillance system that can work without the

Hafsah Aamer, Bahadorreza Ofoghi and Karin Verspoor. 2016. Syndromic Surveillance through Measuring Lexical Shift in Emergency Department Chief Complaint Texts. In *Proceedings of Australasian Language Technology Association Workshop*, pages 46−54.

availability of historical data for a long period of time.

In this work, rather than using machine learning algorithms to classify the new chief complaint of a patient into pre-defined syndromic groups, we explored the lexical content of chief complaints over time, specifically the changes in the distribution of terms, that may indicate an impending event. We hypothesise that when the probability distribution of terms in chief complaints of consecutive time-frames exhibits a large divergence, then there has been a measurable change in the trend of syndromes, which can be used for detecting an outbreak. The Jensen-Shannon Divergence (JSD) (Lin, 1991), also known as information radius, between consecutive time-frames over a chief complaint corpus was used in combination with CUSUM (Cumulative Sum) algorithms (Fricker Jr et al., 2008) to detect any aberrancy in the data. We also experiment with the text segmentation algorithm Link Set Median (LSM) (Hoey, 1991) to find segmentations in the chief complaints texts, as a proxy for lexical shift.

The remainder of this paper is organized as follows. We first describe the *SynSurv* data set used for our experiments. Then, the three different types of time-frames necessary to perform our lexical analyses will be introduced. This is followed by the discussion of how we modeled chief complaints for textual analysis using statistical methods. Finally, we discuss the utilization of abnormality detection algorithms and the results obtained in our experiments.

2 The SynSurv Data Set

The Syndromic Surveillance (SynSurv) data we used for our analyses was collected from two of the main hospitals in Melbourne, Australia; the Royal Melbourne Hospital and the Alfred Hospital. The data was collected on behalf of the Victorian Department of Health, initially to enable monitoring during the 2006 Commonwealth Games held in Melbourne. The data contains 314,630 chief complaints labeled with a syndromic group as well as with disease codes in the ICD-10 and SNOMED terminologies. The original SynSurv data set contained data with respect to eight syndromes *Flu-Like Illness*, *Diarrhea*, *Septic Shock*, *Acute Flaccid Paralysis*, *Acute Respiratory*, *Radiation Injury*, *Fever with CNS*, and *Rash with Fever*. Due to the sparsity of data for many of the syndromes in the data, we chose the top three with the highest numbers of positive cases in the SynSurv data, i.e., *Flu-Like Illness*, *Diarrhea*, and *Acute Respiratory*.

Chief complaints differ from the majority of other free text that can be found in textual documents or social media posts in that they contain medical acronyms, abbreviations, as well as numeric values representing body temperature, blood pressure, and the like. They are notes entered by the triage nurses lacking well-defined linguistic structure. Therefore, some preprocessing was carried out on the set of chief complaints.

We approached the chief complaint preprocessing task first by lowercasing, lemmatization using Stanford Lemmatizer, and removing stop words and the already assigned ICD-10 and SNOMED disease codes from the end of the chief complaint strings. These are the disease codes assigned to each chief complaint upon a patient's discharge. We removed these codes to retain the chief complaint texts in their original format. We also removed all the non-alphanumeric symbols except "/", which plays a meaningful role in the medical domains. For instance, blood pressure is recorded as two numbers, as 120/80, the first number representing systolic blood pressure and the second number representing diastolic blood pressure. If we remove "/", this numeric reading becomes "12080" or "120 80" which results in a loss of information; neither choice may define a good feature to represent a chief complaint with a specific syndrome. In addition, the same symbol "/" is used for shorthand notations while making notes, for example, the chief complaint texts heavily contained "o/a" meaning "overall pale". Other notations included: "r/a", "p/s", "c/o", and "r/t". We also observed that the use of such notation depended on the nurse's own preference; while there were many chief complaints with "o/a" there were also many occurrences of "o a" without the symbol "/" in between the characters. This meant we could not target all the notations consistently.

However, we did not remove any numeric values from the chief complaints because numeric values may be relevant for diagnosing a syndrome. For example, body temperature is one of the main recordings that a doctor (or a nurse) takes when a patient visits the emergency department to observe if there is a serious illness.

3 The Choice of Time-Frame

The number of patients that visit emergency departments varies between weekdays and weekends and also seasonally. To cater for such day-of-the-week and seasonal trends, we used different time-frames to accumulate chief complaints, some of which normalize the frequencies of chief complaints over longer (seven-day) periods. We experimented with the following three time-frames:

Intersecting seven-day windows

The chief complaints over seven days were accumulated as one time-frame window; then, this time-frame window was advanced by one day. The seven-day time-frame windows cater for the varying frequency of patients visiting over weekdays and weekends. The one-day shifting time-frame is popular for supervised syndromic surveillance as it allows for real-time syndromic surveillance at the end of each day (Hutwagner et al., 2003). Note that for lexical shift analysis with seven-day windows shifting by one day, there is a large vocabulary overlap for the six intersecting days; however, the variance would be significant enough to be captured by aberrancy detection algorithms that will be discussed later.

Disjoint seven-day windows

The chief complaints over seven days were accumulated as one time-frame corpus; then, the time-frame was advanced by seven days so the two consecutive windows were disjoint.

Disjoint one-day windows

In this case, all of the chief complaints in a day formed one time-frame corpus of chief complaints. The time shift was for one day and therefore, the chief complaints in the next day formed the next time-frame corpus. Although this time-frame does not normalize day-of-the-week variances, we incorporated this type in order to find some daily patterns in the SynSurv data set.

4 Textual Modeling of Chief Complaints

We examine the distribution of lexical items in chief complaints to find the differences in the terminology used in consecutive time-frames. For this, statistical methods were utilized, as will be discussed in the following sections. Such statistical methods have been previously used for corpus analysis and comparison (Verspoor et al., 2009; Rayson and Garside, 2000). We follow that prior work here, considering chief complaints in the consecutive time-frames as separate corpora to be compared with each other.

4.1 Jensen-Shannon Divergence

The Jensen-Shannon divergence, also known as Information Radius, is a symmetric measure that measures the similarity between two probability distributions P and Q over the same event space. The JSD between the two probability distributions P and Q is a symmetrized and smoothed version of the Kullback-Leibler Divergence (KLD) and is calculated using Equation 1.

$$JSD(P\|Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M) \quad (1)$$

where, $M = \frac{1}{2}(P+Q)$ and D represents the KLD distance between the two probability distributions; on a finite set χ is calculated using Equation 2.

$$D(P\|Q) = \sum_{x\in\chi} P(x)\log_n \frac{P(x)}{Q(x)} \quad (2)$$

When modeling the chief complaint corpora, the union set of the vocabularies of two corpora to be compared was constructed ($V = V_{c_1} \cup V_{c_2}$), where (c_1) and (c_2) were the chief complaint corpora belonging to the two consecutive time-frames. P and Q represent the probability of corpus terms. Since JSD inherently performs probability smoothing, to find the probability distribution of each term t_k over V_{c_i} for each corpus, the conditional probability of term t_k in each corpus was calculated using Equation 3, based on the raw term frequencies (tf) of the terms in c_i.

$$P(t_k|c_i) = \frac{tf(t_k, c_i)}{\sum_{t_x\in V_{c_i}} tf(t_x, c_i)} \quad (3)$$

4.2 Log-Likelihood for Term-Level Filtering

When two corpora are lexically compared with each other, especially in the case of overlapping time-frames, they may share a large number of terms. Therefore, calculating probability distributions over the entire union set of terms contained in the text of the two corpora may not be an effective method, as all the terms will have equal importance. In this case, there is a need for filtering out the terms that do not distinguish the two corpora well (i.e., terms that are common in both corpora).

The log-likelihood score of a term represents the relative frequency difference of that term in the two different corpora under comparison (Rayson and Garside, 2000). This measure is calculated based on the expected value for term $t_k \in V$ using the total frequency of all terms in the corpus and the actual frequency (or the sum of occurrences) of term t_k in the same corpus. Equation 4 shows how the log-likelihood score is calculated for $t_k \in V$ where N_{c_i} is the total frequency of all terms in corpus c_i, and O_{t_k,c_i} represents the observed frequency of term t_k in the same corpus c_i.

$$E_{t_k,c_i} = \frac{N_{c_i} \sum_i O_{t_k,c_i}}{\sum_i N_{c_i}} \tag{4}$$

The expected frequency of term t_k in corpus c_i denoted by E_{t_k,c_i} is a frequency that is expected for t_k if the occurrences were evenly distributed across the two corpora (Verspoor et al., 2009). The log-likelihood of the term is therefore a measure that tells us how different the actual frequency of the term is from the expected frequency of the same term. For this, the log-likelihood is calculated using Equation 5.

$$LL = 2 \sum_i (O_{t_k,c_i} \ln(\frac{O_{t_k,c_i}}{E_{t_k,c_i}})) \tag{5}$$

An alternative to the log-likelihood measure for statistical analysis of textual corpora is Pearson's χ^2 statistic. This measure assumes a normal distribution of terms in the corpora and has been shown in (Dunning, 1993) to be less reliable especially in the case of small textual corpora with rare terms. Given the relatively small size of the chief complaint corpora for each time-frame, the log-likelihood analysis was preferred here.

Once the log-likelihood of each term was calculated, all of the terms with the log-likelihood below a set threshold were filtered out. The texts of the two corpora now contained only the most important terms that participated in the calculation of probability distributions using JSD.

To estimate the best log-likelihood threshold, we calculated the JSD between consecutive time-frames when filtering terms based on different log-likelihood thresholds ranging from 0 to 20. Since we were not comparing two distinct corpora (cf., (Rayson and Garside, 2000; Baldwin et al., 2013)) but consecutive time-frames over a single corpus, we calculated the JSD values between all of the consecutive time-frames and then took the mean of all JSD values for each possible log-likelihood threshold value. It can be seen from Figure 1 that as the threshold increases, more terms are filtered out and eventually, hardly any terms remain and the divergence between consecutive time-frames approaches zero. A threshold this high is not ideal. We therefore applied the elbow method (Kodinariya and Makwana, 2013) to set the log-likelihood thresholds to 1.75, 3.0, and 1.25 for the one-day, disjoint seven-day, and intersecting seven-day time-frames when filtering out the non-important terms.

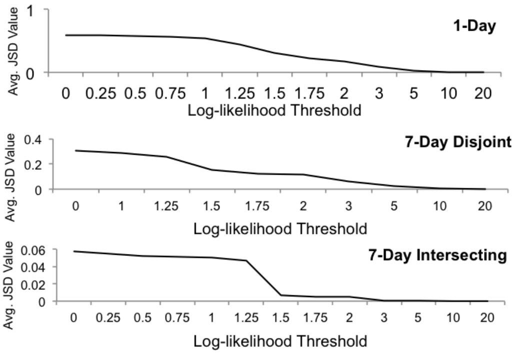

Figure 1: The analysis of the effect of different log-likelihood threshold values on the changes in the JSD distances between consecutive chief complaint corpora in the SynSurv data set

5 Lexical Shift Analysis with Link Set Median

In a separate set of experiments from the statistical JSD distance analysis of the SynSurv data set, we used the Link Set Median (LSM) algorithm (Hoey, 1991) to find segments in the set of chief complaints that suggest lexical shifts in the data set.

The LSM algorithm is based on the idea that a typical text has a cohesive format. Therefore, lexical overlaps can be used to measure the similarity between two sentences. If two sentences share similar terms, then the sentences may be on the same or similar topic. The LSM method identifies lexical repetitions across sentences within a corpus to find textual segments of similar topics. The algorithm has a technique to represent each sentence in the corpus based on its links with other sentences in the same corpus. The lexical link is a repetition between a pair of sentences in the corpus. The set of the links per sentence contains the number of times each sentence has lexical overlaps with any other sentence in the corpus. For instance, if sentence 1 shares two terms with sentence 2 and one term with sentence 4, then “2” is added twice in the link set of sentence 1 as $\{2,2,4\}$. After the link sets have been created for all sentences in the text, the median of the link sets are calculated. This median represents the lexical span of that sentence in the text. If the sentence has a median of “5” for instance, this means that the sentence has a lexical span from $i - 5$ to $i + 5$ where i equals the original position of the sentence. Once each sentence in the text has a corresponding median, an average median of the entire text is calculated. This average median is used as a threshold to find segment boundaries. If the median difference between two consecutive sentences is larger than the threshold, then a segment boundary is placed between two sentences.

In our experiments, we modeled a “sentence” as a set of chief complaints over a specific time-frame and assume a “segment” boundary to be indicative of a lexical shift. We applied the LSM algorithm on the SynSurv data set to understand whether the algorithm will place segment boundaries where the lexical contents of chief complaints deviate and whether such lexical shifts are tied with the changes in the frequencies of syndromic groups.

6 Aberrancy Detection

When using the JSD values to find changes in the term probability distributions in consecutive time-frames, there was a need for a method to detect abnormalities in the probability deviations over time. The Early Aberration Reporting System (EARS) C algorithms (Hutwagner et al., 2003) were designed to detect such changes in syndromic surveillance data. The CUSUM algorithms are based on a statistically detectable change in counts of relevant events over specified windows of time. The EARS implemented three CUSUM algorithms known as C1:mild, C2:medium, and C3:ultra for detecting large, sudden deviations of occurrences of specific events that significantly depart from the norm over time. These algorithms have been named after their level of sensitivity and represent an unsupervised monitoring approach that omits the requirement for long historical data for system training. Since syndromic surveillance systems require real-time analysis of chief complaints, the EARS CUSUM algorithms are ideal. The algorithms have recently been used in another study to find possible disease outbreaks along with machine learning classification techniques (Aamer et al., 2016).

The C1 algorithm calculates the sample mean and the sample standard deviation over rolling windows of samples for $t - 7$ to $t - 1$ days, where t is the current day. C2 adds a two-day lag onto the calculation of the mean and the standard deviation, and C3 is calculated on the basis of the previous two C2 values, details to be found in (Hutwagner et al., 2003). In our work, we used the pre-set threshold values for the C algorithms, i.e., the threshold for C1=3, C2=3, and for C3=2.

Note that the LSM algorithm has an internal method to calculate a threshold based on which textual segment boundaries are identified. Therefore, when using the LSM algorithm, we did not apply the CUSUM algorithms.

7 Results and Discussion

We applied the JSD and log-likelihood filtering algorithms on the chief complaints in the SynSurv data set with the different time-frames described in section 3. Then, the aberrancy detection algorithms were utilized over the JSD measures of consecutive time-frames to retrospectively find any disease outbreaks in the data set. We separately applied the LSM method on the same data set to find likely aberrancies.

Table 1 shows the results obtained (i.e., the number of dates for which an aberrancy was detected) when the aberrancy detection algorithms were applied based on the JSD lexical distribution values, derived from the chief complaints in the SynSurv data set over the different time-frames. As can be seen in Table 1, the C2 and C3 algorithms with higher levels of sensitivity set off a large number of signals. Algorithms that are overly sensitive to fluctuations in disease frequencies are not ideal; they may alert health practitioners too frequently, resulting in alert fatigue and mistrust of the algorithm. On the other hand, algorithms that miss viable shifts corresponding to meaningful events are also not desirable. The C1 algorithm is the least sensitive of the three algorithms while still raising alerts; it appears to balance the two criteria most effectively.

Note that to apply the aberrancy detection algorithms on the resulting JSD values over the disjoint seven-day windows, there was a need for 7 to

Time-frame	Method	#Aberrancies with ADA		
		C1	C2	C3
1-day	JSD+LL	64 (2.4%)	67 (4.0%)	205 (13.2%)
7-day disjoint	JSD+LL	10 (4.6%)	14 (6.5%)	30 (13.8%)
7-day intersecting	JSD+LL	36 (2.4%)	61 (4.0%)	201 (13.3%)

Table 1: The number of aberrancy dates detected by different algorithms in the chief complaints over different time-frames. The total number of windows differed for each time-frame. The 1-day time-frame consisted of 1522 windows starting from July 2005 to August 2009, the 7-day intersecting time-frame had 1516 windows, and the 7-day disjoint time-frame had 217 windows. Note: ADA=Aberrancy detection algorithm, JSD=Jensen-Shannon Divergence, and LL=Log-likelihood.

11 weeks of baseline data which would mean a 7 to 11 week wait before any aberrancy can be detected. For this reason, in practice, it may not be useful to apply the C algorithms on the JSD values for disjoint seven-day time-frames.

Further analysis of the C1 algorithm, however, showed that the weeks for which C1 detected outbreaks were spread throughout the years in the data set, indicating that LSM is highly sensitive and produces too many noisy signals, which is not desirable. Since the texts of chief complaints are all in a single text corpus assigned to various syndromes, a single chief complaint may correspond to more than one syndrome. For example, due to the very similar set of syndromes that *Flu-Like Illness* and *Acute Respiratory* share, a single chief complaint may be labelled *Flu-Like Illness*, *Acute Respiratory*, and *No Diarrhea* at the same time. Therefore, a large lexical shift in the SynSurv data set may correspond to any of the syndromes (including the original set of syndromes introduced in section 2). As a result, there was a need for a more informed method to distinguish between the aberrancies related to each syndrome. For this, instead of using log-likelihood of terms, we collected all the chief complaints signaling the presence of a syndrome into one corpus and extracted the terms with the largest term frequency for each syndrome. Table 2 shows the top 10 terms for the three syndromic groups. We can see is a large intersection for *Flu-Like Illness* and *Acute Respiratory*.

We calculated JSD values over the SynSurv data set with the original chief complaint terms (no filtering) removing stop words only, as well as with the top 10 terms based on term frequencies for each syndromic group. Figure 2 shows the resulting diagrams for *Flu-Like Illness* and *Diarrhea* over the different time-frames. The results for *Acute Respiratory* were not shown since this syndromic group is most similar to *Flu-Like Illness*.

TF rank	Diarrhea	FLI	AR
1	pain	sob	sob
2	vomiting	hr	hr
3	abdo	cough	cough
4	diarrhea	pain	chest
5	hr	o/a	pain
6	nausea	chest	o/a
7	nil	nil	hx
8	o/a	hx	respiratory
9	gastrointestinal	throat	nil
10	hx	respiratory	phx

Table 2: The top TF ranked terms related to each syndrome in the SynSurv data set. Note: TF=Term frequency, FLI=Flu-Like Illness, and AR=Acute Respiratory. The terms include many medical abbreviations such as “sob” for shortage of breath, “hr” for heart rate, etc.

When comparing the effect of the different time-frames, as shown in Figure 2, the results with the one-day window time-frames seem much noisier than those of the two other types of time-frames. It is hardly possible to detect any aberrancies in the one-day time-frame output data whereas with the seven-day windows, both disjoint and intersecting, some clear trends of peaks can be detected around a few dates in the data set.

More importantly, the utilisation of the top terms significantly changed the outputs of the aberrancy detection as indicated by the different patterns in the diagrams in Figure 2 with the original versus top 10 terms. However, it is noticeable that the peaks form at very similar dates for both *Flu-Like Illness* and *Diarrhea* conditions, when the top 10 terms are retained in the analysis. We suspect this is largely due to the significant intersection between the list of top 10 terms for these two syndromic groups, as seen in Table 2. Indeed, there is a 50% overlap between the lists of top frequency terms including “pain, “hr”, “o/a”, “nil”, and “hx”. These overlapping terms are general;

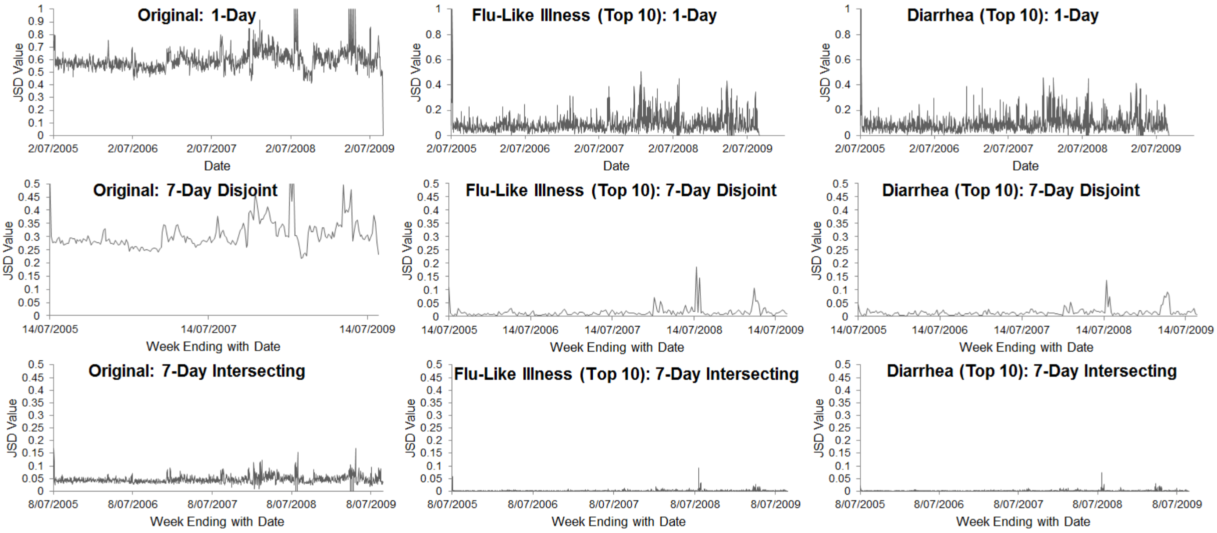

Figure 2: Variation of JSD values in the SynSurv data set with different time-frames when original and top 10 terms retained

they seem non-specific to any syndrome.

We then experimented with the top frequency terms that were not shared between the lists of the two syndromic groups. We took the top 3 terms from the two lists in Table 2 and calculated the JSD values over the disjoint one-day windows of chief complaints. The results are shown in Figure 3 in which the patterns of peaks are different with drastic increases of JSD values at different and separate dates. This suggests that if effective and descriptive sets of terms are found to represent each syndromic group (hence semi-supervised), the lexical shift as measured with deviating JSD values can be utilised as an indication of possible outbreaks of corresponding syndromes.

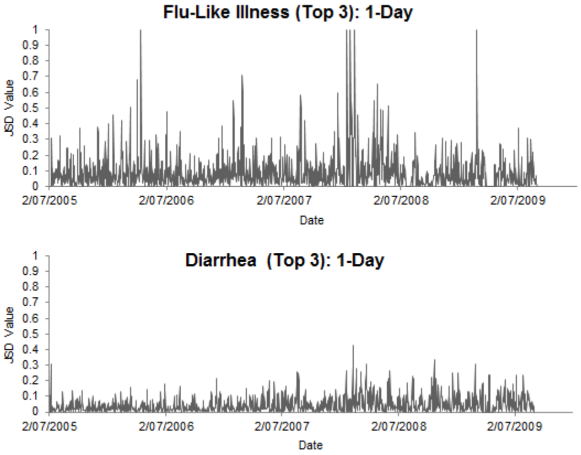

Figure 3: Variation of daily JSD values in the SynSurv data set when only top 3 terms retained

As a proof of concept to understand whether JSD is in fact sensitive to the raw frequencies of positive cases of syndromic groups, we cross-checked deviations of JSD values (over all terms) with reference to the actual positive cases of each syndrome in the SynSurv data set. The results are shown in Figure 4 where JSD is demonstrated to be sensitive especially to the cases where there are no positive labelled chief complaints for any of the syndromes. In such cases, as highlighted in the diagrams of Figure 4, JSD values deviate drastically, indicating a rapid change in the frequencies of syndromes of interest.

In the last round of experiments, we applied the LSM algorithm over the different time-frames of the SynSurv data set. The algorithm detected 77 segments with the disjoint seven-day time-frame, 178 with the one-day window, and 268 with the intersecting seven-day time-frame. When analysing the dates of segmentations, it was observed that although the dates were spread throughout the years, the segments by the disjoint seven-day and one-day time-frames were mostly in the years 2007, 2008, and 2009, similar to the times when the peaks were observed with JSD. However, the 268 segments found using the intersecting seven-day time-frame were distributed over the four years. This may be due to the large vocabulary overlaps where the windows intersect over six days of chief complaints. The intersecting seven-day time-frame, therefore, may not facilitate an effective process of outbreak detection using lexical

Time-frame	Method	# of Segments
1-day	LSM	178 (11.7%)
7-day disjoint	LSM	77 (35.5%)
7-day intersecting	LSM	268 (17.7%)

Table 3: The number of segments detected in the chief complaints over different time-frames. The total number of windows differed for each time-frame. The 1-day time-frame consisted of 1522 windows starting from July 2005 to August 2009, the 7-day intersecting time-frame had 1516 windows, and the 7-day disjoint time-frame had 217 windows. Note: LSM = LinkSetMedian

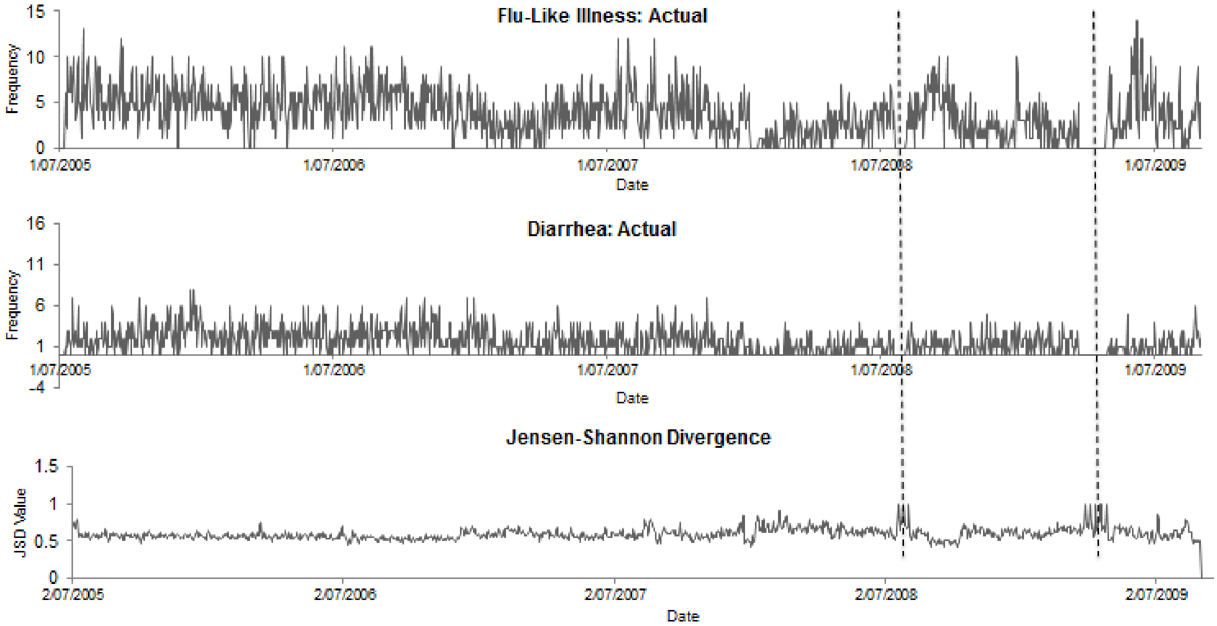

Figure 4: Analysing sensitivity of lexical shifts measured with JSD to actual deviations of the number of daily syndromic chief complaints. Vertical lines point to the dates when drastic disease frequency changes align with large JSD values.

shift analysis with LSM. Based on our results, the large number of deviations/outbreaks detected by the LSM algorithm (in its current form) seems to be an impediment in using the algorithm for syndromic surveillance.

8 Conclusion

We proposed a new semi-supervised method to perform syndromic surveillance over the SynSurv data set containing a large number of emergency department Chief Complaints in the Victoria State of Australia. This new method is based on locating significant lexical divergences in the texts of chief complaints accumulated over consecutive periods of time. We analysed the lexical shifts using Jensen-Shannon Divergence (JSD); a probability distribution divergence measure; and Link Set Median (LSM); a text segmentation algorithm; for the three syndromes *Flu-Like Illness*, *Acute Respiratory*, and *Diarrhea*. The aim was to find whether lexical shifts are tied with possible disease outbreaks in the historical SynSurv data set. We evaluated the lexical shifts with three types of time-frames: i) one-day windows, ii) disjoint seven-day windows, and iii) intersecting seven-day windows advancing by one day.

We found that all three time-frames had some limitations: the disjoint seven-day time-frame when used in combination with the EARS C algorithms is not efficient, and inherently, it will result in longer periods of wait before a likely outbreak is signalled. The seven-day intersecting time-frame, on the other hand, resulted in noisy and frequent signals with the LSM algorithm, mostly as a result of large textual overlaps in consecutive overlapping time-frames. The one-day time-frame produced interesting results but suffers from the day-of-the-week effect.

Our results also demonstrate that if each syndromic group (i.e., disease) is represented with its corresponding distinguishable high-frequency terms, then the JSD measure provides evidence for lexical shifts that is aligned with drastic changes in the frequency of syndromic-labelled chief complaints. The need for distinguishable terms for each syndrome under consideration means that our methods are semi-supervised. Based on our experiments, therefore, the JSD method with syndrome-specific term sets analysed over the one-day time-frames resulted in the most promising outcomes.

In future work, we plan to expand the idea of using high frequency terms into the utilisation of related semantic representations, such as health-related synonyms. In addition, we are planning to analyse lexical shifts in chief complaints using other natural language processing techniques, such as textual novelty detection with Topic Tracking and Detection algorithms. TDT algorithms can track changes in text and find event-level shifts (or first stories) in a corpus. We would like to experiment with such unsupervised algorithms and find whether first story boundaries match drastic changes in the frequencies of chief complaints related to specific syndromic groups.

Acknowledgments

We acknowledge the Australian Defence Science and Technology Group (DST Group) for the support of our work on syndromic surveillance.

References

Hafsah Aamer, Bahadorreza Ofoghi, and Karin Verspoor. 2016. Syndromic surveillance on the victorian chief complaint data set using a hybrid statistical and machine learning technique. In K. Barbuto, L. Schaper, and K. Verspoor, editors, *Proceedings of the Health Data Analytics Conference*.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Colleen A. Bradley, H. Rolka, D. Walker, and J. Loonsk. 2005. BioSense: implementation of a national early event detection and situational awareness system. *MMWR Morb Mortal Wkly Rep*, 54(Suppl):11–19.

Wendy W. Chapman, Lee M. Christensen, Michael M. Wagner, Peter J. Haug, Oleg Ivanov, John N. Dowling, and Robert T. Olszewski. 2005. Classifying free-text triage chief complaints into syndromic categories with natural language processing. *Artificial intelligence in medicine*, 33(1):31–40.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Jeremy U. Espino, John Dowling, John Levander, Peter Sutovsky, Michael M. Wagner, and Gregory F. Cooper. 2007. SyCo: A probabilistic machine learning method for classifying chief complaints into symptom and syndrome categories. *Advances in Disease Surveillance*, 2(5).

Ronald D. Fricker Jr, Benjamin L. Hegler, and David A. Dunfee. 2008. Comparing syndromic surveillance detection methods: EARS versus a CUSUM-based methodology. *Statistical Medicine*, 27(17):3407–29.

M. Hoey. 1991. *Patterns of lexis in text*. Oxford University Press.

Lori Hutwagner, William Thompson, G. Matthew Seeman, and Tracee Treadwell. 2003. The bioterrorism preparedness and response early aberration reporting system (ears). *Journal of Urban Health*, 80(1):i89–i96.

Trupti M. Kodinariya and Prashant R. Makwana. 2013. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Bahadorreza Ofoghi and Karin Verspoor. 2015. Assessing the performance of American chief complaint classifiers on Victorian syndromic surveillance data. In *Proceedings of Australia's Big Data in Biomedicine & Healthcare Conference*, Sydney, Australia.

P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora, held in conjunction with ACL 2000*, pages 1–6.

Fu-Chiang Tsui, Jeremy U. Espino, Virginia M. Dato, Per H. Gesteland, Judith Hutman, and Michael M. Wagner. 2003. Technical description of RODS: a real-time public health surveillance system. *Journal of the American Medical Informatics Association*, 10(5):399–408.

Karin Verspoor, K. Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and Open Access scientific journals are similar. *BMC Bioinformatics*, page 10:183.