MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-437

Using ProMED-Mail and MedWorm Blogs for Cross-Domain Pattern Analysis in Epidemic Intelligence

Avaré Stewart, Kerstin Denecke

L3S Research Center, Hannover, Germany

Abstract

In this work we motivate the use of medical blog user generated content for gathering facts about disease reporting events to support biosurveillance investigation. Given the characteristics of blogs, the extraction of such events is made more difficult due to noise and data abundance. We address the problem of automatically inferring disease reporting event extraction patterns in this more noisy setting. The sublanguage used in outbreak reports is exploited to align with the sequences of disease reporting sentences in blogs. Based our Cross Domain Pattern Analysis Framework, experimental results show that Phase-Level sequences tend to produce more overlap across the domains than Word-Level sequences. The cross domain alignment process is effective at filtering noisy sequences from blogs and extracting good candidate sequence patterns from an abundance of text.

Keywords:

Biosurveillance, Automatic data processing, Medical informatics applications, Investigative techniques

Introduction

Many factors in today's changing societies contribute towards the continuous emergence of infectious diseases. In response, Epidemic Intelligence (EI) has emerged as a type of intelligence gathering which aims to detect events of interest to the public health, from the unstructured text of news and outbreak report.

In a typical Epidemic Intelligence scenario (Figure1), disease reporting events (i.e. victim, location, time, disease) are extracted from raw text. The events which are considered to be relevant for detecting an emerging disease are annotated with additional information (such as threat or severity level) and then aggregated to produce signals. The signals are intended to be an early warning against potential public health threats, and the epidemiologist uses them to assess risk; or corroborate and verify the information locally and with international agencies.

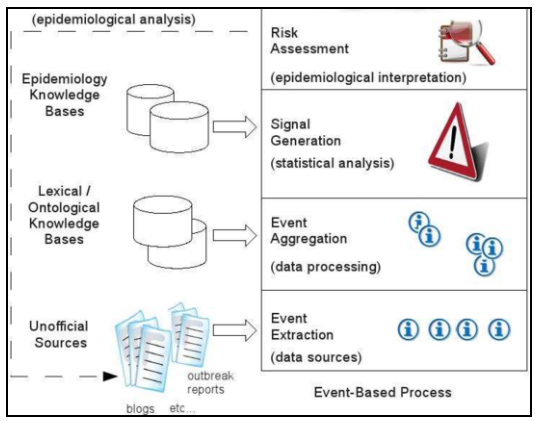

Figure 1- Epidemic Intelligence Scenario

In such a scenario, the diversity of the sources plays an important role in the intelligence gathering process. Medicine 2.0, social medical blogs and other forms of user generated content can be seen as an additional source. These sources are of significance, since those who experience as well as treat disease first hand, and describe their experiences in blogs and other forms of social media.

These present-day systems use news and outbreak reports as sources of information to support intelligence gathering. Moderated systems, (GPHIN, ProMED-Mail [5] ARGUS [9]) rely upon the interpretive and analytical expertise of analysts to filter and extract information about health threats. In automatic approaches such as MedISys [7], HealthMap [2], and BioCaster [1], all stages of the document collection, filtering and processing is done with little or no human intervention. Disproportionately, these systems do not take user-generated content, such as blogs, videos and twitters into account.

In order to fully realize the value of using Medicine 2.0, several challenges must be considered. First, acquiring information relevant to public health events from blogs requires filtering a huge number of irrelevant sentences. Those sentences which are irrelevant, or of poor quality need to be filtered out before any events can be extracted. Second, since natural language is inherently ambiguous, there are many ways to express

the same meaning. Particularly for social media, the complexity is increased given its volume, variety, evolution and informal nature (i.e. interspersing of subjective and factual information, many contributing authors with different styles, topic drifting prose and special lingo) [6]. Finally, abundant well labeled corpora, suitable for supporting many of the tasks in epidemic intelligence – is scarce or non existent.

In this work we seek to overcome these shortcomings, by demonstrating the underlying premise that if two sources discuss the same things - disease outbreak victims – then the languages have common linguistic structures. Thus, the symbols in one corpus can be used to identify them in another one.

In order to better understand which structures are potentially useful for extracting events from the sentences of blogs, we investigate how to transfer sentence structures acquired from a moderated data source (ProMED-Mail), and how this can be leveraged for the structures detection process within a new and less amenable one (MedWorm blogs), based on the commonality of the two used languages.

In this exploratory analysis we seek to understand the following questions with regard to a cross-corpora alignment:

- When do corpora align?
- What type of features (linguistic structures) best characterize such an alignment?

Methods

We cast the problem of Cross Domain Pattern Analysis as a alignment problem where the alignment sought is among the sentences of ProMED-Mail and MedWorm (see ***Figure 2***). In this approach we compare the sentence structures acquired from an auxiliary source (ProMED-Mail), to see if sentences relevant to outbreak reporting can be detected in a target domain (MedWorm blogs).

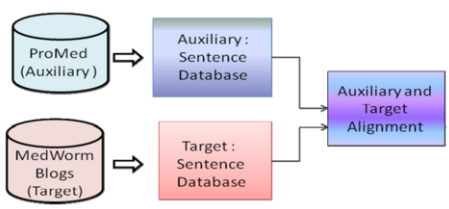

Figure 2- Cross Domain Pattern Analysis Overview

Structural Representations for Sentences

The sentences from each domain can have different levels of structuring (see Table 1). At the token level, the order of the individual words is considered unimportant; and can consist of (1a) the original word of the sentence (1b) a word's stem or (1c) a word's named entity. At the next structural level, one or more tokens can be sequenced lexically. For example, the named entity tokens derived from the Unified Medical Language System[1] (UMLS) concept identifiers (CUI) or the semantic classes (TUI) can be used to define a token sequence based on named entities (2a). An example sequence constructed from named entities using CUI and TUI is shown in ***Figure 3***. Finally, at the highest structural level, the sentence can be represented as a parse or dependency tree; respectively capturing the structure and semantics of the sentence.

Table 1-Structural Representations for Sentences

Structural Level	Example Representation
1. Token	a) Word b) Stem c) Named-Entity
2. Sequence	a) Token Sequence b) Phrase (i.e.: noun, verb)
3. Hierarchical	a) Parse tree b) Dependency tree

Sentence Database

Given a structural representation applied to all the sentences, in both the auxiliary and target domains, we construct a **Sentence Database.** An example sentence from a sentence database using the UMLS concept identifiers and semantic classes is shown in ***Figure 3***.

A 23 **year old** (C0419638, T061) **woman** (C0043209, T098) from Thanh **Hao** (C1435591, T116) **Providence** (C1227418, T005) died of **bird flu** (C0016627, T047) on **Wednesday** (C0585027, T079)[22 **April** (C1332090, T028) 2009]	
UMLS CUI Sequence	<C0419638,C0043209,C1435591, C1227418, C0016627, C0585027, C1332090>
UMLS TUI Sequence	<T061,T098,T116,T005, T047, T079, T028>

Figure 3-Example Sentence with Sequences Constructed from Named Entities using CUI and TUI

Definition: Cross-Domain Pattern Analysis

Given: a) a Sentence Database for an auxiliary domain, b) a second Sentence Database for a target domain and c) a comparison metric; Cross Domain Pattern Analysis is defined as follows:

Find an alignment between sentences having the same structural representation, of the form X -> Y, where X is a structure from the auxiliary domain; Y is a structure from the target domain and X is aligned with Y when the value of the comparison metric used is above a threshold value.

[1] http://www.nlm.nih.gov/research/umls/

Results

Experimental Goals

In order to automatically determine suitable sentences for extracting events from the sentences of blogs, we align sentence structures acquired from ProMED-Mail and MedWorm blogs. The goal of these experiments it to understand: 1) when corpora align and 2) the type of linguistic sentence structures best characterize such an alignment.

Data Collection

ProMED-Mail data was collected directly from the website[2], and MedWorm blogs via RSS. The blog data was collected by We used the blog urls from the RSS feed to get the full content of each blog post. The data for the outbreak reports was collected during an eight month period from January 1, 2009 through August 12, 2009. This coincided with the known outbreak period of the Swine Flu influenza, which reached pandemic status in June, 2009 [3]. The total amount of data used is shown in *Table 2*.

Sentence Structures

To construct the Sentence Databases for the auxiliary and target domains, the documents were first split into sentences using the Stanford Parser[4]. Next, several structural representations for each sentence were produced, this included: a) stems; b) TUI sequences, c) phrase sequences and d) parse tree (see ***Table 1***) . Stems were created using the Porter Stemmer available with Apache Lucene[5], the phrases were obtained by using MetaMap and both the parse and dependency trees were constructed using the Stanford Parser.

Table 2-Summary of the Amount of Data Used

	ProMED Mail	MedWorm Blogs
Documents	1,129	2,171
Sentences	32,518	130,487
Tokens	542,229	1,489,686

Pre-Filtering and Alignment

Pre-Filtering: Not all sentences were used in the Cross Pattern Alignment process. We exploited the characteristic style of ProMED-Mail, where the information about an outbreak is typically expressed in the first sentences of the text. We experimentally used the top-3 sentences. For blogs, we used the top-10. Documents below a 300 Kb size were not included in the experiments.

A*lignment:* To align the structures across the domains, sentence structures of the same type from the target and auxiliary domains were compared using three similarity metrics; one for each structural level; the range of values for all scores were bounded by 0 (not similar) and 1 (equal).

Experiment I: Token Level Alignment

In this section, we present a token level alignment for terms in appearing in both corpora. The metric used for comparing tokens was based on the use of wildcards concatenated with stems. For example , the stem "outbreak" was concatenated with a wildcard to allowing matching morphological variations of the token. The salient terms are visible in the word clouds for MedWorm (***Figure 4***) and ProMED-Mail (***Figure 5***)

Figure 4-MedWorm Stemmed-Word Cloud

Figure 5-ProMed-Mail Stemmed-Word Cloud

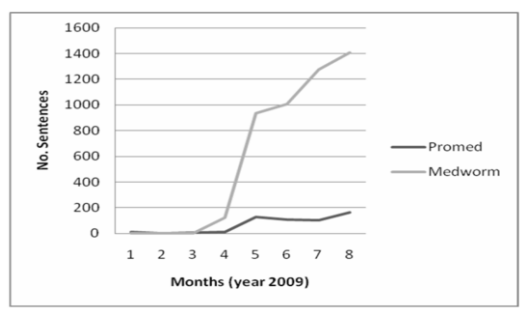

Figure 6-Total Sentences per Month for "h1n1"

[2] http://www.promedmail.org
[3] http://www.who.int/csr/disease/swineflu/updates/en/index.html
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] http://lucene.apache.org

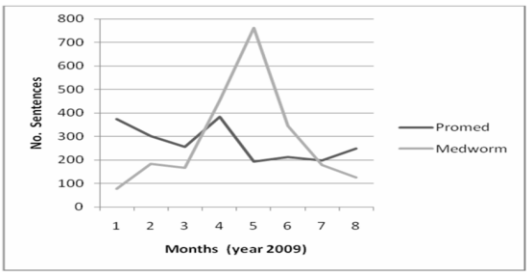

Figure 7-Total Sentences per Month for "outbreak"

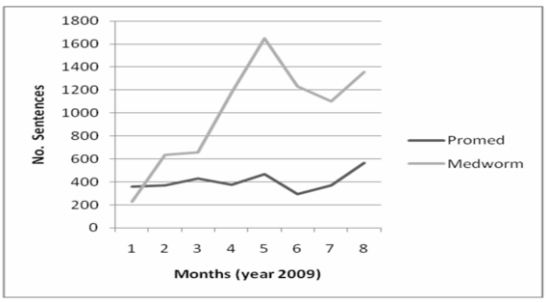

Figure 8-Total Sentences per Month for "virus"

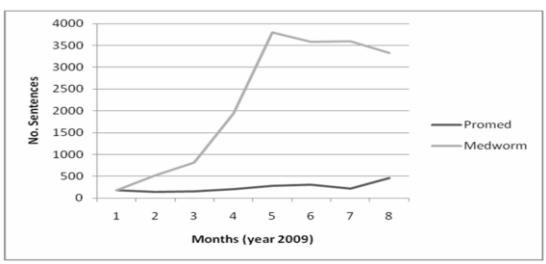

Figure 9-Total Sentences per Month for "flu"

Experiment II: Sequence Level Alignment

In this section, we present a sequence level alignment. The similarity metric used was based on our implementation of a string comparison algorithm which computes the longest overlapping contiguous subsequence of named-entity tokens. The score was normalized by the length of the longest sequence. Pairs with no named entity tokens in common received a score of 0. If all named-entity tokens in the sequence matched, one-for-one, then the pairs received a similarity score of 1.0.

Alignment of Token Sequences

Table 3 depicts the overlap between the sentences in ProMED-Mail and MedWorm for TUI and CUI Word-Level annotations varying the sequence lengths. The overlap between TUI annotated ProMED is far greater in number than the CUI annotations since several CUIs map to the same TUI semantic class. However, we notice that when the length of the sequences exceeds three, the proportion of overlap between ProMED and MedWorm for both types of annotations remains relatively the same.

Table 3-TUI and CUI Sturcture vs. Sequence Length

Number of Aligned Sentences Between ProMED-Mail and MedWorm Using TUI and CUI		
Sequence Length	TUI	CUI
3	6,105	422
4	130	136
5	30	20
6	15	16
7	10	9
8	8	6

Alignment of Phase Sequences

For the Phrase level structures were constructed by using the sentence chunks produced by the MetaMap parser. These chunks were then sequenced together to form the phrase sequences used in the experiments. Pairs of Phrase sequences were considered similar if there was and exact match.

Table 4-Phrase Sturctures vs. Sequence Length

Number of Aligned Sentences Between ProMED-Mail and MedWorm Using Phrases	
Sequence Length	Phase Structure
3	1,180
4	202
5	2

Experiment III: Hierarchical Level

For the hierarchal level, the parse tree for each sentence was used; to compute similarity for this structure, a tree kernel was used (http://dit.unitn.it/~moschitt/Tree-Kernel.htm). The results for tree alignments proved to be much less effective than the above methods, as no scores above .5 were obtained.

Experiment IV: Sequence Quality

In Experiment IV we examine the quality of the sequences with respect to the disease reporting task. ***Table 5*** shows the percentage of overlapping sequences which contain the semantic classes Population Group, Geographical Location, Temporal Concept or Disease. As can be seen, the percentage of sequences containing mentions to population groups and diseases are roughly the same for low support (10 % – 20%). Notably, as the support increases, the proportions change significantly and more overlapping sequences contain mentions to groups (30% support) and geographical location (%40 support).

In addition, we made a human assessment of the cases with a very small number of overlapping sequences. We find that these patterns contain interesting subsequences from which good candidate patterns can be built. One such subsequence

contains the location, disease and public health agency involved in responding to the event: e.g.:

TUI: Vietnam (Geographic Area), **from the disease** (Disease or Syndrome), **Health Ministry (**Health Care Related Organization)

Table 5-Sequence Quality Based on Entity Types

Sequence Quality Bases on Outbreak Report Named Entity Types				
Support Percent	Group	Geographical	Temporal	Disease
10	306	73	150	178
20	31	7	2	82
30	18	2	1	2
40	4	82	0	3

Discussion

The experimental goals of this work were to determine if an increasing level of abstractions in representing sentence sequences play a role in characterizing the sentential patterns between moderated and blog information sources.

In our analysis, we have seen that Phase-Level sequences tend to overlap more than Word-Level sequences. This can be explained by the fact that phases aggregate words and thus there is less variation across domains for these aggregations.
When aligning mined frequent sequences, the support played less of a role than the sequence lengths. Although very small numbers for higher sequence lengths were obtained, we believe this to be encouraging and demonstrate the ability of our approach to filter noise from large amounts of blogs data which a relative small about of human interpreted data.
We also experimented with other types of annotation such as POS and mined rare sequences instead of just frequent ones. Although not presented, these results showed less promise.
A notable limitation of the match phase used in these experiments is that the absence of a „sliding window" to align subsequences with a sequence. More rigorous matching could be used and is considered as future work.

Conclusion

In this work we motivate the use of medical blog user generated content for gathering facts about disease reporting events to support biosurveillance investigation. Given the characteristics of blogs, the extraction of such events is made more difficult due to noise and data abundance.

We address the problem of automatically inferring disease reporting event extraction patterns in this more noisy setting based our Cross Domain Pattern Analysis Framework.

The experimental results show that Phase-Level sequences tend to produce more overlap across the domains than Word-Level sequences and that the cross domain alignment process is effective at filtering noisy sequences and the extracting good candidate sequence patterns from an abundance of text.

As future work, we would like to take into account the significance of the errors associated with the annotation process itself. Also for the future work, we intend to examine how rules can be applied across the domains. Finally, we also intend to compare how the quality of the learnt patterns compares with patterns which are induced from predefined entity pairs in a greedy approach.

References

[1] Collier N, [et al.] BioCaster: detecting public health rumors with a Web-based text mining system [Article] Bioinformatics. - 2008. - Vol. 24.

[2] Freifeld CC [et al.]HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports [Article] J Am Med Inform Assoc. 2008. - 2008. - 2 : Vol. 15.

[3] Ji Heng [et al.] Cross-document Event Extraction and Tracking: Task, Evaluation, Techniques and Challenges [Conference] Proc. Recent Advances in Natural Language Processing. - 2009.

[4] Ji Heng and Grishman Ralph Refining Event Extraction through Unsupervised Cross-document Inference [Conference] Proceedings of the Annual Meeting of the Association of Computational Linguistics. - Ohio, USA : [s.n.], 2008.

[5] Madoff LC ProMED-mail: an early warning system for emerging diseases [Article] Clin Infect Dis. - 2004. - 2 : Vol. 15.

[6] Moens Marie-Francine Information extraction from blogs [Book]. - 2009. - pp. 469--487.

[7] Steinberger R [et al.] Text mining from the web for medical intelligence [Book Section] Mining Massive Data Sets for Security. - Amsterdam : IOS Press, 2008.

[8] Wilson JM [et al.] A heuristic indication and warning staging model for detection and assessment of biological events [Article] J Am Med Inform Assoc. - 2008. - 2 : Vol. 15.

Address for correspondence

Avaré Stewart
L3S Research Center
Appelstr. 9A
30167 Hannover, Germany
stewart@L3S.de