Vol.2012-ICS-168 No.5
2012/9/19

Text Mining in Action: Early Alerting of Disease Outbreaks

NIGEL COLLIER[1,a)]

Abstract: Public health disasters involving the spread of disease such as SARS 2002 and Influenza A H1N1 2009 graphically depict the danger of a global pandemic and the key role of surveillance. Traditionally governments and agencies have used indicators such as over-the-counter sales of medicines and general practitioner surveys as early warning systems together with notifications from the World Health Organisation. Recently though a new class of system is becoming available based on detection from digital Internet media such as news and microblogs. At the heart of these automated information gathering systems is text mining technology. The contribution of this paper is to provide a brief overview of the history and role of such systems with particular emphasis on the BioCaster project at the National Institute of Informatics in Japan.

1. Background

Public health surveillance involves the early identification, assessment and verification of potential public health hazards and the timely dissemination of alerts to appropriate stakeholders. Among the types of methods available, event surveillance techniques emphasize sifting through large volumes of dynamically changing unstructured data such as news reports and government bulletins. Such sources are abundantly available, rapidly updated and freely accessible on the World Wide Web. In most countries though only a few highly trained professionals are available to do this task. Text mining technology [1] has naturally come to be applied to this task [2] since the burden of information overload poses severe restrictions on scarse human resources. Text mining aims to develop high performance algorithms for converting unstructured textual data to a machine understandable format. Computer systems can then perform analysis and deliver results in customised forms depending on the user's interest and requirement. Such systems though tend to be specialised and therefore are not perhaps as well known as more traditional business intelligence or biomedical text mining systems. This paper seeks to provide an outline of the technology being applied, pointers to active and historical systems, and an overview of one complete system that I and my colleagues developed at the National Institute of Informatics in Japan. This system, BioCaster, is in operational use by the public and international agencies around the world.

2. Survey

As shown by Hartley *et al.*'s survey paper [3], event-driven surveillance systems are now widely used by national and transnational public health organisations such as the World Health Organisation (WHO), the Centers for Disease Control and Prevention (CDC) and the European Centre for Disease Prevention and Control (ECDC), Public Health Agency of Canada (PHAC) and many other agencies. In November 2002 at the start of the SARS epidemic, the GPHIN system [4] at Public Health Canada was among the earliest, along with the ProMED-mail network [5], to provide early warning of the impending SARS pandemic. During the Influenza A H1N1 pandemic in 2009 a number of systems are credited with the timely discovery of early events including MedISys [6], HealthMap [7] and BioCaster [8]. Tools such as Riff from Instedd [9] were used to enhance decision support by integrating signals from virtual teams of experts with multiple streams of data from EI systems such as EpiSpider [10], SMS messages and electronic medical records in OpenMRS. Additionally the MEDCollector system also aims to integrate multiple Web-based sources [11]. Of research interest are two early systems: Proteus-Bio [12] and MiTaP [13]. Having timely and well informed information helps governments to take the right actions to reduce the length and severity of an infectious disease outbreak. This information is important not only for pandemic influenza but also for many other diseases such as measles and mumps as well as more exotic diseases like chikungunya.

The challenge for text mining technology is to provide accurate interpretations of real world situations from facts that maybe vague or scattered throughout several reports. Ambiguity caused by a range of factors complicates the analysis. For example, the name of a location such as *Camden* could refer to several towns or cities throughout the world including locations in the UK, Australia and the USA. Toponym grounding is only part of the problem however. Other challenges include:

- Temporal grounding: identifying when the event took place can sometimes be surprisingly difficult, e.g. when an article

[1] National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] collier@nii.ac.jp

talks about the historical background to a pandemic it may discuss the Spannish Flu of 1918.

- Entity grounding: whilst new vocabulary for disease names grows at a relatively slow rate, there is a significant problem with metaphoric terms such as *Obama fever* which cause confusion for automated systems.
- Variant transliterations: there is often considerable variations in transliterated place names particularly for Arabic, e.g. *Tajoura*, *Tajura*, and *Tajoora*.
- Coreference resolution: this is necessary to resolve both entities and quantities such as the number of victims. For example consider the following two statements that might appear in different news reports: (a) *Two British holidaymakers fell ill*, and (b) *Two male pensioners died*. To fully understand that the count of victims was 2 and not 4, *British holidaymakers* needs to be resolved with *male pensioners*, and the relationship between being ill and dying also needs to be explicated.
- Vagueness: catching outbreaks as early as possible requires systems to make decisions with vague reports of mystery illnesses. In such cases other evidence such as the severity of the disease or the number of victims or the speed of transmission needs to play a more important role just as it does for human analysts.

3. BioCaster

3.1 History

In this section I provide an overview and description of the BioCaster system, one of less than ten major automated/semi-automated systems that is currently active for epidemic intelligence in the world. BioCaster has participated as an invited system within the G7 Global Health Initiative's Early Alerting and Reporting (EAR) initiative and makes its output available to a variety of public health agencies across Japan, the European Union (EU), the United States of America (USA), Canada and the World Health Organisation (WHO).

BioCaster [8] begin in 2006 with grant-in-aid funding from the Japan Society for the Progress of Science (JSPS) for a core text mining pipeline to process four languages (English, Japanese, Thai and Vietnamese). Over the next few years an ontology - or conceptual model - was conceived [14] that modeled the key entities and relations within the target domain and across languages. The conceptual modeling drew on a new method from the formal ontology community called OntoClean [15]. Using this we were able to show an empirical improvement in the quality of entity recognition [16] within the text mining system. In 2008 the system greatly expanded after grant-in-aid funding from the Japanese Science and Technologies' (JST) Sakigake fund. The ontology was expanded to define over 300 diseases including those of humans, animals, plants as well as the toxc effects of certain chemicals and radionucleotides. The new ontology bridged terminology across twelve languages and was publicly released [17]. A further milestone was the systematic investigation of automated geo-temporal alerting based on time series analysis on the event frames from the text mining system [18], [19]. More recent work has looked at incorporating health signals from Twitter microblogs [20].

3.2 Platform

The BioCaster system is composed of a backend high performance cluster and front end Web server which provides user services. The Web server is implemented on a 24 x 2.66 GHz Xeon core server running on Ubuntu Linux, Apache, PHP and MySQL. The backend server uses the Rocks cluster middleware (http://www.rockcsclusters.org) on a platform of 48 3.0GHz Xeon cores. Rocks incorporates the Sun Grid Engine, allowing BioCaster to achieve data parallelism thereby providing near-real time text processing capabilities. This is a crucial requirement as any delay in notifying news events will impact on the human analyst's trust in the system.

3.3 Ontology

Our early discussions with domain experts in the public health community revealed that timely information on the following are especially important: (1) outbreaks of newly emerging diseases such as novel pandemic influenza, (2) the moment of transition from animal-to-human and sustained human-to-human transmission, (3) the importation of exotic diseases across international borders and (4) accidental or deliberate release of biological, chemical, radiological or nuclear agents. These are precisely a reflection of the concerns addressed within the revised International Health Regulations (IHR) [21] of the World Health Organization (WHO).

Domain modeling in BioCaster is encapsulated through the BioCaster ontology (BCO), a freely available public health applications ontology designed to integrate laymen's language of disease reporting across twelve languages (http://code.google.com/p/biocaster-ontology). Formal concept analysis [14] was used to organize the BCO around a backbone SUMO (Suggested Upper Merged Ontology) upper-level taxonomy [22] onto which domain entity classes such as DISEASE, PERSON, ORGANISATION, COUNTRY, PROVINCE, SYMPTOM and CHEMICAL were carefully grafted. Root terms - the key concepts that play roles in events - appear as instances of the domain entity classes. The selection of terms centerd on the diseases which were selected from country notifiable disease lists and ranked for public health impact. The resulting ontology is made available for browsing on the BioCaster portal site and also as a free downloadable OWL (Web Ontology Language) file.

The third version of the ontology was released in 2009 and encoded multilingual equivalences between eleven languages: Chinese, English, French, Indonesian, Japanese, Korean, Malay, Spanish, Russian, Thai and Vietnamese. Cross language term equivalences are handled as multilingual synonym sets in a manner similar to EuroWordNet [23]. The 2009 version contains over 300 human and animal diseases.

3.4 Text Analytics

The BioCaster system is constructed from a linear pipeline of processes as described below:

(1) **Data ingestion** of sources can come from a variety of document types such as newswire reports, business reports and

blogs (Web logs). Contents in BioCaster are accepted in RSS (Really Simple Syndication) format, making it straightforward to download directly from a link address on a specific topic.

(2) **Data cleaning** is a technologically mundane process but one which is vital in practice to remove unwanted noise from the text (such as advertisements or links to unrelated news stories) and to join together broken sentences.

(3) **Machine translation** (MT) of the source text into English maybe necessary at this stage if the BioCaster does not have a native fact extraction capability in the source language. Currently we use MOSES, an open source MT system for this purpose.

(4) **Text classification** is applied to group texts into topical categories for either trashing - in the case of documents clearly outside the task definition - or subsequent processing using detailed fact extraction [24].

(5) **Named entity recognition and grounding** are applied using a rule based regular expression language called Simple Rule Language (SRL) which we developed [25].

(6) **Event extraction** is used to obtain structured information about an event such as the name of the condition, the type of agent, the number of victims and time and location where the event happened. i.e. the who, what, where, when and how of the event.

(7) **Evidence assessment** applies a number of checks to remove vague or redundant events. This includes: (a) documents already processed at the same URL, (b) events with no disease, country or species mentioned, (c) events which are clearly historical. Such reports are dropped.

(8) **Event alerting** using statistical aberration detection algorithms [18], [19] is used to obtain significance scores for the rarity of the events using historical geo-temporal information. Such algorithms are widely used within the public health community, e.g. [26].

(9) **Human judgement** is applied by BioCaster users and partners. Human analysis is almost always needed to understand what is abnormal, to discovery rare events that the system may have missed, to make the final decision about vague reports and to link together disparate events. The previous automated stages aims to make human judgments quicker, cheaper and more reliable through data search and visualization on the database of mined facts.

As can be see from the previous description, the BioCaster system comprises a modularised text mining pipeline. The pipeline runs on a dedicated cluster computer linked to the frontend Web server. The modules consist of efficient natural language processing algorithms for classifying documents into relevant or non-relevant as well as dedicated modules for identifying terms and their relationships. Various modules are integrated with a sophisticated knowledge model of the domain defining semantic categories for diseases, species, symptoms, agents etc. and the relationships between them. These relationships are assembled automatically into an event report comprising a slot filler template with a minimum fill of a country, province, disease, species and time element.

The SRL language that BioCaster uses for entity and event recognition is designed to allow users without a background in computer science to quickly build up rule books. Although in general this is laborious we have tried to make the task easier by developing a freely available graphical user interface. SRL has been motivated by earlier pattern based languages such as DIAL (Declarative Information Analysis Language) [1] and incorporates a capability to match to string literals, named entity classes, skipwords as well as word lists. The general SRL syntax is a label followed by a head expression and a body expression. The head expression is output if the regular expression in the body matches to the text. For example

D1: :- name(disease) { list(% disease) }

Rule D1 matches to any phrase in the list ́disease ́and outputs a named entity of type DISEASE.

Our English SRL rule book for biohazard events incorporates approximately 110 rules for entity detection, 12 major word lists containing 870 terms and over 2800 template rules for detecting direct signals such as international travel, zoonosis, category A agents, novel diseases and malformed blood products.

3.5 Data Sources

The abundant and near-real time nature of online news, makes it a cost-effective means of early detection and tracking of health events on a global scale. Open media sources bridge the gap between national and international surveillance as well as providing timely access to sources on the ground. BioCaster surveillances approximately 30,000 news items per day from Google News as well as various public and NPO sources such as the ProMed-mail, Hong Kong SAR Communicable Disease Watch list, the OIE alert lists, the European Media Monitor alerts and AlertNet. News is gathered on a 30 minute cycle which can be shortened as necessary during health emergencies.

3.6 Availability

BioCaster provides outputs to users in a variety of freely available and restricted formats. The open access public portal has various interfaces such as a 30 day Google map of events, customisable graphs etc. The map, called the Global Health Monitor, can be searched according to time, location, disease and special functions show alerted events for bio/chemical/radnuke as well as natural disasters such as earthquakes and typhoons. A searchable database of events is also provided so that other researchers can access aggregated data for their own analysis. The database is freely available for users to view and download data 24/7. Updates to the database take place once every thirty minutes during normal operation but this can be shortened as required during public health emergencies.

A login restricted alerting interface is used by a small test community. Here users can define targeted rules that let them receive email alerts whenever news comes into the system on topics of interest. For example a user could define a rule asking to be alerted on the topic of novel influenza in Europe involving a case of international travel. More advanced news search and analytics are also incorporated into the login site allowing users access to aggregated data on events.

4. Conclusion

In this paper I have briefly described the history and function of the BioCaster disease surveillance system, one of a range of systems in use by the international public health community for surveillancing the world's media. Based on advanced text mining technology we have been providing a freely available service to the global public health community since 2006.

Future advances which we are now working on include: (a) extending coverage to new languages and health threats, (b) working with other system developers and user groups to enhance data sharing and interchange standards, (c) incorporating signals into the alerting algorithms from across media types such as social networking data. This third advance brings significant data handling challenges as we adapt the system for the age of 'big data'.

Acknowledgements

I would like to thank the following postdocs and collaborative researchers who have worked on the BioCaster project. Son Doan, Mike Conway, Reiko Goodwin, Ai Kawazoe, John McCrae, Hutchatai Chanlekha, Dinh Dien, Mika Shigematsu, Koichi Takeuchi, Kiyosu Taniguchi and colleagues at the Global Health Initiative's EAR project. Funding was provided by the Japan Science and Technology Agency (JST) Sakigake fund between 2008 and 2012.

References

[1] R. Feldman and J. Sanger. *The Text Mining Handbook: Advanced Approaches in Analysing Unstructured Data*. Cambridge University Press, Cambridge, 2006.

[2] T. W. Grein, K. B. Kamara, G. Rodier, A. J. Plant, P. Bovier, M. J. Ryan, T. Ohyama, and D. L. Heymann. Rumours of disease in the global village: outbreak verification. *Emerging Infectious Diseases*, 6:97–102, 2000.

[3] D. Hartley, N. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. Brownstein, G. Thinus, and N. Lightfoot. The landscape of international biosurveillance. *Emerging Health Threats J.*, 3(e3), January 2010. doi:10.1093/bioinformatics/btn534.

[4] A. Mawudeku and M. Blench. Global public health intelligence network (gphin). In *Proc. 7th Int. Conf. of the Association for Machine Translation in the Americas, Cambridge, MA, USA*, pages 7–11, August 8–12 2006.

[5] Lawrence C. Madoff and John P. Woodall. The internet and the global monitoring of emerging diseases: Lessons from the first 10 years of promed-mail. *Archives of Medical Research*, 36(6):724 – 730, 2005. Infectious Diseases: Revisiting Past Problems and Addressing Future Challenges.

[6] R. Yangarber, P. von Etter, and R. Steinberger. Content collection and analysis in the domain of epidemiology. In *Proc. Int. Workshop on Describing Medical Web Resources (DRMED 2008), Gotenburg, Sweden*, May 27th 2008.

[7] C. Freifeld, K. Mandl, B. Reis, and J. Brownstein. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J. American Medical Informatics Association*, 15:150–157, 2008.

[8] N. Collier, S. Doan, A. Kawazoe, R. Matsuda Goodwin, M. Conway, Y. Tateno, Q. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi. BioCaster:detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–1, December 2008. doi:10.1093/bioinformatics/btn534.

[9] S. Fuller. Tracking the global express: new tools addressing disease threats across the world. *Epidemiolgy*, 21(6):769–771, 2010.

[10] H. Tolentino, R. Kamadjeu, P. Fontelo, F. Liu, M. Matters, M. Pollack, and L. Madoff. Scanning the emerging infectious disease horizon - visualizing promed emails using epispider. *Advances in Disease Surveillance*, 2(169), 2007.

[11] J. Zamite, F. Silva, F. Couto, and M. Silva. Transactions on large-scale data- and knowledge-centered systems iv, lecture notes in computer science. In *MEDCollector: multisource epidemic data collector*, volume 6990, pages 40–72. Springer, 2011.

[12] R. Grishman, S. Huttunen, and R. Yangarber. Information extraction for enhanced access to disease outbreak reports. *Journal of Biomedical Informatics*, 35:236–246, 2002.

[13] L Damianos, J. Ponte, S. Wohlever, F. Reeder, D. Day, G. Wilson, and L. Hirschman. MiTAP, text and audio processing for bio-security: A case study. In *Proceedings of the Fourteenth Innovative Applications of Artificial Intelligence Conference (IAAI-2002), Alberta, Canada*, July 28th – August 1st 2002.

[14] N. Collier, A. Kawazoe, L. Jin, M. Shigematsu, D. Dien, R. Barrero, K. Takeuchi, and A. Kawtrakul. A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation*, 40(3–4), 2006. DOI: 10.1007/s10579-007-9019-7.

[15] N. Guarino and C. Welty. A formal ontology of properties. In R. Dieng and O. Corby, editors, *EKAW-2000: Proc. 12th Int. Conf. on Knowledge Engineering and Knowledge Management*, pages 97–112, 2000.

[16] A. Kawazoe, L. Jin, M. Shigematsu, R. Barerro, K. Taniguchi, and N. Collier. The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system. In *KR-MED 2006: Proc. Int. Workshop on Biomedical Ontology in Action, Baltimore, USA*, pages 77–85, November 8th 2006.

[17] N. Collier, R. Matsuda Goodwin, J. McCrae, S. Doan, A. Kawazoe, M. Conway, A. Kawtrakul, K. Takeuchi, and D. Dien. An ontology-driven system for detecting global health events. In *23rd International Conference on Computational Linguistics (COLING), Beijing, China*, pages 215–222, August 23–27 2010.

[18] N. Collier. What's unusual in online disease outbreak news? *Biomedical Semantics*, 1(1), March 2010. doi:10.1186/2041-1480-1-2.

[19] N. Collier. Towards cross-lingual alerting for bursty epidemic events. *Biomedical Semantics*, 2(Suppl 5):S9, September 2011.

[20] N. Collier, S. T. Nguyen, and M.T.N. Nguyen. OMG U got flu? analysis of shared health messages for bio-surveillance. *Biomedical Semantics*, 2(Suppl 5):S10, September 2011.

[21] O. Lawrence and J. Gostin. International infectious disease law - revision of the World Health Organization's international health regulations. *J. American Medical Informatics Association*, 291(21):2623–2627, 2004.

[22] I. Niles and A. Pease. Towards a standard upper ontology. In C. Welty and B. Smith, editors, *2nd Int. Conf. on Formal Ontology in Information Systems FOIS-2001, Maine, USA*, October 17–19 2001.

[23] P. Vossen. Introduction to EuroWordNet. *Computers and the Humanities*, 32:73–89, 1998.

[24] M. Conway, S. Doan, and N. Collier. Classifying disease outbreak reports using n-grams and semantic features. *International Journal of Medical Informatics*, 78(12):e47–e58, 2000.

[25] J. McCrae, M. Conway, and N. Collier. Simple rule language editor. Google code project, September 2009. Available from: http://code.google.com/p/srl-editor/.

[26] J. I. Tokars, H. Burkom, J. Xing, R. English, S. Bloom, K. Cox, and J. Pavlin. Enhancing time-series detection algorithms for automated biosurveillance. *Emerging Infectious Diseases*, 15(4):533–539, 2009.