Near Real Time Public Health Protection with DIZIE and BioCaster

Nigel COLLIER, Son DOAN, Bao-Khanh Ho VO

http://born.nii.ac.jp

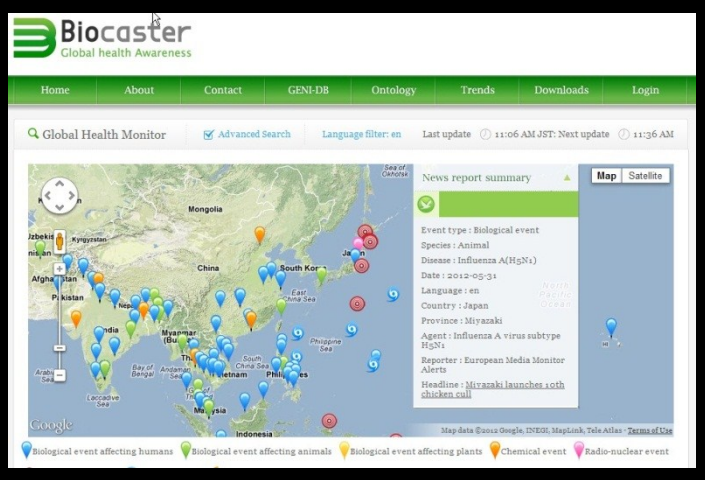
どんな研究？

SARSやトリインフルエンザのような感染症の発生を早期に発見し、監視・追跡するには、様々な言語で書かれたWeb上のローカルニュースを、各国の政府が責任を持ってモニターする必要がある。BioCasterプロジェクトでは、最新のテキストマイニング技術を応用して多言語のニュース記事をフィルタリングし、構造化された形式で現地語に翻訳するWebポータルを開発する。特に、(1) 多言語知識リソース（オントロジー）、(2) 高性能クラスタコンピュータおよびストレージシステム、(3) 感染症に関するニュース記事と、研究文献や遺伝子データベースにある最新の研究成果をナビゲートする、知的なリンケージシステム等の構築に焦点を当てる。

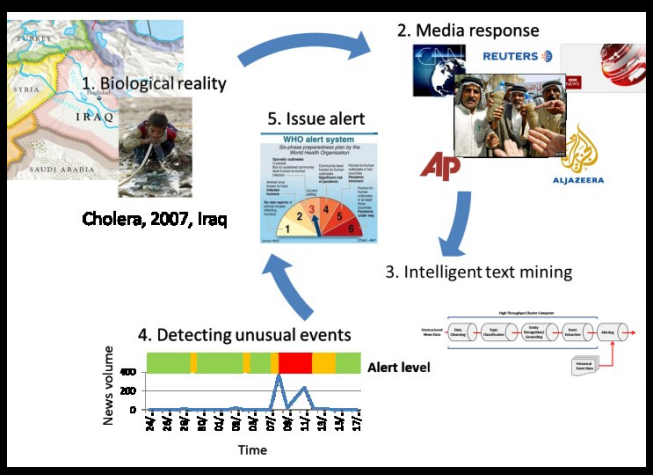
What kind of research?

Early detection and tracking of a possible disease outbreak such as SARS or Avian influenza is a responsibility for governments who are faced with monitoring massive quantities of local news on the WWW in several languages. In BioCaster we are developing a web-portal using the latest text mining technology that can filter news reports in various regional languages and present a summarized translation in the local language. Research is focusing on creating: (1) a multi-lingual knowledge resource (ontology), (2) a high-performance text mining system, (3) an intelligent linkage system for navigating between news about diseases and the latest research findings in the literature and genetics databases.

What is global media monitoring?



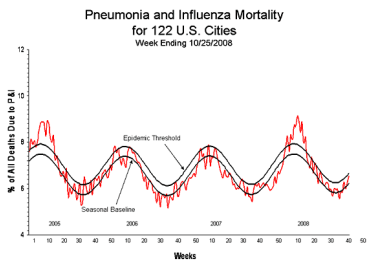
Global media monitoring aims to rapidly detect public health hazards such as disease outbreaks and chemical spillages from open Web based sources such as news and social media sites like Twitter. BioCaster is active in the detection of health threats to humans, animals and plants.



What are the challenges?

Detecting the unusual…

Ebola in Uganda? Salmonellosis in Dublin? How does a computer know when an event is unusual? Our research analyses and evaluates a range of time series analysis algorithms for statistical alerting of event counts.



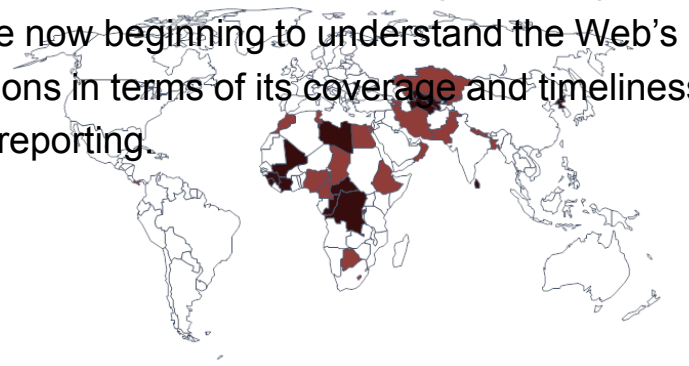
CDC data showing epidemic alerting thresholds for influenza

Understanding ambiguity…

Writers have many different ways of reporting the same health condition such as *influenza, flu, H5N1, bird flu* etc. Reports in multiple languages represents an opportunity but also increase the challenge. A key research result has been the production of a sophisticated ontology for unifying different ways of reporting health conditions.

Understanding the Web's limitations…

News sources vary in trustworthiness and different regions of the world have differing reporting patterns. We are now beginning to understand the Web's limitations in terms of its coverage and timeliness of event reporting.



Combing information sources…

News reports, blogs, search queries … How do we combine signals across media that may differ in temporal and spatial granularity as well as reporting rates and population characteristics? How do we validate them against a gold standard? We have just begun to explore this question in our work within the Grand Challenge funded project DIZIE.

Recent Press Reports about BioCaster

Key Reference

1. Collier, N. et al. (2008) "BioCaster: detecting public health rumors with a Web-based text mining system", BioInformatics, 24(24): 2940-2941, Oxford University Press.

NII
連絡先：Nigel COLLIER ／ 国立情報学研究所 Principles of Informatics, Associate Professor
TEL : 03-4212-2536 FAX : 03-3556-1916 Email : collier@nii.ac.jp

Near Real Time Public Health Protection with DIZIE and BioCaster

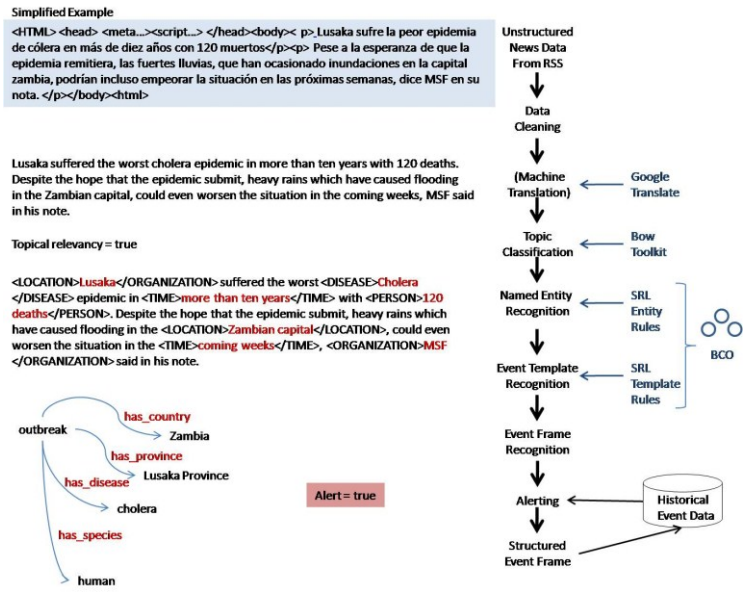
Nigel COLLIER, Son DOAN, Bao-Khanh Ho VO

http://born.nii.ac.jp

What core technologies do you use?

In BioCaster we are exploring a range advanced algorithms for intelligent text processing over very large data sets using optimized feature selection. Key tasks include text classification, terminology recognition, event extraction, event alerting and visualization. Underlying the whole system is a multi-lingual ontology – the BioCaster Ontology or BCO. The BCO is freely available to download and contains a wealth of structured terminology in many languages related to infectious diseases.

BioCaster's Semantic analysis pipeline



How are you using social media?

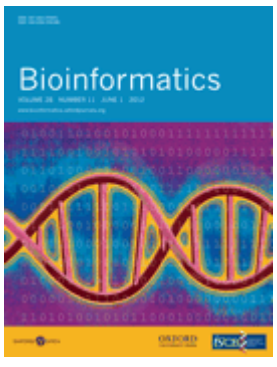
Recent studies by ourselves and others have shown a strong correlation between social networking messages and national influenza rates. In the DIZIE project we have expanded on this to develop an automated text mining system that classifies and aggregates Twitter messages in real time. Messages are classified according to six types of diseases: respiratory, gastrointestinal, neurological, rash, constitutional and hemorrhagic. Results are shown on a novel radial interface for 40 major world cities.

Key Reference

3. Collier, N. and Doan, S. (2012) "Syndromic classification of Twitter messages", Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 91. pp. 186-195. Springer.

How can I use BioCaster data in my own research?

This year for the first time we launched a new database of public health events called GENI-DB. Here you can find downloadable data for over 176 infectious diseases and chemicals affecting human and animal health. The data is an aggregated summary of reports in the world's news media in 10 languages.



Key Reference

2. Collier, N. and Doan, S. (2012) "GENI-DB: a database of global events for epidemic intelligence", BioInformatics, 28(8): 1186-8, Oxford University Press.

Who are you working with?

Partnership is central to our goal in improving health and safety and making sure that our results are accurate and useful. We are working with a number of international public health organizations including: the World Health Organisation, the European Centres for Disease Control, the US Centers for Disease Control, the Ministry of Health in Japan, the Health Protection Agency in the UK, the European Commission's DG SANCO and Public Health Canada. Technology partners include: Kasetsart University (Thailand), Viet Nam National University and Okayama University.

From 2009 to 2012 BioCaster was supported by grant-in-aid from the JST Sakigake fund. DIZIE is supported by grant-in-aid from NII's Grand Challenge Project fund.

NII 連絡先: Nigel COLLIER ／ 国立情報学研究所 Principles of Informatics, Associate Professor
TEL : 03-4212-2536 FAX : 03-3556-1916 Email : collier@nii.ac.jp