BIOCASTER：多言語ニューステキストを利用した感染症の早期発見および監視

BIOCASTER: Detection and Tracking of Disease Outbreaks from Multilingual News Texts

Nigel Collier, Ai Kawazoe, Son Doan, Mike Conway, Reiko Matsuda Goodwin, John McCrae, Dinh Dien, Koichi Takeuchi, Asanee Kawtrakul

要旨

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。

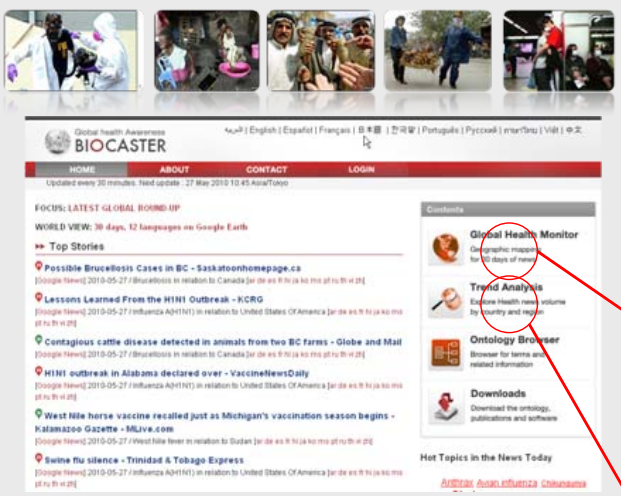
Summary

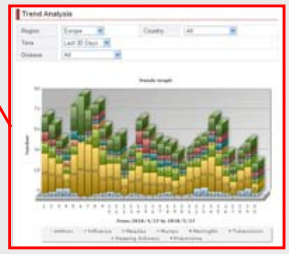
Early detection and tracking of a possible disease outbreak such as SARS or Avian influenza is a responsibility for governments who are faced with monitoring massive quantities of local news on the WWW in several languages. In BioCaster we are developing a web-portal using the latest text mining technology 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.

Purpose

- ■Enable timely access to disease outbreak news to raise government and public health expert's awareness
- ■ Access to multi-lingual news reports on the Internet using text mining technology
- ■ Integration of bio-geographic information to aid in the analysis of disease spread
- ■ Automatic email alerts to registered users for news items on key topics of interest
- ■ Linkage of rich information sources to help users decide on the significance of the outbreak

System Features





- ■ News in 12 languages
- ■ Bio-geographic maps
- ■ Graphical trend analysis
- ■ Automated alerting
- ■ Open source ontology
- ■ Open source text mining system
- ■ Regularly used by national and international agencies

Find out more at: http://biocaster.nii.ac.jp

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

BIOCASTER：多言語ニューステキストを利用した感染症の早期発見および監視

BIOCASTER: Detection and Tracking of Disease Outbreaks from Multilingual News Texts

Nigel Collier, Ai Kawazoe, Son Doan, Mike Conway, Reiko Matsuda Goodwin, John McCrae, Dinh Dien, Koichi Takeuchi, Asanee Kawtrakul

Themes

The BioCaster project encompasses a range of research themes related to the real time semantic analysis of textual content. Key processes include text classification, named entity 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.

References for Ontology Engineering:

1.Collier, N., MatsudaGoodwin, R., McCrae, J., Doan, S., Kawazoe, A., Conway, M., Kawtrakul, A., Takeuchi, K. and Dien, D. (2010), "An ontology-driven system for detecting global health events", Proc. 23rd International Conference on Computational Linguistics (COLING), Beijing, China, August 23-27, (accepted to appear).

2.Kawazoe, A., Chanlekha, H., Shigematsu, M. and Collier, N. (2008), "Structuring an event ontology for disease outbreak detection", *BMC Bioinformatics, 9 (Suppl 3): S8, DOI: 10.1186/1471-2105-9-S3-S8.*

3.McCrae, J. and Collier, N. (2008), "Synonym set extraction from the biomedical literature by lexical discovery", in BMC Bioinformatics, 9:159, DOI: 10.1186/1471-2105-9-159.

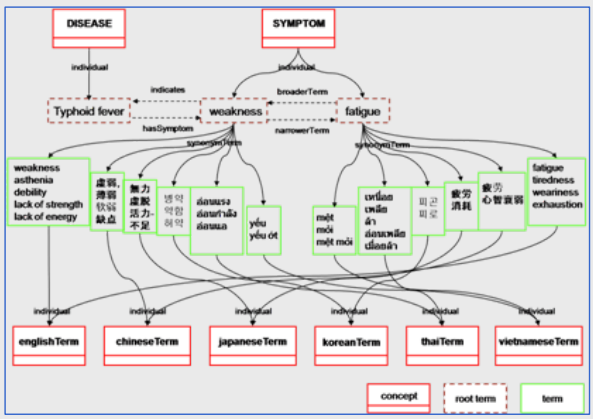


Fig 2. Overview of the BioCaster Ontology

Fig 3. Global health mapping

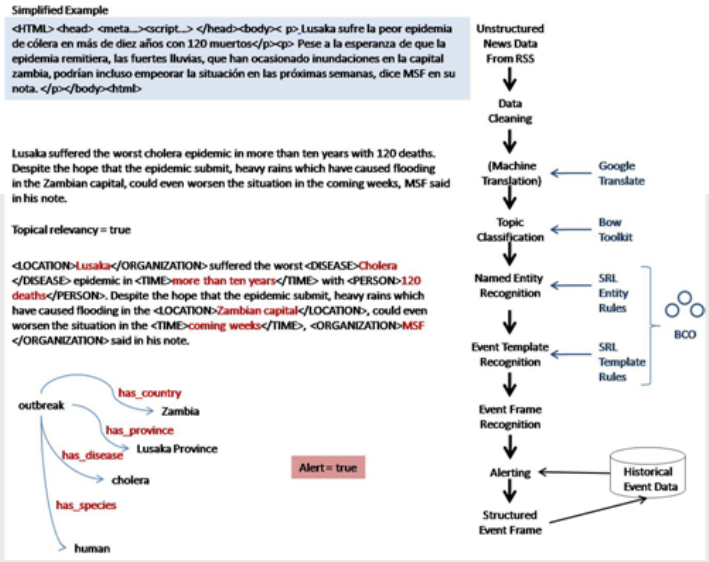


Fig 1. Semantic analysis pipeline

References for Text Classification:

1.Conway, M., Kawazoe, A., Chanlekha, H. and Collier, N. (2010), "Developing a disease outbreak corpus", *Journal of Medical Internet Research (in press).*

2.Doan, S., Conway, M. and Collier, N. "An Empirical Study of Sections in Classifying Disease Outbreak Reports", invited chapter in Annals of Information Systems, Special Issue "Web-based Applications in Health Care & Biomedicine", Springer, 2009.

Reference for Automated Alerting:

1.Collier, N. (2010), "What's unusual in online disease outbreak news?" *Journal of Biomedical Semantics, 1:2, DOI:10.1186/2041-1480-1-2.*

References for Geo-Temporal Tagging:

1.Chanlekha, H. and Collier, N. (2010), "A methodology to enhance spatial understanding of disease outbreak events reported in news articles", International Journal of Medical Informatics (in press).

2.Chanlekha, H. and Collier, N. (2010), "Analysis of syntactic and semantic features for fine-grained event-spatial understanding in outbreak news reports", *Journal of Biomedical Semantics, 1:3, DOI: 10.1186/2041-1480-1-3.*

3.Chanlekha, H. and Collier, N. (2010), "A framework for enhanced spatial and temporal granularity in report-based health surveillance systems", *Journal of Medical Informatics and Decision Making, 10(1).*

Partners

Partnership is central to our goal in improving health and safety. We are working with collaborators at: Okayama University Japan, National Institute of Infectious Diseases Japan, National Institute of Genetics Japan, Kasetsart University Thailand, Viet Nam National University Ho Chi Minh City, Vanderbilt University Medical Center, Pittsburgh University USA, University of Bielefeld Germany, Fordham University USA, and international public health organizations In Europe, and North America.

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