OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Revitalizing the Global Public Health Intelligence Network (GPHIN)

Dave Carter*, Marta Stojanovic and Berry de Bruijn

National Research Council Canada, Ottawa, ON, Canada

Objective

To rebuild the software that underpins the Global Public Health Intelligence Network using modern natural language processing techniques to support recent and future improvements in situational awareness capability.

Introduction

The Global Public Health Intelligence Network is a non-traditional all-hazards multilingual surveillance system introduced in 1997 by the Government of Canada in collaboration with the World Health Organization.[1] GPHIN software collects news articles, media releases, and incident reports and analyzes them for information about communicable diseases, natural disasters, product recalls, radiological events and other public health crises. Since 2016, the Public Health Agency of Canada (PHAC) and National Research Council Canada (NRC) have collaborated to replace GPHIN with a modular platform that incorporates modern natural language processing techniques to support more ambitious situational awareness goals.

Methods

The updated GPHIN platform assembles several natural language processing tools to annotate incoming data in order to support situational awareness; broadly, GPHIN aims to extract knowledge from data.

Data are collected in 10 languages and are machine translated to English. Several of the machine translation models use high performance neural networks. Language models are updated regularly and support external dictionaries for handling emerging domain-specific terms that might not yet exist in the parallel corpora used to train the models.

All incoming documents are assigned a relevance score. Machine learning models estimate a score based on similarity to sets of known high-relevance and known low-relevance documents. PHAC analysts provide feedback on the scoring from time to time in the course of their work, and this feedback is used to periodically retrain scoring models.

Documents are assigned keywords using two ontologies: an all-hazards multilingual taxonomy hand-compiled at PHAC, and the U.S. National Library of Medicine's Unified Medical Language System (UMLS).

Categories are assigned probabilistically to incoming articles (e.g., human infectious diseases, animal infectious diseases, substance abuse, environmental hazards), largely using affinity scores that correspond to keywords.

Dates and times are resolved to canonical forms, so that mentions like *last Tuesday* get resolved to specific dates; this makes it possible to sequence data about a single event that are released at varying frequencies and with varying timeliness.

Cities, states/provinces, and countries are identified in the documents, and gaps in the hierarchical geographic relationships are filled in. Locations are disambiguated based on collocations; the system distinguishes between and correctly resolves Ottawa, KS vs. Ottawa, ON, Canada, for example. Countries are displayed with their socio-economic population statistics (Gini coefficient, human development index, median age, and so on).

The system attempts to detect and reconcile near-duplicate articles in order to handle instances where one article is released on a newswire and subsequently gets lightly edited and syndicated in dozens or hundreds of local papers; this improves the signal-to-noise ratio of the data in GPHIN for better productivity. Template-based reports (where the same document may get re-issued with a new number of cases but no other changes, for example) are still a challenge, but whitelisting tools reduce the false positive rate.

The system provides tools for constructing arbitrarily detailed searches, tied to colour-coded maps and graphs that update on-the-fly, and offers short extractive summaries of each search result for easy filtering. GPHIN also generates topical knowledge graphs about sets of articles that seek to reveal surprising correlations in the data; for example, graphically reconciling and highlighting cases where several neighbouring countries all have reports of a mysterious disease and where a particular mosquito is mentioned.

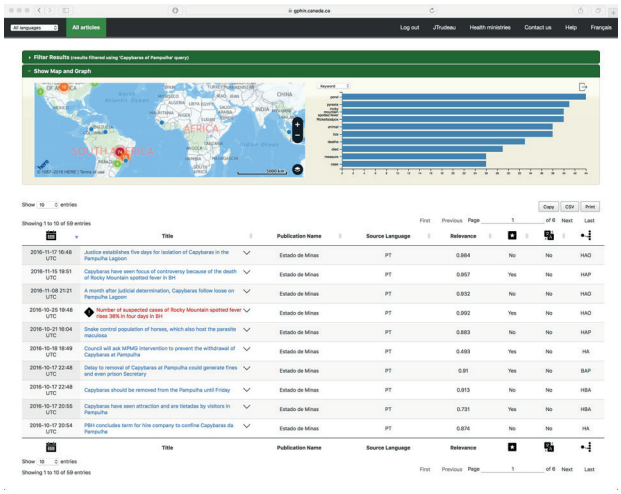
Next steps in the ongoing rejuvenation involve collating discrete articles and documents into narrative timelines that track an ongoing event: collecting all data about the spread of an infectious disease outbreak or perhaps the aftermath of an earthquake in the developing world. Our research is focussing on how to build line lists from such a stream of news articles about an event and how to detect important change points in the ongoing narrative.

Results

The new GPHIN platform was launched in August 2016 in order to support syndromic surveillance activities for the Rio 2016 Olympics, and has been updated incrementally since then to offer further capabilities to professional users in 30 countries. Its modular construction supports current situational awareness activities as well as further research into advanced natural language processing techniques.

Conclusions

We improved (and continue to improve) GPHIN with modern natural language processing techniques, including better translations, relevance scoring, categorization, near-duplicate detection, and improved data visualization tools, all towards the goal of more productive and more trustworthy situational awareness.



GPHIN search interface with map, configurable graph, and a recent alert.

Keywords
natural language processing; software; data annotation

References

1. Mawudeku, A, Blench, M. Global public health intelligence network (GPHIN). 7th Conference of the Association for Machine Translation in the Americas. 2006.

***Dave Carter**
E-mail: david.carter@cnrc-nrc.gc.ca