研 究 報 告 書

「健康被害を監視するための多言語ウェブサーベイランスシステム」

研究期間： 平成 20 年 10 月～平成24年3月
研 究 者： 国立情報学研究所・情報学プリンシプル研究系・准教授ナイジェル コリアー

1，研究のねらい

マスコミなどの公的でないデジタルソースによる、疾患の集団発生に関する情報の収集を行うシステムは今や国内ならびに海外の公衆衛生機関において、重要視されるものとなっている。富裕国では、市販や一般開業医のネットワークといった高度な情報源に溢れているとはいえ、すべての国々にこういったシステムを実施、あるいは維持するソースがあるわけではない。A(H5N1)といった新興疾患の急激なまん延への懸念により、流行性疾患情報システムへの関心が高まってきている。そのシステムは、世界規模での事象を検出し、指標ネットワークを補い、ソースに密接した状態での疾患の対処を可能にするものである。BioCaster BORN(生化学・放射性物質・核)と呼ばれる、この研究プロジェクトの最終目的は、広範囲の感染疾患、ならびに化学、放射性、核による病原物質を早い段階で警戒するため、テキストマイニング技術を基にした、完全稼動状態のウェブ上知的サーベイランスシステムを開発することである。この研究は公衆および動物衛生コミュニティーの研究者らによる緊密な協力のもと、世界規模での健康の向上のため、調査結果の利益が、できる限り広範囲に広がるよう実施されたものである。

2，研究成果

この研究は、いくつか重要な点に関して、我々のこれまでの知識を拡大するものである。(1)歴史的背景に反しての、発生事象の重要性を理解できるよう変化点検出の調査を行った。(2) 化学、放射性、核による食物、水および環境の汚染によって起こる疾患の発生といった広い範囲での健康への脅威に対する知識モデルを製作した。知識モデルには 12 言語での専門用語が含まれる。(3)地理的および時間的類似点のほか、個々の公衆衛生事象の関連性に関する知識を用い、健康への脅威に関する情報を融合した調査を実施した。このワークパッケージに関しては、個別に述べることとする。

この研究を経ての最終的な BioCaster BORN システムは、公共のウェブサーバ(http://born.nii.ac.jp)に組み込む専用の高性能コンピューティング・クラスター上で作働する、モジュール化したテキストマイニングパイプライン(図 1)で構成されている。 システム内のモジュールは言語探知、機械翻訳、文献分類に効率的な自然言語処理アルゴリズムのほか、用語とその関連性を特定し、事象を特定した場合に、そういった事象の危険を世界中の公衆衛生コミュニティーに警告する専用のモジュール類で構成されている。こういったさまざまなモジュールは、疾患、動物種、症状、病原体などに対して語彙の分類および個々の関連性を定めるドメインの高度な知識モデルと統合されているのである。

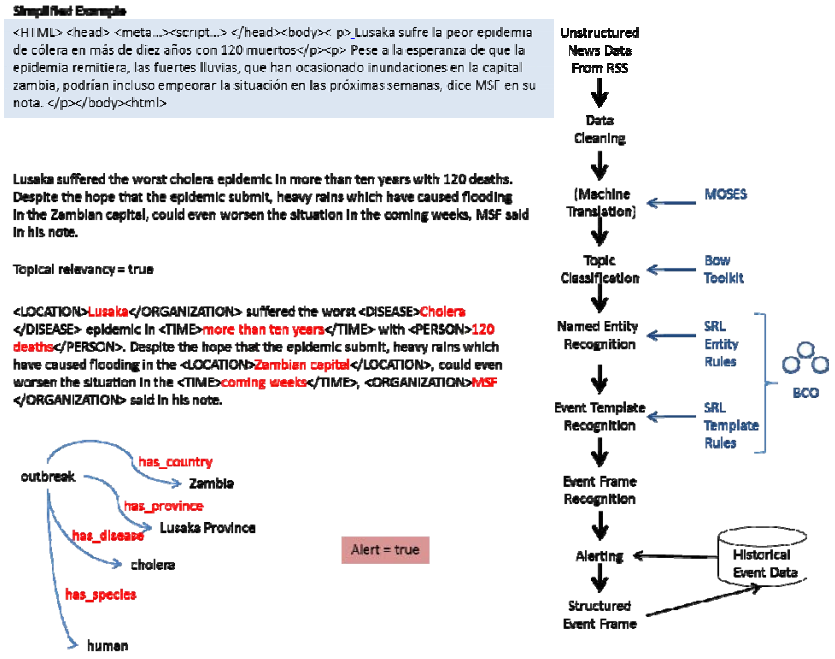

図 1 BioCaster BORN テキストマイニングパイプライン

Google ニュースのほか、ProMed メール、香港特別行政区伝染病注意項目リスト、国際獣疫局警戒項目リスト、ヨーロッパメディアモニター警戒事項リスト、国際獣疫機構(OIE)警戒リスト、 ヨーロッパメディア監視警告事象やアラートネットといった公共または NPO ソースから、通常 1 日約 27,000 のニュース項目が BioCaster BORN によって監理されている。さらに、プロジェクト期間中(2008.10～2012.3)、我々は、eltWater という民間のニュースアグリゲーション会社から新たに 80,000 のニュースソースの使用許諾を受けた。

［変化点検出を用いたニュースの事象警報］

このワークパッケージでは、私は、オンライン衛生関連ニュースの事象を用いて、日々の警報に関して変化点検出アルゴリズムを序列的に評価するうえでの問題点に取り組んだ。毎日の各国の疾患数は BioCaster BORN を用いて、実際の世界中のデータをテキストマイニングした。18 の BioCaster BORN による英語のニュースデータ 18 セットに対して、公衆衛生ドメインで広く用いられる 5 種の異常検出アルゴリズム(EARS C2、C3、W2、F 統計量、EWMA)の、専門家が管理する ProMED メール投稿のシルバースタンダードに対する性能を比較した。ProMED メールは、国際感染症学会の公的プログラムとして、専門家ボランティアが世界規模で、メディア報道、その他のソースを、動物、植物に影響を及ぼす生物的ならびに科学的災害に関する情報についてモニタリングを行っている、毎日 24 時間稼動するボランティアのヒューマンネットワークである。

14 ヶ国における 366 日以上の期間での 18 の発症例に関する 287 件の ProMED メール投稿に関して、システムの感受性、特異性、陽性的中率(PPV)、陰性的中率(NPV)、100 日間の平均警報、および F1 は 95%の信頼区間(CI)で報告された。 F1 に重点を置き、誤警報率および 100 日あたりの平均警報数を公衆衛生分析における最重要基準とした。結果は、W2 に最良の F1 が認められ、C2 に比べてわずかに週の曜日効果を示した。これは、おそらく W2 によって用いた層別サンプリング法によって週の曜日効果が相殺されたためであると考えられた。ドリルダウン分析で

は、国の水準モデルのグラニュラ選択によって起こる問題点ならびに週の曜日効果および報道の偏りによる急激なレポート量の減少について示唆した。選択,その他の研究報告と同様に、私の研究結果は、最終的な検出能力に悪影響を及ぼす事象 1 例に対するニュースの報道が長期にわたる傾向にあることを示すものであった。ニュース報道量に急激な増加がみられた例として、イラクでのコレラ、イギリスでの麻疹があるが、こういった例は、段階的にニュースが増加した米国での A(H1N1)インフルエンザのような事象と比較すると、容易に警戒態勢を取れる傾向にあった。この結果を基に、私は、多言語でのニュースレポートを用いて生じる利益について検討した。一連の新たな実験において、5 種の同じ時間的異常検出アルゴリズムを用い、16 件の疾患発生例の進化を追跡した。さらに、ProMED の報告例 をシルバースタンダードとして用いたところ、129 日以上にわたる試験期間での 13 言語に対する新たなデータの比較分析の結果、さまざまな言語での事象を用いたモデルの多くに、感受性ならびに適時性の向上が認められた。
この結果は、多言語テキストマイニングを用いた自動健康サーベイランスには、インフォームド・チョイスを用いてモデル選択、データソースの管理を行う場合、低価値の情報を警戒事象に変換する可能性があることを示すものであった。 BioCaster BORN ポータル上では多言語でのニュースを用いた C2 警報アルゴリズムの実行が可能となった。

［公衆衛生脅威モデリング］

もうひとつのワークパッケージにおける重要成果は BioCaster オントロジー (BCO)である。これは、疾患報告における一般人向けの言語統一のために考案され無料で提供された公衆衛生アプリケーションの 12 言語のオントロジーである。 (http://code.google.com/p/biocaster-ontology)

BCO の目的

- ニュースでの公衆衛生事象の検出およびリスク評価に必要な、用語ならびにその関連性の説明
- （多言語での） 生体医学上のグレー・リテラチャーと従来の基準のギャップを補充
- 言語間の内容統一を仲介
- 無料提供

BioCaster BORN にとって BCO は主要な知識ソースであり、疾患、薬品、症状、症候群、動物種といったドメイン用語のほかに諸症状を引き起こす疾患、または特定の宿主動物種に作用する薬品といったドメイン感受性の関連事項を含んでおり、これによってテキストマイニングシステムがドメイン内での主要概念や関連性を認識し、ニュースでは明確にされない箇所を補うことができるのである。BCO は、公開メディアの言語での疾患発生サーベイランスに関心のあるシステム開発者に対して、多言語でのサポートを無料で提供しており、我々が知る限り、アプリケーションオントロジーとしてはユニークなものである。
BCO は現在、ヒト、動物の 300 種以上の疾患を、国連が公用語とする 12 ヶ国言語（アラビア語(968 語)、英語(4113 語)、フランス語(1281 語)、インドネシア語(1081 語)、日本語(2077 語)、韓国語(1176 語)、マレーシア語 (1001 語)、ロシア語(1187 語)、スペイン語(1171 語)、タイ語(1485 語)、ベトナム語(1297 語)、中国語(1142 語)）で網羅している。 BCO は、ウェブ上のオントロジー言語(OWL)ならびに単一知識機構システム(SKOS)フォーマットにて、ダウンロードによる入手が可能である。 これは BCO のウェブサイトで無料提供されており 現在までで、各国 350 以上のグループがオントロジーのダウンロードを行っている。

[地理−時間的理解による事象の融合]

このプロジェクトで始めた最大の難課題のひとつがコンピュータに、状況に関してさらなる知識を与えるという試みである。最初のワークパッケージでの作業は、一文章で報告された事実の把握に基づくものであった。さらなる難題として、ある状況の中で、事実が意味を成さなければならない。この問題に取り組むため、私は共同研究者らの協力を得て、解決策となりうる事象オントロジーならびに地理−時間的アノテーション・スキーマを作成した。'入院する'、'検疫'、'治療を受ける' といった事象が疾患の発生における中心的な概念とはいえ、疾患、症状、発生場所といった観察対象物と比較すると、その理解はより困難である。DOLCE（言語と認知工学のための記述的オントロジー）から適用した方法を用いて、BioCaster 事象オントロジー(BCEO)が感染疾患の公衆衛生ドメインで、事象の正式な定義ならびに疾患関連事象を示す表現を提供する。事象オントロジーの初期のバージョンでは、40 例事象の種類に対して、同義語ならびに事象論理上の正式な記述が含まれたものが公開された。

地理−時間の認識は、コンピュータが正確なニュースレポート処理を行うために理解が必須である、また別の複雑な課題である。私は共同研究者とともにアノテーション・スキーマ、ならびに事象発生の場所と期間をニュース文書内より特定するメソッドを作成した。その技法によって、言語の特性による分類に基づき、特殊事象を一般的あるいは仮説上の事象から分類するとともに事象の場所を詳細に特定する手段が与えられた。自動的に対象物およびその関連性(同一性、重複、原因など)に関する記述をニューステキストより検出、統合する機械学習ツールのトレーニング用に、BCEO および地理−時間的標識法に基づき、大規模な、事象に関する注釈つきコーパスの構成を開始した。

3, 今後の展開

3 種のワークパッケージにおける研究から得た見識を基に、今後私は以下の疑問点に答えていこうと考えている。 (a)取り込んだ多変量の事象特性に対する警報の利点は何か。(b)極めて稀な事象に対する警報方法の開発はどのように行えばよいか。(c)'未分類インフルエンザ'のような一般疾患に関する報告などの不特定報告への警報のサポートにはどういった意味的特性が警報最良であるか。(d)発生場所の自動検出アルゴリズムの質はどのように向上させるのか。近年行われたソーシャルメディア分析での調査結果を踏まえると、事象の検出の適時性および達成範囲の向上のためには、今後はおそらく、ニュースによる事象や、ツイッターのようなサイト上での個人的な報告による情報を融合し、取り込んでいく必要があると考えられるであろう。このような方法でのエビデンス結合は今後の研究における大きな課題となってくるであろう。

4, 自己評価

このプロジェクトでは、公衆衛生に関する脅威を、オンラインのメディアソースを用いて検出するシステムの性能向上のために、アルゴリズムとリソースが開発された。私は、本来の目的の重要な部分は達成できたものと考えている。 (a)公衆衛生事象の自動警告に向けての新たな方法の開発への挑戦、(b)コンピュータに公衆衛生を認識させる手助けとなる知識リソースの開発、(c)さまざまなソースからの、エビデンスを融合させる方法の発見 (a)においては、まず最初に、信頼性のある外部基準に対する、自動システム性能を評価する方法を見つける必要があった。研究開

始当初は、基準というものが存在しておらず、ProMED メールの利用によって、自動警告の性能について現実的な見識を得ることができたのである。確立した評価基準によって、自然言語処理および変化点検出アルゴリズムを併用するという、警報に対する新たなアプローチの開拓が可能となった。私自身の、また公衆衛生アナリストらによる独自の研究を踏まえると、私は、この方法によって、分析者は時間と労力を費やすことなく最高水準の結果を得ることができると考えている。この利点は、BioCaster BORN より日本や世界保健機関、米国疾病監理予防センター(CDC)、欧州疾病監理予防センター(ECDC)といった海外の公衆衛生機関のアナリストに送信される警報に見ることができる。私は、(b) において共同研究者とともに、世界初の自由にダウンロードが可能な多言語の公衆衛生オントロジーの開発に成功したのである。このリソース内の用語は公的公衆衛生機関の主導についてアナリストと協議することによって誘導されてきており、今後も改良、拡大が見込まれる。これまでに、世界中で 350 を超えるグループがオントロジーのダウンロードを行っている。最後になるが、(c)では、ニュースレポート内の情報融合を達成するまでには、当初の予測よりはるかに長い時間を要することとなり、また、新たなアルゴリズムの開発だけでなく、知識リソースの産出も必要となった。地理−時間的情報の自動注釈のための新たなスキーマの作成、ならびに報告事象間の関連性を理解するための知識モデルの構築といった本来の目的の一部はすでに達成されている。ひとつの機械学習モデル内にこういったすべてのリソースをまとめることが、今後の研究の当面の主要目標となるであろう。

この研究における最終的な成果は、リアルタイムでの健康関連事象の生物−地理マップや、過去 3 年の新たな事象のデータベースといった、公的に入手可能な BioCaster BORN プラットフォーム上で見ることができる。この研究の開始当初には、データベース化は想定されていなかったのだが、公衆衛生専門家との協議過程で、感染疾患の拡大を調査する上で、その必要性を認識するに至った。プロジェクトを通じてすべての技術ならびにリソースの実現が、世界規模の公衆衛生の保護に利益をもたらすことを願うものである。

5, 研究総括の見解

インターネット上に流れている健康被害に関するニューステキストを分析して健康被害情報に関するアラームを発信するシステムの提案であり、社会的意義が高く、大規模な応用分野をカバーしている課題である。技術的には、イベント系列の解析により情報の重要度を推測し、アラートを出せるようになることを期待していた。

この課題で、公衆衛生に関する脅威を、オンラインのメディアソースを用いて検出するシステムの性能向上のために、アルゴリズムとリソースを開発しており、本来の目的の重要な部分は達成できたものと評価する。また、新たなアルゴリズムの開発だけでなく、地理−時間的情報の自動注釈のための新たなスキーマの作成、報告事象間の関連性を理解するための知識モデルの構築といった本来の目的の一部はすでに達成されていることを評価する。自然言語処理および変化点検出アルゴリズムを併用するという、警報に対する新たなアプローチの開拓が可能となり、開発した多言語の公衆衛生オントロジーは、これまでに、世界中で 350 を超えるグループによりダウンロードされ、利用されている。課題提案時より暫定的に動いていたシステムを中心にした研究開発であったため、完成度が何より重要である。この期待に応える利用実績を示し、この分野に大きく貢献したと考える。

6，主な研究成果リスト

（1）論文（原著論文）発表

1. Collier N. (2011), “Towards cross-lingual alerting for bursty epidemic events”, *BMC Biomedical Semantics*, 2 Supp 5: S10.

2. Collier N *et al.* (2010), “An オントロジー-driven system for detecting global health events”, Proc. COLING 2010, Beijing, China, pp. 215-222.

3. Collier N. (2010), “What’s unusual in disease outbreak news?”, *BMC Biomedical Semantics,* 1(1).

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

5. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo QH, Dien D, Kawtrakul A, Takeuchi K, Shigematsu S, Taniguchi K. (2008), “BioCaster: detecting public health rumors with a Web-based text mining system”, *Bioinformatics*, 24(24): 2940-2941.

（2）特許出願

（3）その他の成果（主要な学会発表、受賞、著作物等）

[1] Collier N. “GENI-DB: A database of Web-based global event disease reports”, *Bioinformatics* (under review).

[2] Collier, N. “An overview of text mining for epidemic intelligence”, *Global Public Health* (under review).

[3] “Text mining in action: early alerting of disease outbreaks from online media”, Talk given at the American Association for the Advancement of Science Annual Meeting, Vancouver, special track on Web Surveillance: Fighting Terrorism and Infectious Diseases, Canada (2012.2).

[4] “BioCaster: Web sensing for real time disaster detection and tracking”, Invited talk given at the Workshop on the Politics of Disease Surveillance: how unofficial reporting is changing official behaviour, Brisbane, Australia (2011.7).

[5] “Analysis of the grey literature including news events and user generated content”, Invited talk given at the EMBL-EBI Industry Programme Workshop on Literature Services, Cambridge, UK (2011.6).

[6] “Web sensing for real time disaster detection and tracking”, Invited talk given at the University of Manchester, School of Computer Science, UK. (2011.6).

[7] “Web sensing for real time disaster detection and tracking”, Invited talk given at the University of Tokyo Institute of Science and Technology, Department of Computer Science, Japan (2011.6).

[8]Collier N *et al.* (2010), “Navigating the Information Storm: Web-based Global Health Surveillance”, in BioSurveillance: Methods and Case Studies, Kass-Hout, T. and Zhang, X. (eds), Chapman and Hall.

[9] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk at the University of Zurich, Department of Informatics, Switzerland (2010.10)
[10] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk at Cambridge University, Computer Laboratory, UK (2010.10).
[11] “Web signals and sensors: an overview of public health alerting in BioCaster”, Invited talk at Oxford University, Department of Zoology, UK (2010.10).
[12] “Online text analysis for early alerting of disease outbreaks”, Invited talk at the National Institute of Public Health, Japan (2010.10).
[13] Doan, S., Conway, M. and Collier, N. (2010), “An Empirical Study of Sections in Classifying Disease Outbreak Reports”, *Annals of Information Systems*, Special Issue ″Web-based Applications in Health Care & Biomedicine″, Springer.
[14] Hartley D, Nelson N, Walters R, Arthur R, Yangarber R, Madoff L, Linge J, Mawudeku A, Collier N, Brownstein J, Thinus G. and Lightfoot N. (2010), “The landscape of international event-based biosurveillance”, J. *Emerging Health Threats Journal*.
[15] “BioCaster: early detection of public health events on the Web”, Invited talk at the Japan Science and Technology Agency, Austria-Japan ICT Workshop, Japan
[16] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk given at the Centre for Epidemiology and Risk Analysis, Veterinary Laboratories Agency, UK (2010.7).
[17] “Text mining in action: Global disease surveillance and alerting from online news”, Invited talk given at JEITA, Japan Electronics and Technology Industries Association, Japan (2010.6).
[18] 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.
[19] 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*, 79(4): 284-296.
[20] “High throughput analysis and alerting of disease outbreaks from the grey literature”, Invited talk given at the European Bioinformatics Institute, Cambridge, UK (2010.1).