厚生労働科学研究費補助金

特別研究事業

健康危機情報の積極的収集と分析および健康危機管理行政への情報提供のためのシステム開発と運用に関する研究

平成 20 年度　総括・分担研究報告書

平成 21 年 3 月

研究代表者　ナイジェル コリアー

（国立情報学研究所）

厚生労働科学研究費補助金

特別研究事業

健康危機情報の積極的収集と分析および健康危機管理行政への情報提供のためのシステム開発と運用に関する研究

平成 20 年度　総括・分担研究報告書

平成 21 年 3 月

研究代表者　　ナイジェル コリアー

(国立情報学研究所)

平成20年度　特別研究事業

健康危機情報の積極的収集と分析および健康危機管理行政への情報提供のためのシステム開発と運用に関する研究

班員名簿

氏　　名	所　　属	職　　名
ナイジェル コリアー	国立情報学研究所　情報学プリンシプル研究系	准教授
竹内　孔一	岡山大学大学院自然科学研究科	講　師
谷口　清州	国立感染症研究所 感染症情報センター	室　長
重松　美加	国立感染症研究所　感染症情報センター	主任研究官
ソン　ドアン	情報・システム研究機構	特任研究員

目　次

平成 20 年度総括・分担研究報告書

総括・分担 研究報告書

厚生労働科学研究費補助金（特別研究事業）
総括研究報告書

健康危機情報の積極的収集と分析および健康危機管理行政への情報提供のための
システム開発と運用に関する研究

WEBベースの国際的健康危機情報システムに関する検討

- 将来的な日本のシステム設計における方法論的提言-

研究代表者　ナイジェル・コリアー　国立情報学研究所 准教授

研究要旨

背景：インフルエンザ・パンデミックなど、世界の公衆衛生におけるリスク評価の情報収集は、訓練された分析者を必要とする高度なタスクである。ワールド・ワイド・ウェブ（World Wide Web）上のニュース報道など、リアルタイムに近い情報が豊富なことは魅力的であり、自由に利用できる情報源であるが、その巨大な規模、曖昧さ、そして多言語性によって、情報を持つ分析者を圧倒するリスクがある。言語処理技術（コンピュータを使用して構造化されていないテキストを、データベースに保存可能な構造化情報に変換することを支援するソフトウエア）を使用することで、適時性、処理能力、および精度を高めることができ、この問題を軽減できる可能性がある。本報告書で我々は、言語処理技術が、分析者の支援において担う潜在的な役割および日本の公衆衛生コミュニティにおけるグローバル・ヘルスのサーベイランス要件にどのように組み込まれるかについての予備研究の結果を報告する。

方法：いくつかの基本的な言語処理技術を概説し、一部の政府システムおよび民間システムにおける開発の背景、対象範囲、および方法の調査について構造化したケーススタディの結果を示す。6つのシステムのケーススタディとして、2008～2009年の4ヵ月間に、重要な情報提供者との面談およびメールによる追跡を行った。また、現在の日本の世界的サーベイランスの方針および実情について歴史的な背景を簡単に検証した。

結果：我々は、考察すべき主な要素を要約し、将来的な協議のために意思決定者による技術的な提言を行い、広範な方法論的な要件を考察する。協議において考察する分野は、言語対象範囲、自動分析、人間による分析、通信方式、ハードウエア基盤、技術サポート、標準、およびパートナーシップを含む。また、今後の進歩のため、自動警報および状況認識など、研究レベルのサポートを必要とする多くの課題分野についても言及する。

結論：GPHINにより、SARSが世界保健機関へ適時に通知されたことから示されるように、現在のグローバル・ヘルス情報システムは、早期警戒情報の提供、米国、カナダ、および欧州連合における公衆衛生対応への通知、そして国際的な公衆衛生ネットワークへの情報提供という点で成果を上げている。本報告書は、慎重に設計された動作パラメータによって世界的な公衆衛生に関する認識および対応の強化が可能であることを示すことで、日本の将来的な国家システムに関する協議プロセスに貢献できることを望む。

研究分担者

ナイジェル・コリアー	国立情報学研究所 准教授	谷口　清州	国立感染症研究所　室長
竹内　孔一	岡山大学大学院 講師	重松　美加	国立感染症研究所主任研究
		ソン　ドアン	情報・システム研究機構 特任研究員

A．研究目的および背景

第一に、グローバル・ヘルス情報システム（GHIS）は、次の特定ニーズに対処する必要がある。公衆衛生上の危機が世界中に拡大する前に検討するため、訓練を積んだ専門家による膨大な公開メディア情報の精査を支援すること。これにより、各国および世界の公衆衛生コミュニティは、罹患および死亡の減少のため、入手可能な最良の情報に基づき、災害管理のために適切な対応をする機会を得ることができる[1]。

2008年6月現在、14億6000万人が630億ページ以上のウェブ情報にアクセスしていると推定されている[2,3]。ウェブ閲覧ソフトウエアを使用することで、誰でも、ほぼ瞬時に、広範囲におよぶ公衆衛生上重要である可能性のある電子ニュース報道、企業リリース、ブログ、個人的メモ、学術速報、政府および行政文書に低コストでアクセスすることができる。しかし、ウェブ自体は、無秩序で混沌としており、規模も巨大であり、誤解を招きやすく、また、使用される言語も増加している。つまり、自動化された支援無しに、一人、あるいは、グループがリアルタイムに追跡するにはあまりにも情報が多すぎる。SARS、ウェストナイルウイルス、あるいは将来的なインフルエンザ・パンデミック[4]などの公衆衛生災害の国際的な拡散に対する社会的関心の高まり、さらに、従来のサーベイランスリソースの配分に関する専門家の懸念[5,6]などにより、ウェブから全世界の公衆衛生データを収集するためのシステムが注目されている。現在据えられているGHISの総合的な目標は、ウェブ上の多言語の情報をほぼリアルタイムで検出し定量化すること、ならびに適切な当局への報告である。

本研究は、健康安全の維持において多くの領域が役割を担っているとの理解に基づいて実施された。本研究では、主に、ウェブベースの情報収集が、インフルエンザ・パンデミックなどの健康安全上のリスクを体系的に検出し、評価にいかに役立つかについて検証している。この点に関して、言語処理技術を使用したWorld Wide Webから情報を収集するグローバル・ヘルス情報システムにおける最新技術の現状について検証した。本研究の目的として我々は、言語処理技術とは、コンピュータを使用して自然言語で記述された情報を構造化したデータベースに変換し、保存、検索、そして人間による意思決定において使用可能なグラフあるいは地図など、効果的な方法で表示するためのソフトウエアと定義している。我々は、日本の公衆衛生システムにおいて感染症アウトブレイクの適時的な報告を行うために、本技術をどのようにデザインすることができるかについて検証した。しかし、本報告書の研究は、決して完全なものではなく、統合された方法論ではないことを認識することが重要である。本研究には、時間的制約があり、これまでほとんど検証されていないことから、稼働中のシステムのケーススタディを検証することで最適な実践方法について示唆する程度に限られる。

考察では、GHIS技術における現在の限界を明らかにし、既存システムの有効性を向上させるために将来の研究に役立つと考えられる分野の示唆を試みている。我々が検証した重要な問題は以下の通りである。GHIS技術が、どこで、どのような目的で使用されているか。現在の自動化と人の手による処理の境界線はどこか。サーベイランスにおける2つのモード（自動および人の手）がどのように相互に補完するか。また、自動化されたヘルスサーベイランスシステムに早期警報システムを組み込む方法についてもある程度

触れている。前述の通り、本研究における時間的な制約のため、世界中で行われている努力のすべてを取り上げることはできない。このため、本研究では、我々が知る最も重要な国、多国間、および民間が行う努力のうち、いくつかを取り上げる。この点に関し、我々は、我々の同僚、とくにGHSAG（世界健康安全保障グループ）のリスク連絡・管理コミュニティから多大な助言およびフィードバックを得ている。

本報告書で示すように、公衆衛生組織では、届出疾患の報告および市販薬（OTC）の販売量の監視などの従来の代用方法を拡張し、ウェブおよび検索エンジンデータからの自然言語データの自動分析[7]を考察する動きが高まっている。この背景には、大量のほぼリアルタイムのオンラインニュースを利用することができること、また、世界規模でヘルスイベントを特定でき、追跡可能である費用効率が高い手段であることがある。同時に、大量のメディアデータを精査することは、極めて時間がかかるタスクでもある。我々は、公開メディアによる診断および診断前報告の懸念や限界について認めながらも、風説に基づく情報が、(a) サーベイランスシステム基盤が存在しない地域で起こりうる発生を人の手で監視することに対する支援、および (b) 他の情報ソースが存在する場合において、既存システムの機能を強化する。そのようなシステムを適時的かつ適切な公衆衛生対応を可能にするための広範なソリューションの一部として考える必要がある。

疫学者は、ある集団に疾病が発生した状況、それらの発生、拡散、認知、および抑制に影響する因子を懸念している。調査の一環として行われた本分野の専門家との協議において、以下に関する適時的な情報は、特に重要であることが明らかにされた。(1) 新型インフルエンザ・パンデミックなどの新興感染症のアウトブレイク (2)「動物から人への感染」から「感染した人から人への感染」への移行 (3) 国境を越えた外来病の流入、および (4) 偶発的または意図的による生物学的、化学的な放射性あるいは核病原体の放出。これらは、世界保健機関（WHO）の国際保健規則（IHR）[8]に記載される懸念を正確に反映している。激しい経済的圧力により多くの国が直面するコンプライアンス上の実質的な障壁を前提すると、公開メディアソースを使用するグローバル・ヘルスサーベイランスシステムは、世界的な公衆衛生コミュニティに対する保健上の脅威の透明性を高める上で特別な役割を担うものと考えられ、それによって、現在進められている国際保健規則の実現に必要な能力の構築に貢献するものと考えられる。

本研究では、主に、保健サーベイランスの自動化技術について取り上げるが、人的なネットワーク、とくに国家間のネットワークが担う中核的な役割が極めて重要であることを忘れてはならない。これらにおいて主要なものとしては、WHO の Global Outbreak Alert and Response Network (GOARN) [9] がある。このネットワークには、我々が今回検証したいくつかのシステムから公式・非公式の入力がある。GOARN は、国境を越えて拡がる疾病アウトブレイクの脅威に関する国際社会の認識を高めるために、人的および技術的リソースをプールする組織および個人の連携を図っている。さらに、世界健康安全保障イニシアティブ（GHSI）[10]（カナダ、フランス、ドイツ、イタリア、日本、メキシコ、英国、米国および EU を含む）によって、集団的準備を促進するための政府高官のネットワークである世界健康安全保障グループ（GHSAG）が創設されるに至っている。GHSAG 自体も、インフルエンザ・パンデミックのような特定の国際的次元での健

康上の脅威に対応するための多くの作業部会を創設している。これらの作業部会の実り多い協力によって、ベストプラクティスおよび健康安全のためのアプローチおよび技術の評価が推進されている。また、いくつかの自動警報システムと GHSAG の業務、とくに国および多国間システム間においてクロスオーバーがある。

とくに注目するべきは、International Society for Infectious Diseases の公式プログラムとして運営されている専門家ボランティアによる民間ネットワークである ProMED-mail である。ProMED-mail メンバーは、公開メディア上の報告および他のソースをモニタリングし、ヒト、動物、および植物に影響を及ぼす生物学的および化学的災害に関する情報を、インターネットを通して報告している。現在、ネットワークには、165 ヵ国の 40,000 人以上が参加しており[11]、自動化アプローチの重要な標準とされている。本報告書の本文でこのシステムを取り上げなかった理由として、今回の研究は、自動化された警報システムの役割に要点を置いたためである。ただし、近年 ProMED-mail と密接な関係を構築している HealthMap については調査している。

重要なのは、本研究の範囲には将来的に意思決定者が考察する必要がある一定の限界が存在することを認識することである。すなわち、

- 評価：本試験の範囲内で技術を体系的に評価することが不可能であったため、我々は報告可能な結論がある程度制限されることを認識した上で、調査をケーススタディに限定した。これに関する主な理由は、ウェブ情報科学分野の歴史が比較的浅いこと、ならびに、評価基準およびベースラインが確立されていないことである。
- 統合：また対応システムについて、あるいは、我々が取り上げたサーベイランスシステムがどのようにセンチネルサーベイランスや臨床検査機関ベースのサーベイランスなどのより広範な分析フレームワークに組み込まれているかについて、多くを述べることはできない。
- 患者プライバシー：我々のシステムの調査では、個人のプライバシーを保証するための管理について直接的な調査は行っていないが（a）サーベイランスシステムが主に、個人ではなく、イベントの集合に着目していたこと、さらに（b）すべてのシステムにおいて、適切な人員にデータアクセスを制限するアクセス管理機構が備わっていること、が明らかにされた。日本における新規システムの開発における重要な領域として、公的な方針と一貫した適切な管理が実行されることを希望する。
- 観点：本報告書では、技術について利用者の観点から検証している。このため、様々なアプローチの理論的な限界、あるいは、さらに向上するために必要な研究について深く説明することができなった。研究レベルでは、公衆衛生イベントの自動検出に関する研究がテキストマイニングの分野で進められており、既に 1 組のコンポーネントタスクとして集束している。本領域には、それを独特な困難なものにする多くの特徴的課題が残っている。本報告書では、広範な研究分野から、将来の研究を進めるにあたって実り多く興味深いことで多くのことが得られると思われる興味深い研究をいくつ

か取り上げている。これらについては結論において取り上げる。

要約すると、本研究は、異なった学問分野のスキルの融合を必要とする、専門的で新しい分野である公衆衛生サーベイランスにおける現状の言語処理技術の見解をまとめる調査として捉えるのが適切である。我々の見解が、日本における将来的なグローバル・ヘルスサーベイランスシステムの計画に携わる意思決定者の支援となることを目的としている。

本報告書における残りの構成は次のとおりである。

- 技術および科学：本セクションでは、様々なウェブベースの保健サーベイランス システムにおいて今日使用されている重要な技術を概説する。
- ケーススタディ：ここでは、一部の政府および民間システムの調査結果について説明する。方法、入力データ、データの表示および検索のための機能、および人的緩和作用について簡単に比較する。広範な技術的解決策を検証し、可能な場合、付加価値についてケーススタディを用いて具体的に示す。
- 日本における世界的感染症サーベイランスの現状：本セクションでは、日本においてグローバル・ヘルスサーベイランスの現在の見解を形成した主要なイベントを概説し、今日における要件について検証する。
- 提言：本セクションでは、日本固有の観点から問題を総括し、将来的な国のシステムの設計目標を高次レベルで概説する。
- 結論

参考文献・資料

[1] Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., et al. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437, 209–214.

[2] Wikipedia (2009), "World Wide Web", available from http://en.wikipedia.org/wiki/World_Wide_Web (last accessed 29th January 2009).

[3] Wikipedia (2009), "List of countries by number of Internet users", available from http://en.wikipedia.org/wiki/Internet_population (last accessed 29th January 2009).

[4] Crosby, A. W. (2003), "The influenza of 1918 – America's forgotten pandemic", Cambridge University Press.

[5] Butler, D. (2006), "Disease surveillance needs a revolution", Nature, 440:6-7, doi:10.1038/440006a.

[6] Jones, E., Patel, N., Levy, M., Storeygard, A. Balk, D., Gittleman, J. and Daszak, P. (2008), "Global trends in emerging infectious diseases", Nature 451:990-993, doi: 10.1038/nature06536.

[7] Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M. and Brilliant, L. (2008), "Detecting influenza epidemics using search engine query data", Nature, doi:10.1038/nature07634. Available from http://www.nature.com/nature/journal/vaop/ncurrent/suppinfo/nature07634_S1.html (last accessed 30th January 2009)

[8] Lawrence, O. and Gostin, J. (2004), "International infectious disease law – Revision of the World Health Organization's International Health Regulations", in J. American Medical Association (JAMIA), 291(21), June 2nd, pp. 2623-2627.

[9] World Health Organization (2009), "Global Outbreak Alert and Response Network", available from http://www.who.int/csr/outbreaknetwork/en/ (last accessed 1st February 2009).

[10] Health Canada (2007), "Communique of the Eighth Ministerial Meeting of the Global Health Security Initiative", available from http://www.hc-sc.gc.ca/ahc-asc/media/nr-cp/2007/2007_ghsi-minist-issm-eng.php (last accessed 1st February 2009).

[11] ProMED-mail (2009), available from http://www.promedmail.org (last accessed 1st February 2009)

B．研究方法

言語学技術について、学会等を通じて最新の情報を収集する一方で、各システムに関する著作物、内部資料などを参考に、概要をまとめ、本研究班の活動の中で参加した国際会議、学会、さらに国際的システムを訪問し、それぞれの内容についての詳細な情報を収集した。さらに、サーバー上で既存のシステムを基に一部仮説を検討した。

我々のために時間を割きいていただいた方々、特にシステムに関する部分で Abla Mawudeku、Michael Blench、Abdelhamid Zaghlool (GPHIN、Public Health Agency of Canada)、Stefan Molnar-Szakacs (Centre for Emergency Preparedness and Response、Public Health Agency of Canada)、Noele Nelson および David Hartley (Project Argus、Georgetown University Medical Center)、Ralf Steinberger (MedISys、JRC)、Roman Yangarber (PULS、University of Helsinki)、John Brownstein (Harvard-MIT Division of Health Sciences and Technology) の皆様の協力に対し、この場をもって深謝する。

（倫理面への配慮）

本研究では個人情報の取扱いは無く、疫学指針等に係る調査もないため、個人情報等に関しての倫理面の問題は生じない。

研究対象として情報収集した海外のシステムは、各国あるいは地域の機密にかかわる情報の取り扱いがあり、それぞれのシステムに知的財産権の発生する情報があることから、それらに抵触する部分の知り得た情報については関係者に草稿を提示し、メール等の記述により了解を得るか、記述内容から削除した。

結果および考察

1．技術および科学

C．研究（調査）結果および考察

1．技術および科学

本セクションでは、情報源としてウェブを使用する自動 PHAS を支援する技術を簡潔に説明する。

1.1 テキストマイニング

1990 年代中頃以来、テキストマイニングは、構造化されていない膨大な量のテキストデータを構造化情報に自動的に変換する新興技術であり、それによって分析者が統計的動向分析を利用して実用的な洞察を得ている。これらの洞察は意思決定者に伝えられて、意志決定者は十分な知識を得た上で選択をおこなう。例えばこれは、公衆衛生において、気象データの変化と相関して経時的に拡大する疾病パターンの探求を意味することもある。情報を大きなデータセットから動向までに落とし込むことは情報発見プロセスの一部に過ぎない。高度なシステムになると、ドリルダウンと呼ばれているプロセスにおいて、分析者は集合データと個々の文書中の詳細との間を行き来できるようになる。前述の通り、ウェブには、テキストマイニングで直面する規模、信頼性、曖昧さ、そして多言語性の問題など、ウェブ固有の問題がある。テキストマイニングプロセスにおいて欠かすことのできない要素は事実を人の手で確認できることである。

ウェブのような巨大なデータセットを自動的に処理するためには、ソフトウエアとハードウエアの両方で高いレベルの効率が必要とされる。近年、ハードウエアのコストが大きく低下したことによって、高速プロセッサや大規模なデータ記憶サイロに集積された揮発性データセットの保存における効率性が向上している。テキスト情報を、構造化したデータベースレコードに変換するには、情報を組織内あるいは組織間で容易に共有することができるという利点がある。これに取って代わったのは、人間が新聞あるいはウェブサイトを検索し、短く要約した報告を作成して組織内において回覧する旧来のシステムである。そのような方法は遅く、手間がかかり、特に組織が範囲を拡大するにつれてエラーの可能性が高まる。

テキストマイニングと通常のウェブ探索との違いは何であろうか。今日一般的に使用されている Google 検索に代表されるようなインターネット検索技術では、キーワードおよびページ順位を使用することで、利用者が関心のあるトピックの文書を自然に見つけ出すことができる。この種の文書検索は、言語から独立した方法に基づく汎用技術を使用している。言い換えると、この方法で処理される言語あるいは領域に関する特定の知識を使用しないので、多くの一般利用者の検索要求に柔軟に対応することができる。ただし、目的のトピックがある専門的な利用者は、そのトピックに関する情報だけを必要とする。こういったタイプの使用法では、テキストから詳細情報を自動抽出するために、カスタマイズされたアプローチが必要となる。これが、テキストマイニングで想定されるシナリオである。

テキストマイニングシステムは一般に明確に定義されたタスク指定から始まる。例えば「国境を越えた感染が関与するすべての感染症アウトブレイククラスタを特定する」。ウェブ文書からの構造化されていないデータを構造化情報に変換するためには、コンピュータは言語の構文的および意味論的な構造に関する知識を必要とする。この要件により、テキストマイニングは言語および領域特異的な技術となる傾向があり、システムを開発するためには学際的な協力が必要となる。技術レベルで重要とみなされることが多いスキルには、数理言語学、知識工学、機械学

習、データマイニング、統計学、および情報抽出が含まれる。これらの知識の一部あるいはすべてを特定のタスクのためのコンピュータシステムに構築することが経済的であるのは、ウェブのようにテキスト収集が大規模であり、特定発見されるその情報が利用者にとって非常に価値が高い場合のみである。これは制限的な要件のように思われるが、それでもテキストマイニングは、安全監視、ビジネス情報、環境モニタリング、患者ヘルスケアの向上、および創薬などの多様な分野で使われて、成果を上げている（例：[1-4]を参照）。民間企業が提供するテキストマイニングソリューションには、SAS、SPSS、Nstein、およびLexisNexisなどがある。

テキストマイニングに用いられている方法の詳細な説明については本報告書の範疇を超えるが、後述するシステムの検証および設計に関する提言の技術的背景を示すために、主要なプロセスについて簡潔に概説する。

1. 通常、第一段階は、データの入力である。データソースとして、電子メール、ホームページ、RSS（Really Simple Syndication）フィード、マイクロソフト・オフィスファイル、PDF（Portable Document Format）文書など様々なテキストソースが使用される。
2. 不要データの削除は瑣末な技術であるが、テキストから不要なノイズ（広告、無関係なニュース記事へのリンクなど）を削除したり、途中で切れた文章をつなげたりする際に不可欠なプロセスである。不要データが確実に削除されていないと、後の段階において、文章のトピックが不明瞭になる恐れがある。この段階においてシステムは、複数のトピックについて取り上げている大きな文書をそれぞれ個別に扱えるように異なるセクションに分解することが多い。例えば、H5N1インフルエンザのアウトブレイクに関する記事には、1918年のパンデミックに関する歴史的な記述が含まれる可能性がある。これはゾーニングと呼ばれている。
3. 次に、これらの文書群に対してデータ重要度判定を行い、廃棄文書（非重要文書）、または以降の段階で詳細事実抽出を実施する処理用文書のいずれかに分類する。この段階において、同一イベントの複数の報告など、重複した情報が文書クラスタ化によって検出される。この段階では、明白な真陰性の削除が目的であるが、システムは、タスクの定義の境界線上にある微妙なケースの扱いに苦戦することがある。例えば、黄熱ワクチンが原因となった暴動に関する記事では、アウトブレイクについて直接触れない可能性があるが、保健システム窮迫状態の重要な証言となる可能性がある。大規模なシステムではこの段階において、以後追跡調査が必要な場合に備え、すべての文書を頻繁に保存する。
4. 利用可能なデータ結果を使用し、重要度が決定される。それは重要度判定段階の結果のみ、あるいは、事実抽出からの結果で可能である。ハイエンドシステムは、複雑な統計解析を使用して検出された各イベントに警報レベルを割り当てる。現実には、システムにとって、高い精度で自動的に行うことが最も困難な段階である。とくに、情報が不確実かつ曖昧、あるいは意思決

定プロセスにおいてより深いレベルの背景知識のモデル化が必要な、イベントの早期段階では困難である。被修飾範囲が的はずれであるか、正しく見抜いている場合でも「、」の打ち方が悪いために読み手に誤解を与えることが多い。

5. 人間の判断は、プロセスにおいて重要な段階である。何に異常があるのか理解すること、システムが見逃した可能性のある稀なイベントを発見すること、曖昧な報告について最終的な判断を下すこと、そして、共通点がないイベントを関連付けることが必要となる。テキストマイニングプロセスは、データベースのデータ検索および視覚化を通じて、検索された事実に対する人間による判断をより簡単かつ安価に行えるようにし、判断の信頼性を高めることを目的としている。ユーザは、情報を検索し、分析ニーズに適した様々なフォーマットで結果を表示させることで解釈することができる。表示の中で最も単純なものは、オリジナル文書へのリンクとともに一覧表として表示する方法である。これは、システムがそのような結果に至った根拠を分析者自身が確認する上で重要である。表中の要素は、一般的な概念（例：INFLUENZA CASE[インフルエンザ症例]）に標準化するか、オリジナル文書のままの形式（例：29-year-old hospital worker[29歳の病院従事者]）で表示することができる。この段階でシステムの限界がユーザに最も明らかになる。ユーザーは、人にとっては意味が明瞭だがコンピュータソフトウエアにとっては不明瞭なニュアンスを修正するために、判断する必要がある。人間による分析では、現在の自動化されたアプローチでは利用できない新しい経路により、調査につながるデータの規則性を見つけることができる。
6. フィードバックおよび改善。テキストマイニングシステムの運用開始後は、新しい語彙、文書の種類、および情報源について更新し、誤解釈を修正するためにシステムを保守・拡張する必要がある。したがってシステムは、ユーザが現在の出力についてのフィードバックプロセスを通じて関与できるようになる高度なグラフィカルユーザインタフェースを搭載していることが多い。

コンピュータが高品質な情報をテキストから抽出するためには、ある程度の言語学的な理解が必要とされる。システムは、通常2タイプの知識を必要とする。すなわち、関心オブジェクトのクラスとそれらの関係を示す分野依存知識（DISEASE[疾病]と呼ばれる状態を引き起こすPATHOGEN[病原]と呼ばれる病原体群、あるいはPATHOGEN[病原]に感染するPERSON[人]と呼ばれる人口群など）、およびこれらの関係が実際のテキストの言語においてどのように顕在化するかのパターンである。例えば、「PERSON was infected with DISEASE（PERSON[人]がDISEASE[疾病]に感染した）」が、「the 59-year-old farm worker was infected with H5N1 influenza（59歳の農場労働者がH5N1型インフルエンザに感染した）」と言うテキストに相当する可能性がある。前述のプロセスフローにおいて、知識管理はシステムのセットアップ

（第一段階の前）、フィードバックおよび改善段階で扱われよう。

時間および場所を理解することは、高品質なデータを取得するために重要である。しかし、実際には多くの落し穴がある。例えば、文書のタイムスタンプは、報告されたイベントの発生時間を判断するための最良の方法とは限らない。例えば、2008年10月2日付の文書に「先週の火曜日に、ベトナム南部の 2 つの省におけるアウトブレイクの原因として、A型鳥インフルエンザウイルスが確認された」と報告されたとする。我々は、テキストマイニングシステムがこのイベントの日付を2008年9月30 日と記録することを期待するだろう。実際には、場所も曖昧な場合が多い。例えば、2007年夏にCamdenで発生したウマのインフルエンザのアウトブレイクは、オーストラリアのシドニー近郊のCamdenと識別される必要があって、英国ロンドンのCamdenではない。

ウェブ上の文書は様々な自然言語で記述されているため、コンピュータは、異なる語彙および文章構造を理解する必要があるが、これには高いコストがかかる。テキストマイニングシステムが複数の言語の文書を扱うためには、システム設計者によるさらなる戦略的な考察が必要とされる。これらの詳細については後述する。

これまでシステム評価に関する問題については取り上げていない。研究の範囲内において、政府が支援する共有タスクを通じて、様々なタスクベースのゴールドスタンダードができている。これらは、用語、用語の相互関係、およびその他の重要な情報の判定について、主に、感受性および特異性などの標準的な信号処理メトリックスに基づいている。これらの共有されたタスクの例としては、Message Understanding Conferences（MUC）、要約に関するDocument Understanding Conferences（DUC）、Automatic Content Extraction（ACE）プログラム、そしてゲノミクス領域におけるBioCreativeが含まれる（例：[5-8]を参照）。現在、公衆衛生領域においてタスクベースの評価は存在しないが、コミュニティグループが形成され始めており、将来的な評価のインキュベータとして機能する可能性がある。

参考文献・資料

[1] Cody, W., Kreulen, J., Krishna, V. and Spangler, W. (2002), "The integration of business intelligence and knowledge management", IBM Systems Journal, 41(4), DOI:10.1147/sj414.0697.

[2] Esterby, S. (2006), "Trend analysis methods for environmental data", in Environmetrics, Wiley Interscience, 4(4), pp. 459-481.

[3] Hahn, U., Romacker, M. and Shulz, S. (2002), "Creating knowledge repositories from biomedical reports: The MEDSYNDIKATE text mining system", Proc. Pacific Symposium on BioComputing (PSB), Hawaii, USA, pp. 338-349.

[4] Wishart, D., Knox, C., Guo, A. Shrivastava, S., Hassanali, M., Stohard, P., Chang, Z. and Woolsey, J. (2006), "DrugBank: a comprehensive resource for *in silico* drug discovery and exploration", Nucleic Acids Research, Oxford University Press, 34(Database issue): D668-D672, DOI: 10.1093/nar/gkj067.

[5] Grishman, R. and Sundheim, B. (1996), "Message Understanding Conference – 6: a

brief history, Proc. International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 466-471.

[6] Nenkova, A. (2005), "Automatic text summarization of newswire: lessons learned from the document understanding conference", Proc. National Conference on Artificial Intelligence (AAAI), Pittsburgh, USA.

[7] Doddington, G., Mitchell, A., Pryzbocki, M., Ramshaw, L., Strassel, S. and Weischedel, R. (2004), "The automatic content extraction (ACE) program – Tasks, data and evaluation", in Proc. Language Resources and Evaluation Conference, pp. 837-840.

[8] Hirschman, L., Yeh, A., Blaschke, C. and Valencia, A. (2004), "Overview of BioCreAtIvE: critical assessment of information extraction for biology", BMC Bioinformatics, 6(Suppl 1):S1, DOI: 10.1186/1471-2105-6-S1-S1.

1.2 機械翻訳

ウェブ上の文書は様々な自然言語で記述されているため、処理のある段階において、手動または自動的な手段による翻訳が必要となる。手動翻訳では、多くの場合において最高品質が確実に保証されるが、リソースまたは時間が制限される場合の多くでは、自動翻訳が考慮される。したがって、ユーザに海外ニュースメディアにおいて報道される保健関連のイベントのトピックおよび重要性を理解することを支援する機械翻訳は、多言語グローバル・ヘルス情報システムを構築する上で重要な技術であると考えることができる。様々な言語のニュースに対応するために、GHIS に機械翻訳（MT）を組み込む方法にはいくつかある。ここでは、考えられる数多くのシナリオから 2 つ選び、それらについて説明および検証する（図 1 を参照）。図 1 左側のシナリオ 1 では、まず、様々な翻訳元の言語で記述されたすべての文書が 1 つの翻訳先の言語に翻訳されることを示している。その後、これらすべての文書がテキストマイニングシステムによって処理される。図 1 右側のシナリオ 2 では、翻訳元の言語別にテキストマイニングシステムが構築されている。各言語依存テキストマイニングシステムによる処理後、文書は MT を使用して翻訳される。2 つのシナリオにおけるシステムの特徴は以下の通りである。

シナリオ 1 における明白な利点は、1 つのピボット言語（通常は英語）を対象とした 1 つのテキストマイニングシステムのみしか構築する必要がないことである。その利点は、構築時間を短縮可能なことである。すべての言語について構築するには高価であるトピック分類モジュール、自然言語処理ツール（例：形態素解析器および係り受け解析器）、および指定エントリ抽出モジュールから構成される典型的なテキストマイニングシステムの構築や維持には時間がかかる。単一のテキストマイニングシステムしか必要としないことは、多言語 GHIS の構築にかかる負担の軽減という点で極めて有益である。しかし、テキストマイニングシステムの精度は、MT システムの品質に大きく依存することになる。後述する通り、MT システムの品質そのものがそれらの辞書に大きく依存しており、また、疾病名があまり聞かれないあるいは広く知られていない場合には、一部の深刻な疾病に関するニュースを見逃す可能性が高い。

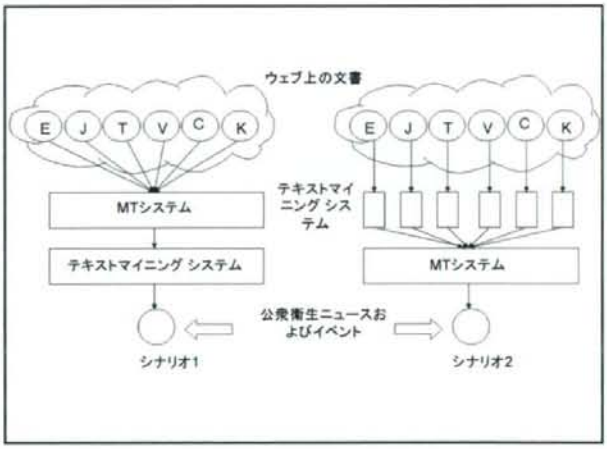

図 1 ：グローバル・ヘルス情報システムへの機械翻訳の適用が想定されるシナリオ。

シナリオ 2 の利点は、テキストマイニングシステムが MT の品質に左右されずに、すべての保健関連のニュースを検出することができることである。疾病またはウイルスの名称など、専門用語の複数言語における同義語情報が、多言語辞書またはオントロジーに収載されている。各言語のテキストマイニングシステムが用語を翻訳元の言語のまま識別する。これは、日本人エンドユーザがベトナ

ム語またはタイ語で記述された関連ニュースを、日本語で検索できることを意味する。日本語で入力された検索語は多言語オントロジーによって翻訳される。翻訳された用語は、翻訳元の言語のニュース記事中特定に特定された用語と比較される。これは、ニュースの検証および完全な理解という観点からは利点となるが、ユーザは機械翻訳を用いて元のニュースを高い精度で翻訳する必要がある。また前述の通り、すべての言語のテキストマイニングシステム、さらに多言語オントロジーの構築および維持は、テキストマイニングシステムが各言語に大きく依存する多くのモジュールから構成されるため、多大な時間および労力を要する。

MT の開発および品質の現状から、MT が組み込まれた GHIS の最終的な性能には以下の 2 つの大きな制約が想定される[1],[2],[3]。1 つ目は各言語ペア間の翻訳のためのソフトウエアがあるかどうか、2 つ目は翻訳の品質である。最初の制約については、ほとんどの MT システムは母国語と英語の翻訳用に開発されているため、ピボット言語として英語を使用する必要がある。多くの MT システムが日本語を扱うことができるものの、商用 MT のリストによると、一部の言語（例：ベトナム語）については、日本語に翻訳したり、日本語から翻訳したりする MT システムが存在しない[4]。したがって、上記のシナリオにおいては、すべてのニュースを英語に翻訳する必要がある。日本人エンドユーザが、出力されたニュースを日本語に翻訳することを求める場合、処理の最終段階において英日 MT を適用することができるはずである。

この英語を介した日本語への翻訳という制限があるため、MT の品質を保証することがユーザの経験に影響を及ぼす大きな課題であり、評価基準という疑問を提起する。人間の参照基準に対する MT の自動評価は、これまで極めて困難なタスクと考えられてきた[5]。過去 10 年間、MT 研究者は、自身のモデルの実証的な品質測定法に対するニーズから、評価メトリックスの設計などにおいて大きな成果を上げている（例：[6]を参照）。これらの例としては、現在では、コミュニティによって広く受け入れられている Blue[9]などがある。Blue は優れた MT システムの簡易評価方法であるが、分かりやすさや可読性などの実用的な品質を評価しない。これまでの機械翻訳に関する研究では、出力の用語対象範囲、分かりやすさ、可読性、および忠実さを反映する測定値に基づいて品質の評価が試みられてきた[7-8]。多くの場合、人間が出力を評価する必要があるため、多大な時間と労力を必要とする。さらに、MT システムによって評価者が異なる場合には、MT システムを評価者の分野依存知識に依存するスコアで単純に比較することはできない。また、エンドユーザが翻訳を使用する様々な目的（例：分類のための要約、あるいは、総括のための完全な理解）あるいは、翻訳先の言語に同等の言葉が存在しない概念については翻訳元の言語において関心概念を要約あるいは拡大する必要があることなどについて考慮する必要がある。一言で言えば、完全な翻訳など存在しないということである。

本報告書では、時間的制約のため、MT 品質の実証的スコアの算出、あるいは、MT システムの詳細な比較は行わないが、MT システムの品質については実例を使用して検証する。つまり、サンプル文章に適用することによって GHIS の観点から翻訳の品質を検証する。我々は、MT が GHIS 処理パイプラインにどれだけ貢献することができるか、そして、新しい記事の人間による分析における最終的な結果を活用する方法について考察する。図 1 に示したテキストマイニングシス

テムは多くの詳細な段階から構成されている。最も基本的な段階は、文書からの構造化されていない単語を使用したトピックの自動分類である。ここでは、大部分の生物学的用語を正しく翻訳することができれば、MTが役に立つと考えられる。テキストマイニングにおける最終段階は、文章の構文法（例：動詞-名詞）関係に基づいてイベントを抽出することである。このレベルの情報抽出を円滑に行うためには、極めて高い品質の翻訳が必要とされる。用語あるいは文法の翻訳先の言語への翻訳におけるすべての間違いが、テキストマイニングの品質に影響を及ぼす可能性がある。新しい記事の人間による判断にも高い品質の翻訳が必要となる。

以下は、WHOテキストを富士通のAtlas機械翻訳ソフトウエアを使用して英日翻訳したものである。

<WHO例1>

Epidemiological investigations uncovered exposure to infected poultry as the likely source of infection in both cases. To date, Indonesia has reported 7 human cases of H5N1 avian influenza. Four of these cases were fatal.

<Atlasの結果>

疫学的調査はどちらの場合も、ありそうな感染源として感染した家禽への露出の覆いを取りました。これまで、インドネシアはH5N1 鳥インフルエンザの7件の人間の症例を報告しました。 これらの4つのケースが致命的でした。

<人間による翻訳>

疫学的調査により、両方の症例で、感染家禽への曝露が最も疑わしい感染源であることが明らかになった。本日までにインドネシアは、7例のH5N1亜型鳥インフルエンザの症例を報告している。うち4例が死亡している。

「avian flu」および「Epidemiological investigation」などの専門用語が上手く翻訳されていることが分かる。領域に依存する表現である「human case」における「case」および「source of infection」における「source」は、それぞれ正しく「症例」および「源」と翻訳されている。これらの結果から、領域依存性が高い単語が含まれる一般的な保健関連の専門用語は、正しく翻訳されることが期待できる。この成功の理由の1つは、MTの手動で構築された領域特定辞書の存在である。一方、「exposure to」などの複数の単語から構成される表現は、単一の辞書表現として認識されず、ソフトウエアは間違った日本語翻訳を出力している。また、動詞「uncovered」の翻訳も適切ではない。しかし、この例では、ユーザは意図された意味を理解することが可能であると思われる。さらに、「fatal」および「these」の翻訳が正しくなく、結果、「Four of these cases were fatal」の意味を翻訳から理解することはできない。原因は、MTソフトウエアは、「these」が何を指しているかのような、離れた文脈関係の理解を困難とみたためだと思われる。

<WHO例2>

The patient visited her husband in Nonthaburi Province, north of Bangkok, where backyard chickens had begun to die a few days earlier.

<Atlasの結果>

患者は、裏庭の鶏が数日より早い状態で死に始めたバンコクの北のノンタブリー州で彼女の夫を訪問しました。

<人間による翻訳>

患者は、バンコクの北にあるノンタブリ県の夫を訪ねたが、そこでは2－3日前から裏庭で飼育していた鶏が死に始めていた。