

情報学が、感染症の 広がりを食い止める

「テキストマイニング」という言葉をご存知だろうか。

体系化されていないテキストデータの中の、

言葉の意味やそれらの相関関係を分析することで、

特定の目的に対するそのテキストの有用性を評価する言語処理の手法のことである。

このテキストマイニングを用いて感染症問題の解決に取り組むプロジェクト、

BioCasterが各国の研究機関の連携によって進められているという。

その活動の最前線を紹介しよう。

Webテキストの利用価値

感染症と聞いて、まず思い浮かべるのは昨年の新型インフルエンザの大流行だ。流行情報に関して重要なのは、発生地域をできるだけ正確に知ることである。

そんな中、情報システムによって感染症の拡大を防ぐことを目的としたプロジェクト、BioCasterが注目を集めている。Web上に存在する、感染症に関するさまざまな情報をシステムで自動的に収集・解析し、警告および対策や治療のための参考情報として

Web上で公開すること。それが、プロジェクトのメインの活動だ。膨大な量のテキストから感染症関連の必要な情報を抽出するために、システムにテキストマイニングの手法が使われているところが特徴である。

「鳥インフルエンザが流行したときに、自然言語処理分野における私の研究が、感染症の監視システムというかたちで社会に貢献できるのではないかと考えました」と、BioCasterの研究責任者であるNIIのコリアー・ナイジェル准教授は言う。

BioCasterの概略はこうだ。まず、Web上から収集してきた大量のテキストデータから感染症関連の情報だけを抽出する。その後、症例、病原体、日時、場所などの重要概念を認識し、構造化されたイベント情報を取り出す。それらの緊急度に応じて文書にランクを付け、Web上に公開し、専門家に感染症対策のために活用してもらう。以下、フェーズごとの機能を詳しく紹介していく。

ニュースレポートから Twitterまで幅広く情報を収集

BioCasterのシステムによる感染症監視は、情報の収集からはじまる。ニュースレポートや公的機関から発表される文書はもちろん、公のメーリングリストでのやりとりに至るまで、Web上に公開されたさまざまな情報を集めてくるのだ。誰もが個人単位で情報発信ができる、というのがWebの大きな特徴であるが、コリアー准教授はその点を最大限に生かそうとしている。「ブログ、TwitterやFacebookなどのソーシャルメディアも今後は情報収集の対象として考えていく必要があると思います。というのも、地域によっては、公的機関が



コリアー・ナイジェル

Nigel Collier

情報学プリンシプル研究系
准教授

情報を提供するためのインフラがほとんど整備されていないためです。確かに、個人が発信する情報ですから、情報の信憑性は高いとは言えません。しかし、そこから必要な情報を抽出する仕組みを開発するのは、私にとって大変やりがいがあることなのです」

多言語情報に対応する

「英語、タイ語、ベトナム語、日本語の4カ国語からスタートし、現在12カ国語対応に向けて動いています」と、コリアー准教授はBioCasterで扱う言語の多様化について語る。さまざまな言語で書かれた情報は、どの言語のものも、分類の前にシステム内でいったん英語に翻訳される。この過程には感染症分野特有の問題がつかまとう。感染症は、1つの決まった名前と呼ばれることがほとんどない。豚インフルエンザがそのいい例で、「Swine Flu」「Swine Influenza」「Pig Flu」など、英語だけでもこれだけの呼び名がある。そのため、英語に翻訳する際、英語圏で慣用的には使われない単語に訳されるという問題が頻発してしまうのだ。単語同士の組み合わせを1つ1つシステムに登録することにより解消していくことになるが、地道な作業が必要なことは想像に難くない。この過程は、まだ人手を介さなくてはならない部分なのである。

即時性の向上が課題

英語に訳された情報は、テキストマイニングの手法で分類されていく。BioCasterのシステムでは、「感染症の名前」「発生している地域」など、情報を分類する際にキーとなる語句に注目し、それらの関係性などから1つのカテゴリーを特定していくという。現在、文書の分類という観点では、人間の専門家に比べて70%くらいの正確性で情報のカテゴリーを特定することが可能であるそうだが、「解決すべき課題はまだたくさんある」という。「感染症の流行拡大を防ぐには即時性がまだまだ低いのです。テキストマイニングと



■ BioCasterシステムによる感染症流行地域表示例(アジア地域中心)

というのは、情報がそろわなければならないことですが、感染症の発生当初は、ごく限られた情報しかWeb上に現れません。情報がそろってきた段階では、すでに感染症が広がっているということです。情報が少ない状態で、その感染症の発生を検知し、流行の兆しを予知することができるのがベストなのです」。分類された情報は、データベースに登録され、BioCasterのWebサイトから閲覧が可能となる。「世界中の公衆衛生に携わる専門家への情報提供を主目的としています、それと同時に一般の方の感染症への関心も高まれば嬉しいですね」

国内外の組織との連携、今後の展望

BioCasterの活動はさまざまな組織とのつながりの上に成り立っている。国立感染症研

究所、国立遺伝学研究所、岡山大学、ベトナム国立大学ホーチミン市校、タイのカセサート大学などと連携して、システムの性能向上に日夜勤しんでいる。「私たちのシステムには、まだ人の手を介さなければ成り立たない部分もあります。しかし、着実に成長していることも確かです。このシステムが、情報収集をサポートすることで、感染症の流行拡大を防ぐことが可能となり、社会がより安全になる。その実現のため、今後さらに力を入れて研究に取り組んでいきたいです」

BioCasterのシステムが、性能向上、利用範囲拡大により、世界の公衆衛生情報インフラとなること。それは、医療技術や情報技術の未熟さが原因で、感染症に苦しむ世界各地の人々を救うことにつながっている。

(取材・構成 工藤拓也)