バーは、計算機言語学、知識工学、ソフトウエア工学、および公衆衛生の専門知識を有している。MedISys は、JRC においてメディアをモニタリングしているより大きなグループの一部であるため、開発をより大きなグループで行うことが可能である。そのため、この数字は、システムが実際に利用可能な人的リソースを過少に表すものである。PULS チームは、計算機言語学、ソフトウエア工学、および専門用語処理分野の専門知識を有する約 5 名で構成されている。

システムの歴史：PULS

PULS は、Dr. Roman Yangarber (Department of Computer Science, University of Helsinki, Finland) が学術的プロジェクトの一環として運用しているプロトタイプシステムである。プロジェクトの目的は、自然言語処理技術である高度な知識獲得アルゴリズムの公衆衛生および他領域における応用の可能性を検証することである。

PULS 基本概念は、MITRE グループと共同開発されたサーベイランスシステムの早期プロトタイプである Proteus-BIO プロジェクト[7]に触発されたニューヨーク大学によって 2001ー2003 年に発案された。2004 年に、University of Helsinki に移った Dr. Yangarber は、知識を自動的にパターンおよび領域オントロジーとして特定するための PULS システムの研究を開始した。2006ー2008 年に、Department of Computer Science による出資の下、PULS のプロトタイプ版が運用された。MedISys との連携は 2007 年に開始された。このインターネットを介したデータ処理システムの統合では、PULS が比較的ハイエンドの言語理解にコンポーネントを提供し、MedISys が文書入力および早期段階のトピックのフィルタリングを提供する。これによって 2008 年には、フィンランドにおける技術革新のための資金提供機関である TEKES が、PULS に研究助成金を提供することとなった。

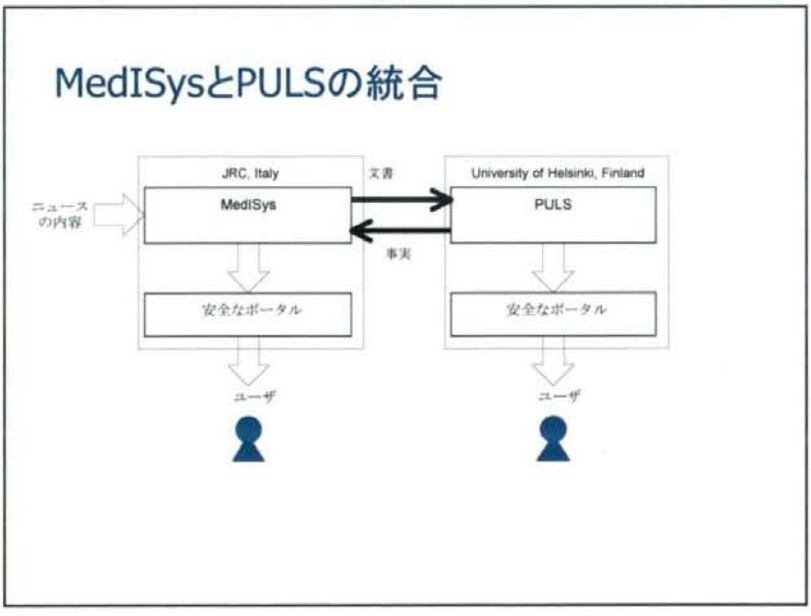

図 6：MedISys と PULS の統合

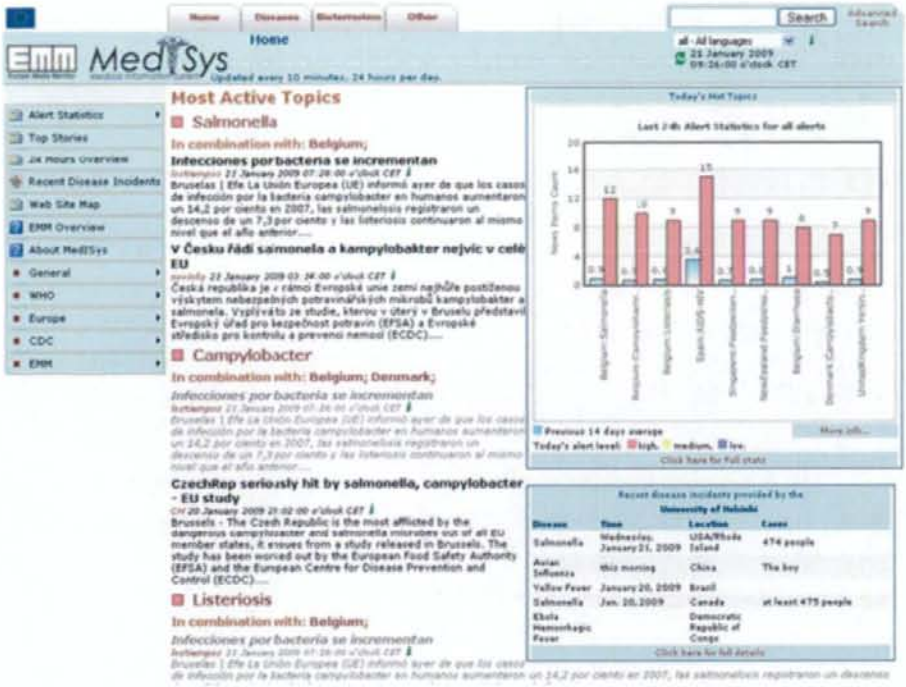

図 7：MedISys ウェブポータル。分類された公衆衛生記事へのリンクが表示されている。PULS から得た疾病発生分析が画面右下の「Recent disease incidents provided by the University of Helsinki」に表示されている。

現在のユーザ

MedISys へのアクセスには、(1) 一部のニュース報道、地図、および収集された警報に対する一般公開アクセス、(2) 欧州委員会外の公衆衛生コミュニティのためのパスワード制限されたアクセス、および (3) 欧州委員会内の公衆衛生コミュニティのためのパスワード制限されたアクセスの 3 つの特権レベルがある。パスワード制限されたユーザが利用可能なサービスには、核あるいは化学汚染などのより広範なカテゴリが含まれる[5]。レベル 3 は、情報サービス企業である Lexis-Nexis からのニュースソースを含む、広範囲のニュースソースへのアクセスを提供している。

MedISys の一般に公開されているページには、平均で 1 日 1,000 名のユーザがアクセスしており、さらに多くが自動メールおよび RSS サービスを利用している。MedISys のパスワード制限されたエリアのユーザには、欧州委員会 (例：DG SANCO)、WHO、European Centre for Disease Control (ECDC)、さらに、フランスの Institute de Veille Sanitaire (INVS)、スペインの Instituto de Salud Carols III、カナダの Global Public Health Intelligence Network (GPHIN)、および米国の疾病対策センター (CDC) などの 20 の公衆衛生当局および国立機関が含まれる[3]。

また MedISys は、Health Emergency Disease Information System (HEDIS) と呼ばれる JRC

が開発した他システムをリアルタイムで更新している。なお、HEDIS は、WHO および World Organization for Animal Health（OIE）を含む多くのソースからの情報を統合するウェブベースのポータルである。

PULS は、MedISys との連携に加えて、European Centre for Disease Control（ECDC）およびフィンランドの National Department of Health（KTL）などの EU 内の他組織に対してもサービスを提供している。

2.3.2 対象範囲

言語

ウェブポータルからは 26 言語しか利用できないが、40 すべての言語のニュースが定義されたカテゴリについて処理されている。ポータルから利用可能な 26 言語は以下の通りである。アラビア語、ブルガリア語、中国語、クロアティア語、チェコ語、デンマーク語、オランダ語、英語、エストニア語、フィンランド語、フランス語、ドイツ語、ギリシア語、ハンガリー語、イタリア語、ラトビア語、ポルトガル語、ルーマニア語、ロシア語、スロバキア語、スロベニア語、スペイン語、スウェーデン語、リトアニア語、マルタ語、およびポーランド語。多言語オントロジーを使用することで、ユーザが入力した関心トピックが自動的に他の言語に展開されて複数言語のニュース報道が検索されるため、検索語の数を大きく増やすことができる。

PULS は現在英語に対応しているが、近い将来においてフランス語およびスペイン語にも対応する計画がある。

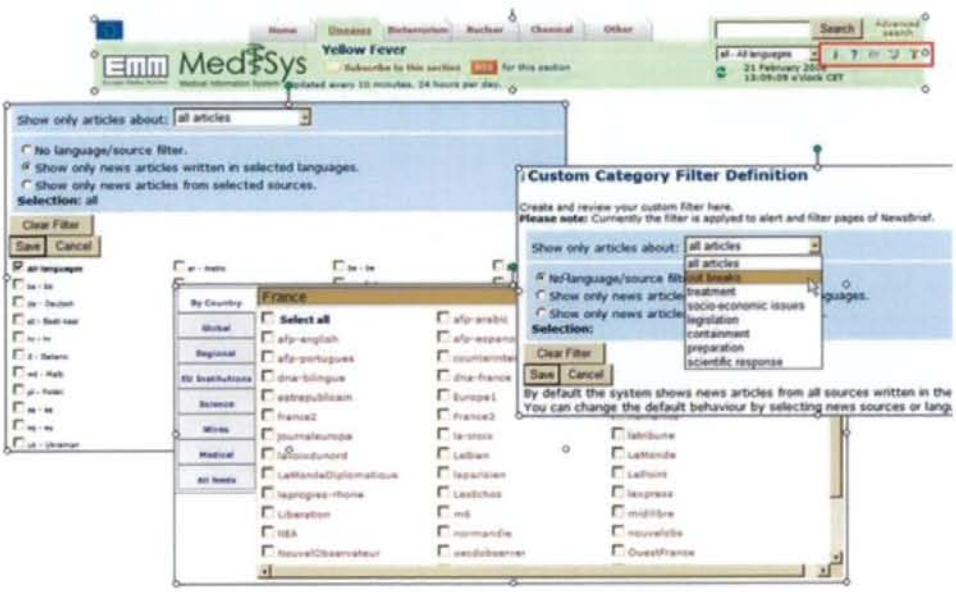

図 8：ニュースのフィルタ設定をカスタマイズするための MedISys ユーザインタフェース

疾病

MedISys は、約 300 の保健関連のカテゴリにおよぶ様々な CBRN 脅威について対象としている[3]。これは、何千もの検索語、さらに、何百もの場所などの汎用カテゴリに相当する。MedISys におけるニュース記事の分類は、とくに疾病アウトブレイクに集中している訳ではなく、疾病、治療、予防接種および他薬剤、方針、公衆衛生組織、および症状などの関連概念が含まれる。これらの概念は、複数の言語における医学的表現、およびニュース報道で使用される可能性があるそれらの専門用語の一般的な表現として与えられる。合計で、人間の健康に影響を及ぼす可能性がある動物病を含めた約 10 の主要な症状と 99 の症状が対象とされている。さらに、MedISys には、疾病が特定されない報告のための「unknown」疾病カテゴリがある。

図 8 の「Custom Category Filter Definition」ウインドウに示されるように、MedISys では、ユーザが柔軟にフィルタ設定を変更することができ、疾病のアウトブレイクに関心があるユーザは罹患患者、入院患者、あるいは死亡患者などの任意のニュース報道を検索することができる。これは、MedISys からトピック フィルタリングされた報告を毎日受け取るユーザに対する負担を軽減する上での最大の利点であると考えられる。

主に、病原体および状態に注目した PULS オントロジーには、ProMED-mail などの一般に公開されている公衆衛生に関するニュースを収集したものから収集・確認され、さらに、手動で整理された広範囲に及ぶ専門用語が含まれている。病原体に関するルート用語の合計数は約 1,000 語と推定される。さらに、PULS オントロジーには、少数の症状および生物体の一般表現が含まれている。

その他の公衆衛生上の脅威

前述の通り、MedISys はあらゆる CBRN 脅威を対象としている。

地理

MedISys は、対応する言語による市、町、村、州、県、省、地域、国などの名称が含まれた多言語辞書を使用することで世界中の地域を網羅している[8]。地名の多くは曖昧であるが、MedISys は発見的規則を適用してこれらに対処している。例えば、最高粒度でロンドンを選択すると、ロンドン村よりも首都ロンドンが優先される。テキスト中の他の地名も、重要な手がかりとなる。例えば、テキスト中にオーストラリアが含まれている場合、「Camden」は、ニューサウスウェールズの町と判断される。

PULS の地理的知識は、CIA ファクトブック[9]などの一般公開されている大きなデータベースから収集されている。対象範囲は、国ごとに粒度が異なるが、ほぼ全世界を網羅している。例えば、米国の位置情報は極めて詳細に対応しており、場合によっては、特定の建物、あるいは、湖あるいは橋などの地理学的特徴まで対応する。

2.3.3 方法

データソース

Europe Media Monitor プログラムは、MedISys にデータを提供する効率の高いウェブ巡回エンジンを開発している。現在、ProMED-mail を含む 43 言語の 4,000 ニュースサイトから 80,000ー90,000 記事／日のニュースが処理されていると推定される。これには、Lexis-Nexis などの情報サービス会社が提供する約 20 社の通信社からのニュースが含まれる。ニュース記事の約 6%が EMM 警報カテゴリの少なくとも 1 つに一致すると推定されている（下記の考察を参照）。ソー

スは、地理的対象範囲および欧州のニュース対象範囲を最大にするように選択されている。ソースデータの入力には、RSS フィード、さらに、独自に開発した専用の画面スクレイピングソフトウエアなどが使用されている。

PULS は安全なインターネット接続を介して、頻繁に更新されるニュース見出しデータの送信に広く使われているウェブフィードフォーマットである RSS フィードを使用して、短いサイクルで MedISys からデータを取得している。

ハードウエア：MedISys

ハードウエア要求は、処理されているデータの規模に比較しても控え目である。[5]において、MedISys 単体での実行には、約 3 台のマルチコアサーバと同等なシステムが要求されると報告されている。以下に示されるように、これは主にトピック分類および警告に軽量なアルゴリズムが使用されているためである。

ハードウエア：PULS

PULS は、Helsinki University の Department of Computer Science が所有する Linux クラスタ上で運用されており、大学の技術スタッフによってハードウエアの保全が行われている。

アルゴリズム：MedISys

図 9 に示されるように、MedISys は、様々な脅威および国別による動向のキーワード分類法および統計学的分析を使用する全自動の言語学的に軽量なアルゴリズムによって特徴付けられる。

データ入力では、ニュース記事の見出しが過去に分析した見出しと比較され、新しい記事のみが受け入れられる。特別に開発されたアルゴリズムを使用してダウンロードされたウェブページからノイズを削除し、その後、テキストはユニコードに変換される。これにより、システムプロセスを通じた複数の言語の標準規格データのエンコードが可能になる。

同一トピックが様々なニュース提供者から報道される傾向があるため、MedISys はクラスタリングアルゴリズムを使用して、同一トピックの内容を 8 時間以内にグループ化している。クラスタリングアルゴリズムは言語に依存しないため、新しい言語および領域に柔軟に適応できる。

確かな記事が公衆衛生上の脅威の対象範囲を保証するために開発された何百ものブールクエリを使用してリアルタイムに選択される。クエリ言語は、屈折的なばらつき、否定、単語の類似性、さらに、正負の重み付けに対応する。また、ログインユーザが独自のクエリを指定することができる。

収集された統計情報は、原則、すべての脅威カテゴリに適用することができる。現在これは、MedISys が特定カテゴリにおける想定外の急上昇を過去のベースラインと比較して検出できるように、疾病および場所情報について算出されている。特定の疾病－位置ペアに対する警告ベースラインは過去 2 週間のデータを使用し、過去 24 時間における疾病－位置頻度に対する有意レベルを計算して、短期のイベントサイクルで算出されている。警告には、低、中、高の 3 つのレベルがある。高レベルの警告は、頻度に週末の影響を調整するための補正因子が適用された 2 週間の平均から少なくとも 3 つの標準偏差がある場合に発令される。MedISys のアプローチの重要な利点の 1 つは、統計がすべての言語ソースにわたって集計されている点であり、これにより、希薄なデータに関連した問題を解決し、ユーザの言語能力外のソースからの警告へアクセスする事ができる。

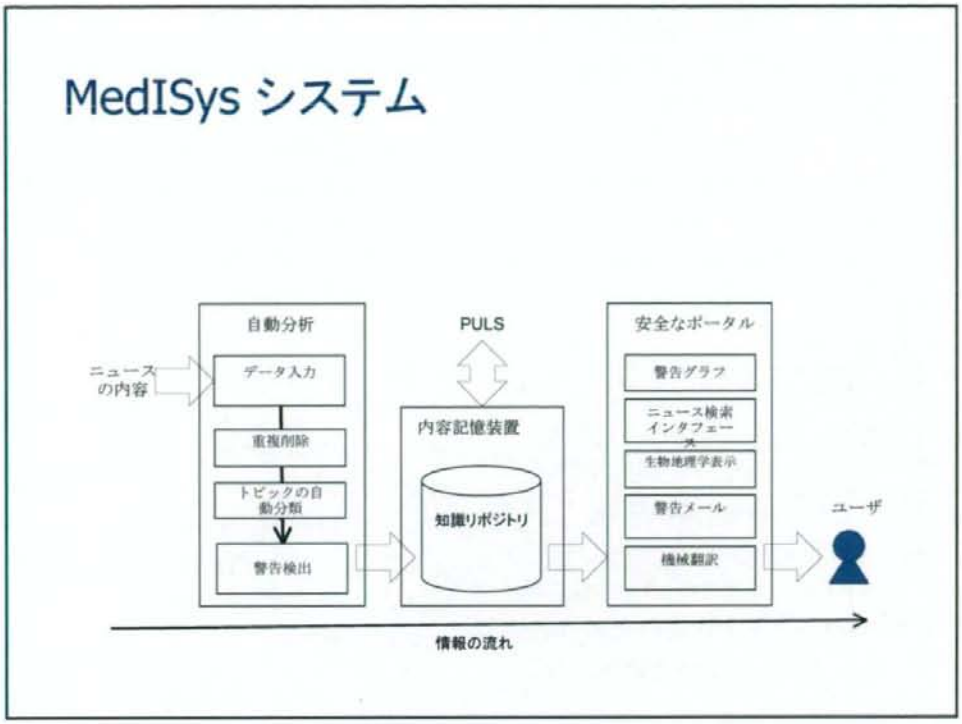

図 9：MedISys システムにおける情報の流れ

MedISys は、登録ユーザに対して、最新ニュースの表示、検索機能、生物地理学マップ、警告グラフ、RSS フィード、KML（Google Earth）フィード、および保健警告メールを含む様々なインタフェースを介して情報を提供している。初期設定では、最新ニュース表示は、情報を保健上の脅威および領域別に分類表示する。インタフェースおよびデータ視覚化ツールは、密接に統合されていることが特徴である。これにより、ユーザは、原文検索、グラフおよびマップの間でシームレスに切り換えることができる。ユーザに提供される機能の特徴には以下が含まれる。

- イベント発生位置がハイライト表示される世界地図
- 警告カテゴリにおける疾病－位置別に収集されたニュースの数を示すグラフ
- 過去 24 時間における有意な疾病－位置ペアを表示するグラフ
- 各地域における警告の統計
- 言語、疾病、あるいは位置によるニュースのフィルタリング
- 病原体用語についての Wikipedia エントリへの広範なリンク
- 「アウトブレイク」「治療」「法律」などの直交カテゴリによるフィルタリング
- Medical Subject Heading（MeSH）オントロジー用語[10]へのリンク

- 人、組織、および検索語などの特定エンティティのニュース記事中の表示

グラフデータを含むすべてのニュースは、10 分ごとに更新されている。

ニュースのユーザ言語への翻訳は、ユーザにリクエストされた場合、一般に公開されている Google 翻訳を使用して行われる。

また、JRC は、登録ユーザに対して、Rapid News Service（RNS）と呼ばれているサービスを介して特定の疾病、国、または言語などの特定トピックの情報を提供している。ユーザは、この情報を使用して独自のニュースレターを作成することができ、作成したニュースレターをオンラインで公開したり、電子メール、SMS、あるいは RSS を使用して配信したりできる。

アルゴリズム：PULS

PULS は、MedISys がもたらし潜在的な警告として特定した記事に、高精度のテキストマイニング分析を追加する。図 10 に示されるように、PULS は、MedISys から文書を取得し、疾病アウトブレイクに関する構造化された事実を抽出し、情報をデータベーステーブルに入力し、その後、ユーザインタフェースにおいて、ハイライトされたテキストへのアクセスを許可する[3]。

PULS の特徴の 1 つは、信頼性スコアを抽出イベントに組み込めることである。これは、特定のニュース報道からのイベントフレームにいくつの候補がスロットに当てはまるかを測定することで行われる。そのイベントに多くの異なる候補が適用可能な場合、イベントの信頼性レベルは低下する。

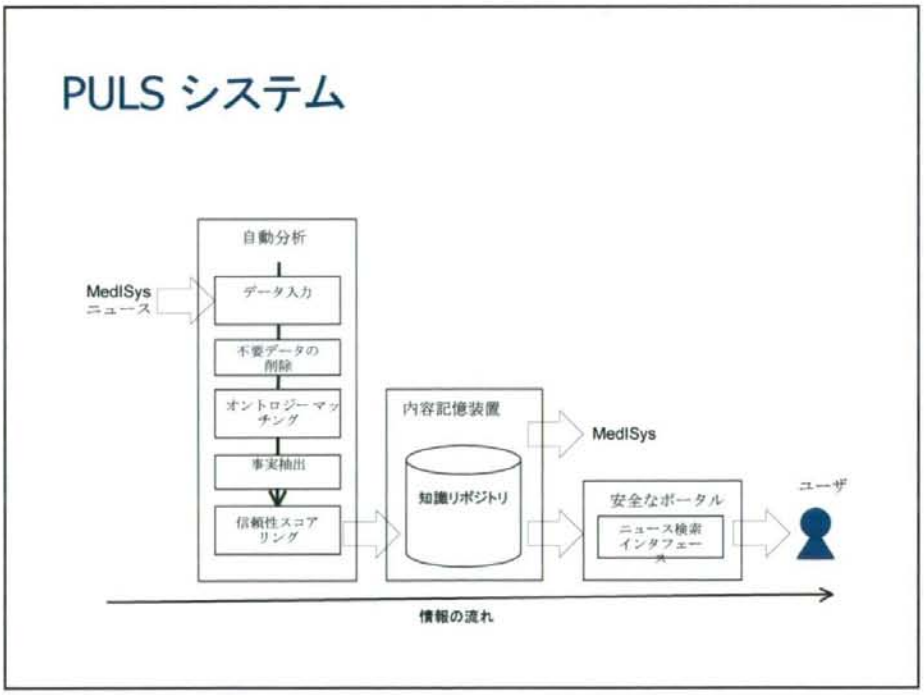

図 10：PULS システムにおける情報の流れ

図 11：ニュース報道クラスタの位置を表示している生物地理学マップ

2.3.4 考察

MedISys の開発は継続して急速に進められており、重複文書の特定など、他の Joint Research Centre メディアモニタリングアプリケーション、および生物地理学マッピング（図 11 参照）などのヒューマンインタフェース技術の恩恵を受けている。

PULS チームは、近い将来にフランス語およびスペイン語に対応することを予定している。学術的観点からは、システムが研究を前進させることが重要であるーこれに関するチームの主要戦略は、テキストマイニングシステムにおけるオントロジーの拡張およびパターンの適用の自動化に機械学習を検討することである。これらは、知識獲得のテーマの下に集約され、人間の作業負荷を軽減することによって将来のテキストマイニングシステムの拡張を高速化することを目的としている。

参考文献・資料

[1] Joint Research Centre (2009), the MedISys system is available from http://medusa.jrc.it (last accessed 17th February 2009).

[2] Joint Research Centre, (2009), "Overview of the Europe Media Monitor", available from http://press.jrc.it/overview.html (last accessed 17th February 2009).

[3] Steinberger R., Flavio F., van der Goot, E., Best, C., von Etter, P. and Yangarber, R. (2008), "Text Mining from the Web for Medical Intelligence", In: Françoise Fogelman-Soulié, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds.): Mining Massive Data Sets for Security. IOS Press, Amsterdam, The Netherlands.

[4] Steinberger, R., Fuart, F., Pouliquen, B. and van der Goot, E. (2008), "MedISys:A Multilingual Media Monitoring Tool for Medical Intelligence and Early Warning", in Proceedings of the International Disaster and Risk Conference (IDRC'2008), pp. 612-614, Davos, Switzerland.

[5] Yangarber, R., Best, C., von Etter, P., Fuart, F., Horby, D. and Steinberger, R. (2007), "Combining Information about Epidemic Threats from Multiple Sources", in Proceedings of the Workshop Multi-source Multilingual

Information Extraction and Summarization (MMIES'2007) held at RANLP'2007, pp. 41-48. Borovets, Bulgaria, September.

[6] Best, C., van der Goot, E., Blackler, K., Garcia, T. and Hornby, D. (2005), "Europe Media Monitor – system description", EUR, Technical Report 22173 EN.

[7] Grishman, R., Huttunen, S. and Yangarber, R. (2003), "Information Extraction for Enhanced Access to Disease Outbreak Reports", J. Biomedical Informatics, Elsevier, v. 35, pp. 236-246.

[8] Atkinson, M., Piskorski, J., Pouliquen, B., Steinberger, R., Tanev, H. and Zavarella, V. (2008), "Online-Monitoring of Security-Related Events", in Proc. COLING 2008: Companion Volume – Posters and Demonstrations, pp. 145 – 148, Manchester, UK, August.

[9] Central Intelligence Agency (2008), "The World Factbook", available online at https://www.cia.gov/library/publications/the-world-factbook/ (last accessed 29th January 2009).

[10] Lowe, H. and Barnett, G., (1994), "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches", in J. American Medical Informatics Association (JAMIA), 271(14): 1103-1108.

民間システム

2.4 HealthMap （Harvard-MIT Division of Health Sciences and Technology、米国）

2.4.1 背景

システムの歴史

2006 年 9 月から稼働している HealthMap プロジェクト[1-4]は、一般に無料で公開されている自動サーベイランスツールであり、年中無休、ウェブベースのソースからほぼリアルタイムでニュース報道を収集している。プロジェクトは、Dr. John Brownstein が責任者を務める Harvard-MIT Division of Health Sciences and Technology（米国）の Children's Hospital Informatics Program によって運用されている。HealthMap の主要な目標は、(a) 一般集団から人間や動物の健康の専門家に至る範囲におよぶ多種多様なユーザに様々な感染症のアウトブレイク情報を統合的観点で示すこと[1]、(b) 影響の大きなニュースに焦点を絞ること、および (c) 既存の公衆衛生サーベイランスのソースを補完することである。

2007 年以降、HealthMap プログラムは、専門家ボランティアによる民間のネットワークである ProMED-mail と密接に連携している。2 つのグループの間に発展した協力の 1 例として、HealthMap は、ProMED ウェブページからのニュースを取得し、未処理テキストを構造化された表示に変換している。ProMED-mail の専門家は、HealthMap（図 12）を利用することで、ウェブニュースソースからの風説的な報道に早期にアクセスすることができ、その後、40,000 名のユーザが購読するメインメールリストに報告することができる。これにより、HealthMap は、出力を検証する機会を得る。

技術的観点からは、HealthMap は、マッピングソフトウエア（Google マップ）、オペレーティングシステム ソフトウエア（Linux）、ウェブおよびデータベースサーバ（Apache／MySQL）、さらに Wikipedia を一部利用して地名辞書などの多くのオープン プラットホーム アプリケーションを活用している。

図 12：HealthMap ユーザインタフェース

財源

HealthMap は、今回調査した PULS および BioCaster と同様に民間のシステムであり、National Library of Medicine、National Institutes of Health（NIH）、および Canadian Institutes of Health Research の助成金で開発されている。2008 年に、HealthMap および ProMED-mail は共同で、インターネット検索企業である Google 社が設立した慈善団体である Google.org から 3 年間にわたる多額の研究助成金を取得している。さらに、HealthMap は、開発資金として、Google.org からさらに 3 年間の助成金を取得している。

現在のユーザ

HealthMap は、毎月 20,000 名以上のユーザが利用しているが、これらのユーザには、アトランタにある疾病対策センター（CDC）、Department of Health and Human Services、WHO、Wildlife Conservation Society、さらに、国、州、および地方レベルの公衆衛生従事者が含まれる。民間の HealthMap プロジェクトは、米国内の疾病アウトブレイクに関する情報を収集することが許されているため、Project Argus が実施する国際サーベイランスを補完すると考えられる。

2.4.2 対象範囲

言語

言語対象範囲に関しては、現在、英語、中国語（標準語）、フランス語、ロシア語、およびスペイン語の主要な 5 言語に対応している。

疾病

現在、170 以上の人間および動植物の疾病を対象とすると推定される。

その他の公衆衛生上の脅威

現在、サーベイランスは主に疾病を対象としているが、環境災害および衝突などの発生の早期警戒インジケータもモニタリングしている。

地理

基本的に全世界を対象とするが、下記の通り、調査によって一部の重点地域において対象範囲が制限されていることが明らかにされており、それは今後の開発で対応する予定である。しかしこれは HealthMap に限定される問題ではなく、他のシステムにおいても重要な要件であると思われる

2.4.3 方法

データソース

Google ニュース、ProMED-mail、さらに、WHO および EuroSurveillance などの公的機関からの公式の報告を含む 14 の主要なインターネット上のニュースソースから取得されている。

知識源

HealthMap は、様々なソースの手動分析で疾病および地名の大規模な構造化辞書を開発している。辞書は、分析対象となる各言語について個別に開発されている。辞書リストは、実際のニューステキスト中のアウトブレイクの名前および位置を決定するために、大規模な言語依存パターンに統合されている。

警告は、主に、アルゴリズム的であるが、いくつかのデータソースの信頼性を判定するための手動で構築された規則が含まれる。例えば、検証されたソースである WHO からの報告の信頼性については高く重み付けがされ、地元メディアの報道には、専門知識が限定されている可能性を考慮し、信頼性については低く重み付けしている。

人間による分析

現在、すべての報告について、HealthMap サイトにおいて公示後、手動で確認している。これは、プロジェクトチームの 1 名の分析者、

さらに毎日の自動電子メールの外部購読者（ProMED、WHO、および CDC からのユーザ）によって行われている。キュレーションは、オンライン データ管理ツールによって行われる。

ハードウエア

HealthMap のアルゴリズムは、言語学的な分析に関して比較的軽量なことが特徴である。単一サーバ上で実行可能なシステムのため、ハードウエア要求も低い。

アルゴリズム

図 13 に示されるように、HealthMap には 6 つの主要な段階があり、自動分析は 5 つの言語それぞれについて個別に実施されている。

第 1 段階：データ入力。ニュースは、RSS フィードを使用してインターネットから短いサイクルでダウンロードされる。

第 2 段階：トピックの自動分類の 1。ニュース記事は、パターンベースの分類、構造化された辞書の検索、および経験的ルールを使用して国および疾病別に分類される。

第 3 段階：トピックの自動分類の 2。ニュース報道は、ベイズ分析を使用して、速報、警告、コンテキスト、古いニュース、または疾病非関連のニュースに分類される。

第 4 段階：重複削除。関連したイベントについて記述するニュース記事を決定するために、追加的な分析が行われる。

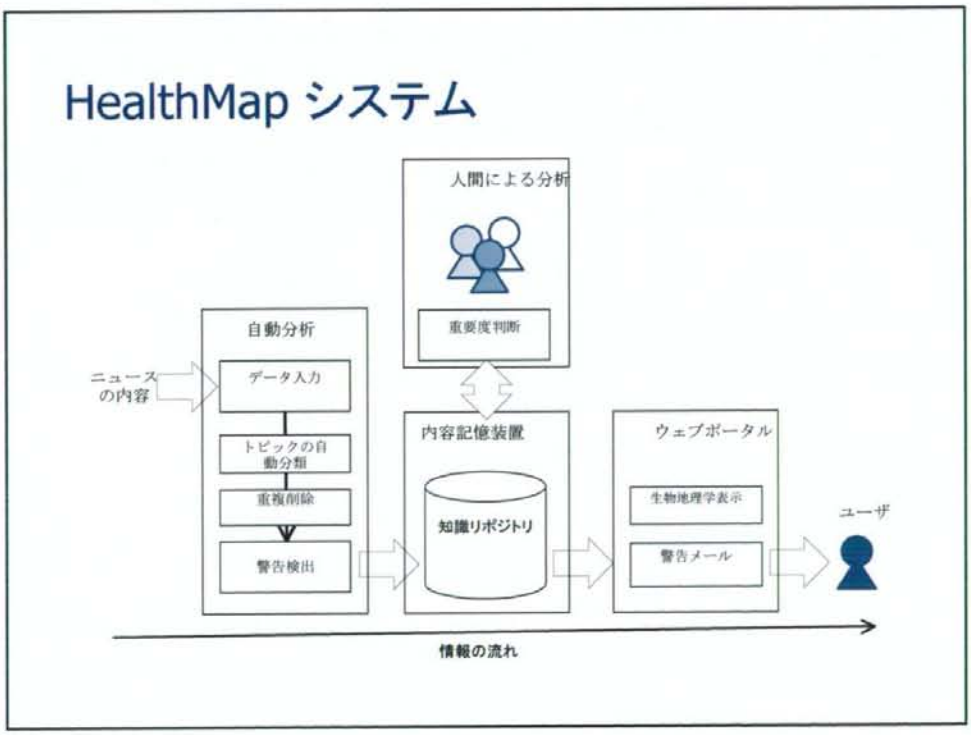

図 13：HealthMap システムにおける情報の流れ

第 5 段階：警告検出。警告では、影響の大きなニュースに注目している。そのため、HealthMap は、統計学的評価のためのコンポーネントがいくつか組み込まれている。これらには、特定の場所におけるソースの数、警告の新しさ、およびニュースソースの信頼性が含まれる。複数のニュースソースによって確認されたイベントに対してもより高い重みが付けられる。

第 6 段階。配信。エンドユーザは、HealthMap のウェブページから、様々なフォーマットの情報にアクセスすることができる。図 12 に示されるように、最も視覚的なコンポーネントは、ユーザへの情報過負荷を解消するための生物地理学マップである。マップは Google マップを使用して構築されている。Google マップは、Google 提供の自由に利用できる軽量なマッピングサービスであり、HealthMap ポータルサイトに組み込まれている。マップには衛星画像およびマッピング情報が含まれており、ユーザは、過去 30 日以内にアウトブレイクが発生した場所の地理的情報を取得することができる。追加機能によって、ユーザは疾病、日付、国、およびデータソース別にフィルタリングすることができる。二重フィルタリングの開発が早期段階のため、マップには冗長情報が表示されるが、ユーザはサイトの別ページから関連した記事にアクセスすることができる。その他の配信オプションには、RSS フィードおよびメールリストが含まれる。

2.4.4 考察

研究は現在、言語処理技術を使用して行っているイベント分析のレベルを深めること、さらに、アラビア語、ヒンディー語、およびポルトガル語を含めて対応言語を増やすことに集中している。HealthMap および他多くのシステムのデータソースについては、アフリカおよび南米などの重点地域における新興疾病に関するニュースの対象範囲が課題である。その他の重点は、精度の向上、報告の検証、情報の地理的解像度の向上、さらに、各言語の訓練セットのために人的ネットワークの組み込みである。

参考文献・資料

[1] Brownstein, J., Freifeld, C., Reis, B. and Mandl, K. (2008), "Surveillance san frontières: Internet-based emerging infectious disease intelligence and the HealthMap project", in PLOS Medicine, 5(7), pp. 1019-1024.

[2] Brownstein, J., Freifeld, C., Reis, B., Mandl, K. (2007), "Evaluation of online media reports for global infectious disease intelligence", Advances in Disease Suveillance, v.4, pp. 1.

[3] Freifeld, C., Mandl, K., Reis, B. and Brownstein, J. (2008), "HealthMap: Global infectious disease monitoring through automated classification and visualization of Internet media reports", in J. Amercian Medical Informatics Association 15:150-157.

[4] HealthMap (2009), available from www.healthmap.org (last accessed February 1st 2009).

2.5 BioCaster（国立情報学研究所、日本）

2.5.1 背景

システムの歴史

BioCaster は、年中無休で、グローバル・ヘルス情報をほぼリアルタイムで収集することを目的とした研究用プロトタイプシステムである。このプロジェクトは、国立情報学研究所（日本）の Dr. Nigel Collier が率いる最新テキストマイニング技術の可能性を検討する学術的プロジェクトの一環として進められている。研究の主要な目標は、(a) 文書の意味論的アノテーション アルゴリズムの模索、(b) 自然言語処理技術を向上するための知識の獲得、および (c) 言語学的信号に基づく早期警告システムの検証である。

BioCaster の概念[1,2]は 2006 年にまで遡る。この年、助成金によって、疾病アウトブレイク関連のニュースに意味論的な索引付けを行うための高性能システムが構築された[3]。当初から BioCaster は、アジア太平洋地域における新興および再興の保健上の脅威の潜在的リスクのため、この地域における言語に重点を置いていた。このため 2006 年には多言語オントロジーの構築が開始された[5]。この構造化された公衆衛生語彙集は、コミュニティリソースとして一般に公開されている。

国立情報学研究所における BioCaster の開発は、計算機言語学およびソフトウエア工学を専門とする約 4 名が中心となって行われた。2006 年に、岡山大学（日本）、国立遺伝学研究所（NIG、日本）、Kasetsart University（タイ）、および Vietnam National University（VNU、ベトナム）を含む学術的パートナーのネットワークが急速に確立された。これらのグループの協力により、複数の言語によるソフトウエア工学、公衆衛生、遺伝学、および計算機言語学に関する専門知識が提供された。現在では、システムにはより深い言語理解のためのコンポーネントが含まれ、英語、日本語、および、ベトナム語の文書を処理することができる。

財源

BioCaster は、日本学術振興会（JSPS）および新領域融合研究センターの育成融合プロジェクトを含む国の支援組織からの助成金によって開発されている民間のシステムである。BioCaster は、コンピュータによる保健上の脅威の理解の向上について検討するために、科学技術振興機構（JST）のさきがけプログラムの下、2009 年から 3 年間にわたる助成金を得ている。

現在のユーザ

BioCaster の出力は、いくつかの形で利用することができる。ウェブポータルには、(1) Global Health Monitor と呼ばれる一般に公開されたマッピングおよびグラフ作成インタフェースと (2) パスワードで制限された警告インタフェースの 2 つのモードがある。現在、警告インタフェースは、国立感染症研究所（日本）および Health Protection Agency（英国）の小さなテスト用コミュニティの使用に限定されているが、今後、より広く公開される予定である。さらに、一般に公開されている多言語オントロジー（図 13 を参照）は、8 言語による構造化された用語を提供しており、WHO を始め、北米の 25 組織、アジアの 32 組織、欧州の 14 組織、オセアニアの 1 組織）の世界中の 73 の学術、業界、および公衆衛生関連組織でダウンロードされている。

2.5.2 対象範囲

言語

BioCaster は現在、3 言語（英語、日本語、

およびベトナム語）によるニュースを分析しているが、近々タイ語にも対応する予定である。上記の通り、システムは8言語（中国語（標準語）、英語、フランス語、スペイン語、日本語、韓国語、タイ語、ベトナム語）による多言語オントロジーがあり、将来的には12言語に拡張されてこれらの言語に対応することが可能になる。

疾病

BioCasterの病原体には、アジア太平洋地域、欧州、および北米の主要な国の保健省関連省庁が定める届出疾患によって優先順位が付けられている。現在、認識される120以上の疾病は、主に、人間の健康への脅威となるものであるが、人間の健康に影響する可能性がある一部の動物病も含まれている。

その他の公衆衛生上の脅威

これまでは人間の疾病に重点を置いてきたが、現在、動物の保健に影響する病原体および一部の化学物質に拡大する方法を検討中である。

地理

BioCasterは、アジア太平洋地域に重点を置いているが、国および行政区レベルでの地名のタクソノミーを介して世界中の疾病アウトブレイクを対象としている。なお、タクソノミーには5,000以上の英語による地名が記録されている。これらは、主にWikipediaなどの一般に公開されているソースから抽出された。BioCasterのアルゴリズムは、新しい地名を高い精度で検出することができるが、現在BioCasterには、町やランドマークなどの詳細なレベルの辞書が搭載されていない。

図14：BioCasterのGlobal Health Monitor

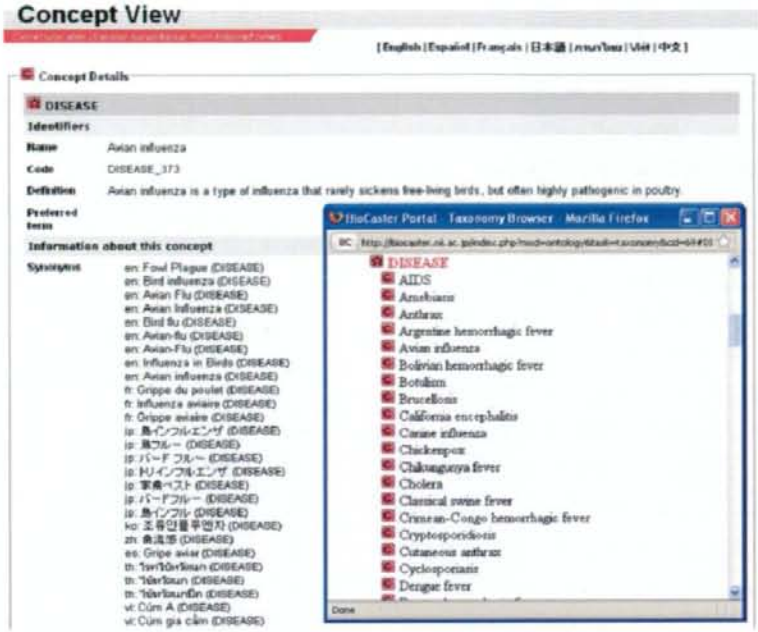

図 15：**BioCaster** のオントロジービューア

2.5.3 方法

データソース

データは、一般に公開されている Google ニュース、News Tin、European Media Monitor、および ProMED-mail を含むニュース、公衆衛生、政府、およびビジネスに関する 1,700 以上の RSS フィードを使用して 1 時間の短いサイクルで取得されている。2009 年からは、ニュースに関して情報サービス会社と契約している。これが、パスワード制限されたポータルを介して、ログインユーザに対して提供されている。

知識源

上記の通り、BioCaster の専門用語の知識源の中心は、8 言語による疾病、病原体、症状、症候群、および生物などの領域専門用語、さらに、特定の宿主種に影響する疾病の原因症状あるいは病原体などの領域に依存する関係を含む多言語オントロジーである。これにより、テキストマイニングシステムは、ニュース報道で明示的に記載されていないギャップを埋めるための領域にある重要な概念および関係に関する基本的な理解が可能になる。オントロジーには、Wikipedia に加えて、Medical Subject Headings（MeSH）、SNOMED CT、および ICD-10 などの外部語彙集およびオントロジーへのリンクに関する情報も多く含まれる。上記の通り、オントロジーには地名専用のセクションが含まれている。現在、これには約 5,000 の地名が含まれている。

文書から知識を抽出するために、BioCaster は、表面的な表現を構造化情報に変換するための一連の規則パターンを維持している。例えば、「man exposes airline passengers to

measles（男性が航空機の乗客をはしかに曝露）」というフレーズは、そのような特別なパターンを使用して「species(human)およびdisease(measles)（種（ヒト）および疾病（はしか））の構造化された表現に変換する必要がある。これらのパターンの記述には時間がかかるため、BioCaster プロジェクトでは、独自の Simple Rule Language（SRL）と呼ばれる軽量な規則言語およびパターン構築用インタフェースを開発している[6]。これらは、オープンソースライセンスの下に研究コミュニティに対して公開されている（BioCaster が使用する規則については開示されていない）。現在、BioCaster は、英語で表記された関心イベントを特定するために、約 4,000 の SRL 規則を使用している。

ハードウエア

BioCaster のテキストマイニングシステムは、カリフォルニア大学サンディエゴ校が開発したクラスタコンピューティング用の Linux ベースのオープンソースプラットホーム上に構築されている[7]。このスループットの高いシステムを使用することで、大量のニュースの意味論的処理を可能な限り短時間に行うことができる。

アルゴリズム

図 15 に示される通り、BioCaster は、モジュール式のプロセスパイプラインを採用している。

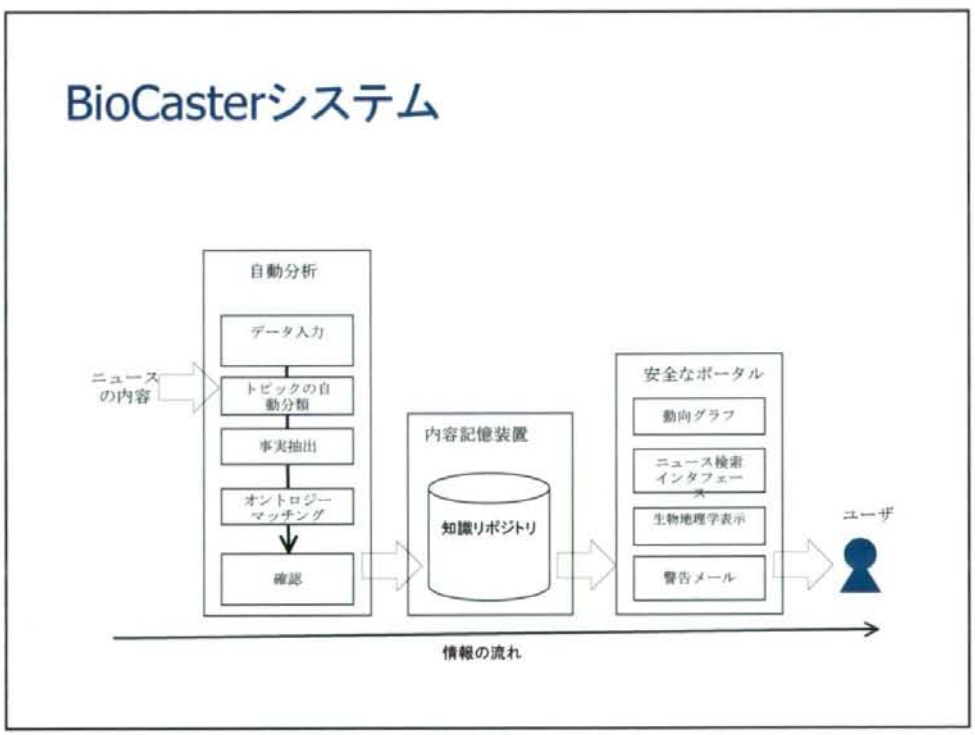

図 15：BioCaster システムにおける情報の流れ

主要な段階は以下の通りである。

第1段階：データ入力。ニュースは、RSSフィードを使用してインターネットから1時間の短いサイクルでダウンロードされる。その後、発見的フィルタを使用して、ニュース報道から他記事へのリンクなどのノイズが削除される。

第2段階：トピックの自動分類。文書の分類は、人間が注釈を付けた関連および非関連文書に基づいて訓練された機械学習アルゴリズムを使用して行われる[8]。このプロセスの指針は、WHOのIHR Annex 2意思決定装置、および公衆衛生に関する研究パートナーとの協議に基づいている。

第3段階：事実抽出。機械学習を使用した用語認識が、疾病、ウイルス、細菌、症状、および場所を含めた18種類の用語について行われる。その後、SRLパターンを使用して、ヘルスイベントに関する事実が特定される。これらには、時間、場所、病原体、海外旅行、薬物耐性、病院従事者の感染、あるいは新しい種類の病原体などの直接観察可能な証拠、ならびに、イベントが公共サービスの混乱に関係するか、あるいは、経済に影響しているかなどの間接的なインジケータが含まれる。事実抽出は、基本的に、下流において起こり得る潜在的な警告についての検索に詳細な証拠を提供することができる信号を特定することを目的としている。

第4段階：オントロジーマッチング。テキスト中の用語は、ニュース記事間、および言語間でイベントに関する理解を統一するために、BioCasterオントロジーの概念と比較される。

第5段階：確認。BioCasterは、出力の手動による統計学的な確認を使用しないため、一般に公開されているウェブポータルページに表示する報告の特異性を向上するために、いくつかの発見的フィルタを使用している。なお、このプロセスでは信頼性スコアは算出されない。発見的フィルタが特異性を低下させる可能性があるため、これらはログインが要求されるポータルページに表示されるニュースには適用されていない。

Biocasterの警告検出は現在、手動で行われている。ポータルのログイン制限されたセクションにアクセス可能なユーザは、重大度インジケータに対応するカスタマイズされたトピックのニュース警告を購読することができる。ユーザは、直接的および間接的なインジケータ、さらに、場所、記事のジャンル（ニュース、ビジネスレポート、公式報告など）に対応する15のトピックカテゴリを使用してニュース警告を設定することができる。

ユーザに提供されるその他の機能には以下が含まれる。

- Google マップを使用した生物地理学マップ
- 警告前ニュースに関する疾病ニュースは、オントロジーの8言語でフィルタリングおよび視覚化することができる。
- ニュース報道の PubMed、HighWire、および Google Scholar などの外部検索エンジンへの動的リンクは、専門家がニュートピックの背景情報の確認に要する時間短縮に役立つ。
- 多言語オントロジーの検索および閲覧、ならびに、外部語彙集およびオントロジーへのリンク

- 地域および国別の過去 1 年間において収集されたニュース報道の数を示す動向グラフ

2.5.4 考察

近い将来、BioCaster は、タイ語および中国語を含むオントロジーを使用したシステムで稼働する予定である。自動警告検出は、報告の確認のためのユーザフィードバックの組み込みと同様に重要な優先事項である。さらに、プロジェクトチームは、結果として得られたニュース記事を各自の言語で要約するために機械翻訳を使用する方法について検討している。

参考文献・資料

[1] BioCaster (2009), available from www.biocaster.org (last accessed February 17th 2009).

[2] Collier, N. Doan, S., Kawazoe, A., Matsuda Goodwin, R., Conway, M., Tateno, Y., Ngo, Q., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M. and Taniguchi, K. (2008), "BioCaster: detecting public health rumors with a Web-based text mining system", Bioinformatics, Oxford University Press, DOI: 10.1093/bioinformatics/btn534.

[3] Collier, N., Kawazoe, A., Son, D., Shigematsu, M., Taniguchi, K., Jin, L., McCrae, J., Chanlekha, H., Dien, D., Hung, Q., Nam, V., Takeuchi, K. and Kawtrakul, A. (2007), "Detecting Web rumours with a multilingual ontology-supported text classification system", Advances in Disease Surveillance, ISDS, vol. 4, pp. 242.

[4] 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.

[5] Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D. Barrero, R., Takeuchi , K.and Kawtrakul, A. (2007), "A multilingual ontology for infectious disease surveillance: rationale, design and challenges", Language Resources and Evaluation, Elsevier, DOI: 10.1007/s10579-007-9019-7.

[6] McCrae, J., Conway, M. and Collier, N. (2009), the Simple Rule Language interface is available from http://code.google.com/p/srl-editor/ (last accessed February 17th 2009).

[7] Rocks cluster computing (2009), available from http://www.rocksclusters.org (last accessed February 17th 2009).

[8] Doan, S., Kawazoe, A., Conway, M. and Collier, N. (2009), "Towards role-based filtering of disease outbreak reports", J. Biomedical Informatics, Elsevier, DOI: 10.1016/j.jbi.2008.12.009).

[9] Intergovernmental working group on revision of the International Health Regulations (2005), "Decision instrument for the assessment and notification of events that may constitute a public health emergency of international concern – Report of the Ad Hoc Expert Group on Annex 2", published by the World Health Organization, available from http://www.who.int/gb/ghs/pdf/IHR_IGWG2_ID4-en.pdf (last accessed February 19th 2009).

結果および考察

3．日本における国際的感染症サーベイランスの現状