プロジェクト名： 感染症への適切な対応の基盤となる多言語オントロジーの開発

１．プロジェクト体制

研究代表者
［国立情報学研究所］ Nigel Collier
共同研究者
［国立情報学研究所］藤山秋佐夫
［国立感染症研究所］谷口清洲　重松美加
［岡山大学］　竹内孔一
［カセサート大学（タイ）］ Asanee Kawtrakul
［ベトナム国立大学（ベトナム）］ Dinh Dien
［国立遺伝学研究所］ 舘野義男

２．研究目標

概要

アジア太平洋地域における監視システム基盤の欠如は、最近のトリH5N1の流行など、感染症の急速な蔓延に対する包括的コントロールを妨げると考えられる。その地域における監視の改善の一環として、BioCasterプロジェクトは、いくつかの地域言語で自動的にインターネットニュースや他のオンラインソースをモニタリングするため、オントロジー中心のテキストマイニングに基づくシステム開発を試みている。そのシステムの中心はオントロジーであり、混在した事実の高度検索を可能にし、イベントの優先度を評価するためにシステムによる知的推論を可能にするうえで役立つ。オントロジーは、我々のニーズに対して必要な言語範囲または意味的特異性のない既存の分類体系を拡張する。本報告では、我々のニーズを概説し、優先病原体とそれらに起因する疾患に焦点を当てた、新しい概念構造および多言語用語資源を構築する根拠と方法に関して詳細に検討する。オントロジーは8言語で病原体102種の役割に注目し、オンラインデータベースおよびダウンロード可能なWeb Ontology Language（OWL）ファイルとして自由に利用できる。昨年に初めて公開して以来、オントロジーが世界中で73以上のグループによりダウンロードされ、毎月ポータルにアクセスする平均1,500人のユーザーにより定期的に参照されていると推定される。オントロジーは、ウェブ上で自由に入手できる非専門用語の唯一の多言語公衆衛生オントロジーとなることにより、テキストマイニングおよび公衆衛生の分野に対して特異的に寄与している。

背景

ヒト（SARS）および動物（トリインフルエンザ）における最近の伝染病は、アジア太平洋諸国の疾病監視に明らかな差があることを示している。感染症監視は、そのような急速に蔓延する伝染病に対して予防の中核となるべきであるが、時宜を得た情報が欠如すると、公的

機関の管理努力が妨げられると考えられる。2008年の黄熱病およびトリインフルエンザの時期に関して図1,2のデータが示すように、BioCasterプロジェクトは、インターネットニュースや学術文献からの集団発生監視に対するテキストマイニングシステムを開発しており、感染症の集団発生が急速に拡大する可能性のある集団を認識するうえで公衆衛生の専門家を支援することができる。全体的な有益性とは、脅威への認識を高めることおよび情報に基づく介入を実施するために不確定性を低下させることである。

カナダのPublic Health Agencyが運営しているGPHINシステムは現在、国際的に蔓延する伝染病の早期発生を監視する数少ない積極的監視システムの1つである。このシステムは最先端の監視システムを代表するものであり、SARS（重症急性呼吸器症候群）の流行を最も早期に察知した、世界保健機構（WHO）が信頼しているシステムである。しかし、そのシステムには現在、日本語、韓国語、タイ語、およびベトナム語を含む一部の言語で専門用語を網羅していないなど、いくつかの制限がある。また、システムの知識源は公共的に利用できず、それらを検討または拡張するユーザーの能力を制限している。

死亡および罹患を最小限に抑える情報の価値として適時性が重要な要素の1つであることを考慮すると、地域言語処理能力が必要であることは明らかである。集団発生あるいは発生は地域メディアで最初に公共の場で言及されると考えられるが、それが仮に公表される場合、さらに時間が経過してからそのニュースは国際メディアで翻訳され、公表される。

BioCasterの中核は、テキストマイニングシステムで計算可能なセマンティクスとして役立つ多言語オントロジーである。オントロジーは、混在した事実を高度に検索できるように、また警告を自動的に発信するため、イベントの優先度を評価するうえでシステムが知的推論を実行できるように役立つ必要がある。しかし、そのことはプロジェクトの早期で明確にされていたが、既存の分類体系には我々のニーズに対して必要な言語範囲または意味的特異性が存在しない。

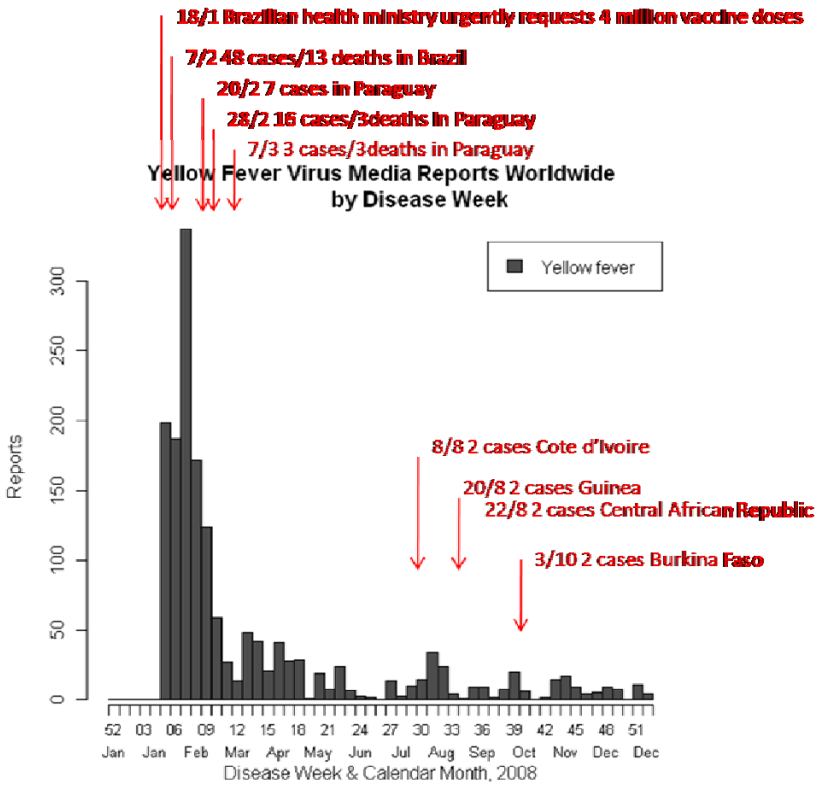

図 1　BioCaster は 2008 年の黄熱病に関連するメディアの報告を察知した。WHO 集団発生報告の通知は赤で示した。

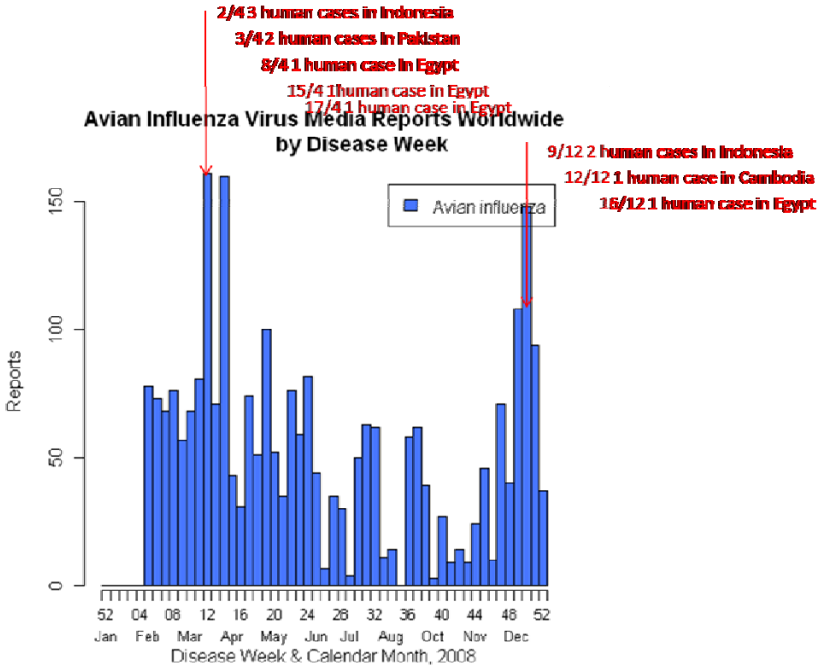

図 2　BioCaster は 2008 年のトリインフルエンザに関連するメディアの報告を察知した。WHO 集団発生報告の通知は赤で示した。

例えば、意味情報を用いたそのようなオントロジーの検索能力を説明する将来のシナリオを次のように想定する：公衆衛生の専門家は、ベトナムにおいてH5N1トリインフルエンザウイルスが再集合する可能性を調査することに関心がある。BioCasterポータルにログインし、ベトナム、日付範囲および英語ニュース記事の内容と共にH5N1トリインフルエンザを検索対象用語として入力する。内部的にBioCasterは、最初の用語が疾患概念階層である高病原性H5N1トリインフルエンザにおいて基本語の英語異表記であり、様々なクエリーとして使用されるH5N1疾患、HPAI（H5N1）およびA（H5N1）インフルエンザなどの英語同義語が存在することを認識する。検索は実行されるが、結果はユーザーが必要とする情報とは無関係である。次にシステムは、ユーザーに<hasSymptom>関係子に基づく関連症状および<causedBy>関係子により認知される病原体を用いて検索の選択幅を提供する。ユーザーはこの選択肢を選択し、咳、肺炎および急性呼吸窮迫などの症状と、病原体名すなわちA型インフルエンザウイルスH5N1亜型を用いて検索を再実行する。この時点で、記事は発見されるが、その報告は既に2週間経過しており、場所名に関する極めて重要な情報が失われている。次いで、ユーザーはベトナム語のニュースを発見するためにベトナム語に相当する用語で検索を再度実施する。システムがベトナム語ニュースを検索した後、基本語に対する<SynonymOf>関係子に従って混在した情報を要約する各イベントに対して英語で構造翻訳文が生成され、それらから英語で<preferredTerm>が認知される。各用語は<preferredTerm>型で得られ、イベントの比較を容易にする。次に専門家は、場所名が明らかに特定されている場所を検索するイベントを得る。このシナリオにおいてシステムは、意味的に関連する用語のクエリーを拡張することにより、ユーザーが関連情報を迅速に見つけ出し、言語の壁を越えることを支援する。

３．平成２０年度の研究進捗

結果

図3のウェブポータルビューが示すように、BioCasterオントロジーは公衆衛生用語の構造化した専門用語を8か国語で提供する。2007年10月に最初にリリースするBCOの目標は、定義と関係を有する50の基本語（同義語クラスター）を構築することである。上記で述べたように、2回目のリリースではこれを102基本語まで拡張することである。動物からヒトへ感染する可能性のある伝染病（人畜共通疾患と呼ばれている）は今後、新たな大規模感染を引き起こす可能性が最も高く、これらの伝染病を網羅する必要があると考えられることから、第2版では動物およびヒト病原体が含まれている。これは確立された分類と比較して適当であるが、8か国語にわたる数千の表面レベルの用語を我々に提供し、1つのドメインおよびアプリ

ケーションに注目したコンパクトな構造が得られる。これに続いて、我々は毎年オントロジーおよび用語バンクを拡張できると考えている。

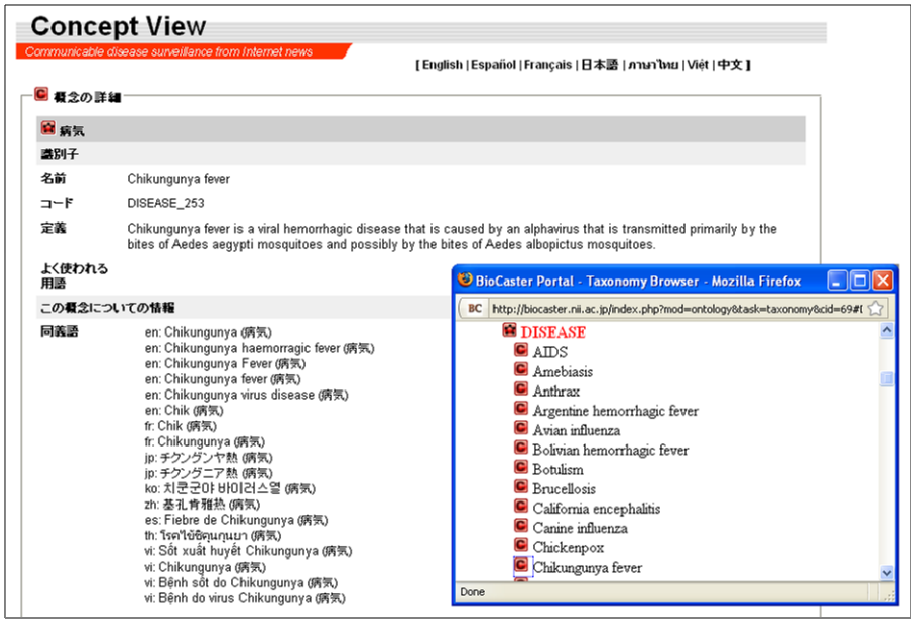

図３　８か国語で用語を示している多言語オントロジーのウェブポータルビュー

Dr. Ai Kawazoe（以前はNII、現在は筑波大学に所属）との協力により、イベント階層の形式でオントロジーに新しいレベルの概念化を追加することもできた。この概念化のレベルは、流行時に発生しうる感染症の集団発生のタイプを説明するために必要である。例えば、死亡、入院あるいは病院閉鎖である。

オントロジーの評価は、評価指標がないうえ「優れた」オントロジーのあるべき姿の概念が様々であるため、非常に困難である。我々は最初に、用語のコーパス収録範囲を用いて従来の方法とは異なる評価方法で評価した。32 週間の評価期間中、我々は適切な 29,443 件の報告で言及されている英語用語の割合を観察し、1 回以上出現する疾患用語の 78%、病原体用語の 69%、および症状用語の 76%が網羅されていると推定した。

2009 年 2 月までに、（イベントオントロジーを除く）主要な疾患オントロジーは、世界保健機構（WHO）や北米（25 団体）、アジア（32 機関）、欧州（14 機関）およびオセアニア（1 機関）に位置する他の機関など、世界中で 73 以上の研究、産業および公衆衛生機関によりダウンロードされた。

我々のシステム内で多言語公衆衛生オントロジーを利用するため、そのオントロジーをBioCasterポータルウェブサイト上の検索インターフェースに組み入れた（図4）。これにより、現在ユーザーは8か国語（中国語、英語、フランス語、日本語、韓国語、スペイン語、タイ語、

ベトナム語）で概念および専門用語を検索でき、他の言語での訳語を見ることができるうえ、他の概念から概念定義や関連情報を参照することができる。また、我々は、ユーザーが疾患、症候群および場所に基づいて様々な言語で新しい報告を選別できるように、そのオントロジーを検索インターフェースに組み入れた。

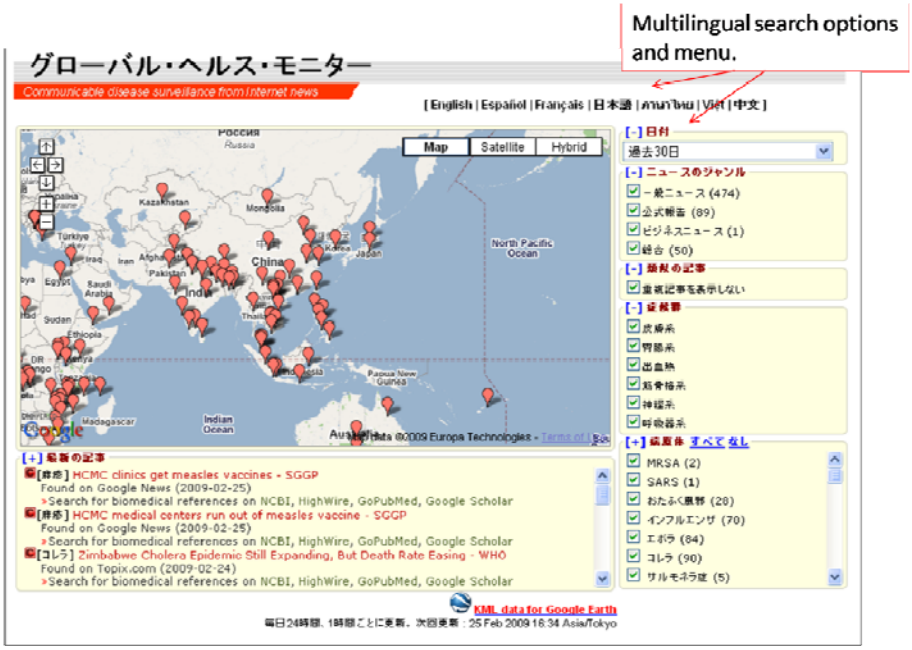

図４　７か国語でオプションを示した検索インターフェースのウェブポータルビュー（日本語オプションを選択した場合）

方法論

BCOにおいて、我々は最初に6種類の言語、すなわち中国語（簡体字）、英語、韓国語、日本語、タイ語およびベトナム語を対象とした。Version 2では、フランス語とスペイン語を、主要な国際語として考慮されたことから追加した。実質的に非語彙化ドメイン非依存クラスから構成されている上位レベル構造を数理言語学者が構築した後、言語学者がOntoClean法を用いて解析したドメイン依存エンティティクラスに対応するリーフクラスを構築した。

次いで、用語を、疫学者、遺伝学者および数理言語学者の協力を得て生物学者が英語、韓国語、日本語および中国語で収集した。 ベトナム語やタイ語の専門用語サポートはこれらの言語に堪能な言語学者が行った。

ドメイン依存クラスおよび関係において、EuroWordNetプロジェクトと広範に類似する作業工程に従った。我々は最初に、保健省のウェブサイト上の届出疾患リストから収集した優先病原体のリストに基づき、オントロジーフラグメントを特定する。これは、最も価値のある専門資源に

集中するためにデザインされている。優先病原体リストにより、必然的に語彙有効範囲の仕様が決まり、病原体が宿主で引き起こす疾患やそれが示す症状などの専門用語が収集できる。次に用語はコード化され、それらの同等性および連想関係が確認される。これに続いて、我々は品質検査を行い、意見公募および公共評価のため新しいバージョンをリリースする。第2段階では、オントロジーフラグメントを比較、媒介および再構築する。

ツール支援では、記述論理形式にエクスポートし、バリデーションの推論機構との統合を可能にするWeb Ontology Language (OWL) プラグインと共にProtegeオントロジーエディターを用いてBCOを開発する。

有効範囲

オントロジーの重要な要素は使いやすい分類階級である。語彙および関連性の有効範囲は、数理言語学者とドメイン専門家が協議して決定した。感染症監視に対していくつかのシナリオを示した。優先度の高いシナリオは、(a) 病原体の動物-ヒト感染から限定的または持続的なヒト-ヒト感染までの移行期間、(b) 国境を越えた感染性および病原性病原体の蔓延、(c) 病原性病原体のヒト集団への人為的散布である。WHO会議報告書が裏付けているように、我々の考察から、個々の症例よりむしろ異常クラスターの検出および追跡に注目する必要があることが判明した。

ウイルスDNA/RNAおよびそれらと宿主遺伝子との相互作用は、病原体に対する感受性あるいは耐性を決定する際に重要な役割を果たしていることから、遺伝疫学により他の局面が情報ニーズに追加される。したがって、我々は、病原体と遺伝子およびそれらの産物を含む宿主に関してさらに詳細なレベルで追加する予定である。BCOにおいてそのような情報を取り入れることの背後にある戦略は、各病原体の全体像をその寿命期間に関して取得すると共に、専門家の参照能力を増強し、MEDLINEにおけるライフサイエンス文献データベース内の論文を理解する可能性を向上させることである。

デザイン

BCOの上位レベルは、OWL仕様のSuggested Upper Merged Ontology (SUMO) から採用した基本オントロジーから構成されている。SUMOオントロジーは、他のオントロジーとの潜在的な統合源を与える下位クラスであるAttribute、Quantity、ObjectおよびProcessと共にEntityなどの非常に一般的なクラスを提供する。またSUMOは、SUO-KIFおよびOWL fullにおいてより広範な分類体系および豊富なaxiomizationを有している。我々の目的に合わせ、不必要な詳細部分を除外するため、全てのnon-leafクラスに2つ以上のchildrenを持たせることでSUMOの階層を簡易化した。

BCOの中位レベルは、ターゲットエンティティクラスのdisjointセットから構成されている。これらは、自動用語認識およびグラウンディング手法を用いて達成できる粒度のレベルを考慮して選択される。この結果、表現関係を有する比較的浅いオントロジーが得られる。

前述したように、我々は<synonymTerm>関係子により言語固有用語と結びつく、および<preferredTerm>関係子によりこれらの各言語の優先語を結び付ける言語の中で、および言語の間で言語間ピボットとしての役割を果たす基本語の概念を採用した。これを図5に図示した。各基本語は、固有識別子、定義、エディタノート、スコープノートおよびICD10、MeSH、SNOMED CTやWikipediaなどの外部語彙や資源に対する様々なリンクを含むさらなる特性を有している。言語固有用語は、固有識別子、ISO639言語識別子、およびそれが略語か口語表現かなどの特性を有している。

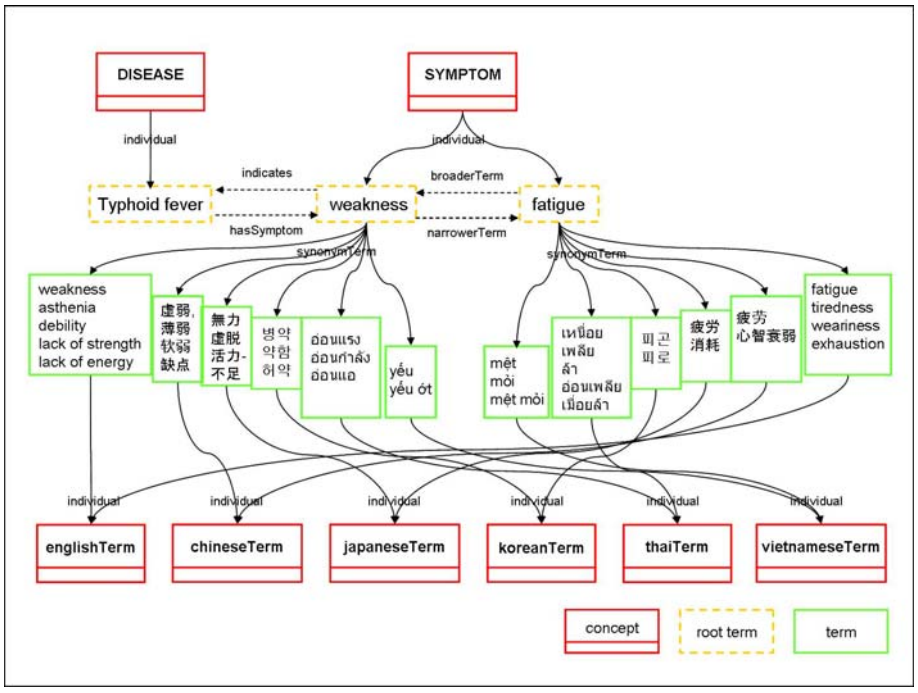

図５　DISEASE および SYMPTOM と腸チフス、衰弱および疲労の基本語に対する概念を示した BCO の概要基本語は、様々な言語の同義語とリンクしている（この図では６つ）。

４．次年度以降の研究展開

BCO は、感染症監視システムの開発を支援するために多言語オントロジーが必要であることから計画された。BCO は感染症監視システムを支援するために透過的かつ協力的に開発されたアプリケーションオントロジーである。同時に我々は、単一言語の生物医学的テキストマイニングの開発を自力で行うために、専門用語が必要なアジア太平洋言語に対してそれが使用できると期待している。BCO の第２版は http://biocaster.nii.ac.jp にて 2008 年にリリースされた。我々は、オントロジーの改善と拡張に対するフィードバックを求めている。

今後の開発において我々は、新たな健康への脅威を網羅するため、オントロジーにおいて用語数および言語数を拡張する予定である。また、我々は、様々な言語の用語を収集するプロセスの速度を上げる際に自動知識獲得アルゴリズムがどのような役割を果たすのか、に関してさらに詳細に検討する予定である。

５．研究経費

平成１８年度実績　：7,000 千円
平成１９年度実績　：7,066 千円
平成２０年度実績　：7,500 千円

６．平成２０年度の研究成果

（１）主要成果物 main result（知見 knowledge、成果物 result、知的財産権 intellectual property right, etc.等）

Nigel Collier, Ai Kawazoe, Reiko Matsuda-Goodwin, Son Doan, Quoc Hung-Ngo, Lihua Jin, Mika Shigematsu, Dinh Dien, Roberto Barrero, Koichi Takeuchi, Asanee Kawtrakul, “The BioCaster Ontology database”, available at http://biocaster.nii.ac.jp.

（２）成果発表 presentation of the result

＜論文発表 presenting thesis＞

〔学術論文 academic thesis〕

1. Kawazoe, A., Jin, L., Shigematsu, M., Bekki, D., Barrero, R., Taniguchi, K. and Collier, N. (2009), “The development of a schema for semantic annotation: Gain brought by a formal ontological method”, J. Applied Ontology, IOS Press (in press).
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. Kawazoe, A., Chanlekha, H., Shigematsu, M. and Collier, N. (2008), “Structuring an event ontology for disease outbreak detection”, in BMC Bioinformatics, 9 (Suppl 3): S8, DOI: 10.1186/1471-2105-9-S3-S8.
4. 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.

5. 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 Multingual Ontology-Supported Text Classification System”, Advances in Disease Surveillance, Vol. 4, pp. 242.

〔会議録 conference record〕

1. Collier, N., Kawazoe, A. Shigematsu, M., Taniguchi, K., Jin, L., McCrae, J., Dien, D., Hung, Q., Takeuchi, K., Kawtrakul, A. (2007), "Ontology-driven Influenza Surveillance from Web Rumours", in proceedings of the 2007 Options for the Control of Influenza VI (Options), Toronto, Ontario, Canada, June.
2. Kawazoe, A., Jin, L., Shigematsu, M., Barerro, R., Taniguchi , K. and Collier, N. (2006), "The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system", Olivier Bodenreider (ed)., Proceedings of the International Workshop on Biomedical Ontology in Action (KR-MED 2006), Baltimore, Maryland, USA, November 8, pp. 77-85.

〔解説・総説 explanation of the outline, abstract〕

〔研究ノート research note〕

〔その他 others〕

＜会議発表等 conference presentation, etc.＞

〔招待講演 invitation lecture〕

〔一般講演 general lecture〕

＜著書等 one’s writings

（３）その他の成果発表 other result presentation