科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　３　日現在

機関番号：　１２１０２
研究種目：　基盤研究(B)
研究期間：　2011～2013
課題番号：　２３３０００３３

研究課題名（和文）トピックの特性の多観点把握に基づく多言語ウェブテキストの言語間対照分析システム

研究課題名（英文）A System for Multi-Faceted Topic Analysis and Cross-Lingual Comparative Web Text Analysis

研究代表者
宇津呂　武仁（UTSURO, TAKEHITO）

筑波大学・システム情報系・教授

研究者番号：９０２６３４３３

交付決定額（研究期間全体）：（直接経費）　　15,600,000円、（間接経費）　　4,680,000円

研究成果の概要（和文）：本研究では，ウェブ上で収集可能な多言語ニュース・ブログ・電子掲示板等の文書を情報源として，多言語での報道内容，関心動向や，意見の分布を分析し，国・文化・言語の間でどのような違いがあるのかを発見する過程を支援する方式について研究を行った．本研究では，以下の(i)～(iv)の多様な観点における差異に着目し，各観点における差異を発見する過程を支援する方式を実現した．(i) 一つのトピックの中での詳細な話題・関心事項の差異．(ii) 国・文化・言語の間の時系列特性の差異．(iii) 書き手の実体験に関する差異．(iv) 賛否・主観の差異．

研究成果の概要（英文）：In this project, given a collection of multilingual documents available on the Web, we cross-lingually and cross-culturally compare facts and opinions that are observed in the collected documents. We especially focus on various facets including the following four of (i) to (iv), and realize a framework of assisting the process of discovering cross-cultural differences: (i) differences in specific issues and concerns within a topic, (ii) differences in time series features, (iii) differences in real life experiences, (iv) differences in opinions.

研究分野：　情報工学

科研費の分科・細目：　情報学・メディア情報学・データベース

キーワード：　ディレクトリ・情報検索　ファセット検索　多言語処理　トピック分析　ニュース・ブログ

１．研究開始当初の背景

(1) 宇津呂，吉岡は，本研究課題の前身として，基盤(B)「トピックの特性を言語間で比較・対照分析する多言語ウェブテキストマイニングの研究」(代表:宇津呂,H20～22)，基盤(B)「異なる特徴を持つニュースサイトを比較対照する世界ニュース分析システムの研究」(代表:吉岡，H 21～23)において，ウェブ上の外国語文書における情報爆発という二重の壁を克服するため，

① システム側は，外国語文書に対して情報検索・集約技術を適用し，利用者が自力で言語間の差異に相当する個所を特定するために必要な手がかりを提示することに徹する．

② 利用者は，言語間の差異に相当する個所を特定した後，人手による翻訳結果を参照して正確な情報を得る．

というアプローチをとった．そして，各言語のブログマイニングにおいて既存の検索エンジンを上回る性能を達成するとともに，多言語ニュース・ブログにおいて，言語間の差異に相当する個所を利用者が特定する過程を支援するプロトタイプ方式を開発した．

(2) その一方で，(1)の研究課題において発見された差異を詳細に分析した結果，その大部分は，トピックについての詳細な話題・関心事項，時系列特性，実体験に関する情報の有無，意見の分布，といった複数の因子における差異およびその組み合わせに起因することが判明した．さらに，国・文化・言語の間の差異に相当する個所の提示にとどまらず，これらの多様な観点から構成される複合的因子を特定・類型化し，分かりやすく利用者に提示することによって，利用者による差異発見の効率を大幅に改善できると確信し，本研究課題を提案するに至った．

２．研究の目的

本研究では，ウェブ上で収集可能な多言語ニュース・ブログ・電子掲示板等の文書を情報源として，多言語での報道内容，関心動向や，意見の分布を分析し，国・文化・言語の間でどのような違いがあるのかを発見する過程を支援する．そのために，以下の(i)～(iv)の多様な観点における差異を自動で特定・類型化する技術を実現する．

(1) 一つのトピックの中での詳細な話題・関心事項の差異．
(2) 国・文化・言語の間で関心が集中した時期が異なる，といった時系列特性の差異．
(3) ブログ・掲示板・レビューサイト等における書き手の実体験に関する記述の有無の差異．
(4) 一つのトピック，あるいは，その中での詳細な関心事項に対する賛否・主観の差異．

宇津呂，吉岡のこれまでの研究では，このような多様な観点を考慮できておらず，単に，言語間の差異の候補として検出された個所を利用者に提示するだけであった．一方，本研究では，(1)～(4)の各観点における差異を利用者に提示し，利用者が言語間差異を発見する過程を支援する言語間対照分析システムを作成する．そして，提案方式によって，利用者による差異発見を促進する．

３．研究の方法

(1) Wikipedia には，多言語トピックモデルの情報源としては，最大規模のトピック数を収録しており，新規の流行的トピックへの対応も極めて早い．そこで，多言語 Wikipedia の一つのエントリの記述をトピックモデルとみなして，入力文書の記述内容・詳細な話題を特定する手法の研究を行う．これまでに，小規模なプロトタイプ方式を提案済みであるので，これを大規模化して実装しその有効性を評価する．

(2) 国・文化・言語の間の差異を正確にとらえるためには，ニュース記事・ブログ記事等の時間情報を把握し，特定の言語・時期にのみ観測される特異な話題・関心を特定することが重要である．そこで，時系列解析において著名な Kleinberg のバースト解析モデルと(1)の多言語 Wikipedia トピックモデルを併用して，話題のまとまり単位でのバースト解析を実現し，各言語特有の時系列特性をとらえる方式を実現する．

(3) ブログ・掲示板・レビューサイト等を対象として，書き手の実体験に関する記述の有無および意見の分布特性を正確に特定する手法を確立する．

(4) (1)～(3)の各観点において把握した単言語でのトピックの特性に対して，言語間の差異度を定式化する．翻訳資源として Wikipedia の項目間対訳関係リンク，既存の対訳辞書を用いる．次に，各観点についての差異を手がかりとして，利用者が言語間の差異を発見する過程を支援する言語間対照分析システムを構築する．

４．研究成果

(1) 広義には同一のトピックについてのニュース記事・ブログ記事であっても，各言語での記述内容における詳細な話題・関心事項を正確に特定し，その微妙な差異を言語間で検出することが重要であるという知見が得られた．例えば，「臓器移植」の例では，日本語特有の現象として，ニュー

ス・ブログにおいて，特定の話題「臓器移植法」への関心が高く，英語ブログ特有の現象として，「euthanasia(安楽死)」への関心が観測される．そのため，これらの詳細な話題・関心の差異の検出が重要な手がかりとなる．そこで，多言語 Wikipedia の一つのエントリの記述をトピックモデルとみなして，入力文書の記述内容・詳細な話題を特定する手法の研究を行った．この研究に対して，電子情報通信学会 言語理解とコミュニケーション研究会より学生研究賞を授与された．

(2) トピックの俯瞰的分布を表現するための数理モデルを実現するために，Wikipedia を知識源として,「分野トピックモデル」というトピックモデルの一種を開発した．これによって，分野の粒度での分布を表現し，通常のトピックモデルの上位概念として位置する数理モデルとして利用することが可能となった．この研究に対して，情報処理学会データベースシステム研究会より学生奨励賞を授与された．

(3) 時系列解析において著名な Kleinberg のバースト解析モデルと (1) の多言語 Wikipedia トピックモデルを併用して，話題のまとまり単位でのバースト解析を実現し，各言語特有の時系列特性をとらえる方式を実現した．従来のバースト解析では，個々のキーワードの時系列特性を独立に解析するために，話題のまとまりとしてのバーストを自動認識することが困難であった．一方，提案方式では，同一話題の文書集合を同定したうえで，話題単位でバーストを検出することを実現した．

さらに，バースト時期の類似性，トピックモデルの対訳関係を考慮して，言語間の対応・差異を定式化した．具体的な分析対象として，日本語ニュースおよび中国語ニュースを取り上げ，時系列トピックモデルの推定，および，時系列トピックモデル間の日中対応を同定する方式を実現することにより，言語間の差異度を測定した．

(4) トラブルに巻き込まれた実体験者等がその実体験を語る中で発信する意見・主観情報の事例を分析し，特徴的な言語表現を類型化した．具体的には，すでに起こった被害事象に対してその被害状況を説明する表現，および，被害に遭って生じた感情を説明する表現において特に重要な特徴があり，実体験と意見との相関の分析に有効であることを示した．

また，質問・回答サイトを対象として，多様なトピックに渡って，実体験の中でも特に重要性の高いものであるトラブルの有無の分析を対象として，トピックの特性を把握する技術について研究を行った．まず，多様なトピックに渡って，トラブルの実体験に関する記述例を多数収集するために，種となる情報源として，国民生活センター(消費生活センター)におけるトラブル相談およびトラブル解決策指導事例のテキストを用いた．そして，一般の相談・回答サイト中のテキスト集合を対象として，相談・回答サイト中のトラブル実体験相談事例の候補を収集した．次に，意見表現とトラブル実体験の有無との間の相関の分析を行い，トピックを横断して両者の相関を観測することができた．

(5) 実体験の中でも特に重要性の高いトラブルの有無，および，トラブル周辺での関心事項を対象として，言語間の対応・差異を分析した．具体的な分析対象として，日本語および中国語の質問・回答サイトを取り上げ，特定の話題に関連して，実体験の有無に関して，言語間でどのような差異が認められるかを分析する方式を実現した．

(6) 特定の話題に関するコミュニティにおいて強い関心を持つブロガーを対象として，ブロガーの持つ関心事項・意見を日本語と中国語の間で比較・対照分析し，差異の発見を支援する方式を実現した．ここでは，日中両言語において，大規模にブロガーのブログ記事データを収集するとともに，両言語でトピックモデルを適用することにより，ブロガー・コミュニティを作成し，日中間でコミュニティ比較を行う方式を実現した．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① Kensaku Makita, Hiroko Suzuki, Daichi Koike, Takehito Utsuro, Yasuhide Kawada, Tomohiro Fukuhara, Labeling Blog Posts with Wikipedia Entries through LDA-Based Topic Modeling of Wikipedia, Journal of Internet Technology, 査読有, 14巻, 2013, 297-306, 10.6138/JIT.2013.14.2.13

② Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, Yoji Kiyota, Applying a Burst Model to Detect Bursty Topics in a Topic Model, Lecture Notes in Computer Science, 査読有, 7614巻, 2012, 239-249, 10.1007/978-3-642-33983-7_24.

③ Daisuke Yokomoto, Kensaku Makita, Hiroko Suzuki, Daichi Koike, Takehito Utsuro, Yasuhide Kawada, Tomohiro Fukuhara, LDA-Based Topic Modeling in Labeling Blog Posts with Wikipedia Entries, Lecture Notes in Computer Science, 査読有, 7234

巻 , 2012, 114-124, 10.1007/978-3-642-29426-6_15.

④ Daisuke Yokomoto, Kensaku Makita, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara, Utilizing Wikipedia in Categorizing Topic related Blogs into Facets, Procedia - Social and Behavioral Sciences, 査読有, 27 巻, 2011, 169-177, 10.1016/j.sbspro.2011.10.595

⑤ Taichi Katayama, Akihito Morijiri, Soichi Ishii, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara, Comparing Similarity of HTML Structures and Affiliate IDs in Splog Analysis, Lecture Notes in Computer Science, 査読有, 6637巻, 2011, 378-389, 10.1007/978-3-642-20244-5_36.

〔学会発表〕(計 33 件)

① Liyi Zheng, Tian Nie, Ichiro Moriya, Yusuke Inoue, Takakazu Imada, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. Comparative topic analysis of Japanese and Chinese bloggers., 7th International Symposium on Mining and Web, 2014 年 5 月 15 日, ビクトリア(カナダ)

② Shota Arai, Tian Nie, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. Collecting and classifying examples of consumer troubles on contract and cancellation" in a question-answer site, the 10th International Symposium on Natural Language Processing, 2013 年 10 月 28 日. プーケット(タイ).

③ Daichi Koike, Yusuke Takahashi, Takehito Utsuro, Masaharu Yoshioka, and Noriko Kando, Time series topic modeling and bursty topic detection of correlated news and twitter., 6th International Joint Conference on Natural Language Processing, 2013 年 10 月 16 日, 名古屋国際会議場(愛知県)

④ Liyi Zheng, Takehito Utsuro, and Masaharu Yoshioka, Bursty topics in time series Japanese/Chinese news streams and their cross-lingual alignment, 13th Conference of the Pacific Association for Computational Linguistics, 2013 年 9 月 2 日, 慶應義塾大学三田キャンパス(東京都).

⑤ Shuo Hu, Yusuke Takahashi, Liyi Zheng, Takehito Utsuro, Masaharu Yoshioka, Noriko Kando, Tomohiro Fukuhara, Hiroshi Nakagawa, and Yoji Kiyota, Cross-lingual topic alignment in time series Japanese/Chinese news, 26th Pacific Asia Conference on Language, Information and Computation, 2012 年 11 月 9 日, バリ(インドネシア).

⑥ Yusuke Takahashi, Shuo Hu, Takehito Utsuro, and Masaharu Yoshioka. Applying a burst model to detect bursty topics, 12th China-Japan Natural Language Processing Joint Research Promotion Conference, 2012 年 7 月 17 日, ハルビン(中国).

⑦ Takehito Utsuro, Shuo Hu, and Yusuke Takahashi. Bursty topic detection and cross-lingual topic alignment, 12th China-Japan Natural Language Processing Joint Research Promotion Conference, 2012 年 7 月 17 日, ハルビン(中国).

〔その他〕

・ツイッター・ブログ・ニュースの話題を集約・俯瞰する検索エンジン
http://nlp.iit.tsukuba.ac.jp/research/list03.html

・東日本大震災に関するニュース・ブログの分析支援
http://nlp.iit.tsukuba.ac.jp/research/list04.html

・安心・安全な社会を支える情報アクセス
http://nlp.iit.tsukuba.ac.jp/research/list05.html

6．研究組織

(1)研究代表者

宇津呂 武仁（UTSURO TAKEHITO）
筑波大学・システム情報系・教授
研究者番号： 90263433

(2)研究分担者

吉岡 真治（YOSHIOKA MASAHARU）
北海道大学・情報科学研究科・准教授
研究者番号： 40290879

乾 孝司（INUI TAKASHI）
筑波大学・システム情報系・助教
研究者番号： 60397031

(3)連携研究者

中川 裕志（NAKAGAWA HIROSHI）
東京大学・情報基盤センター・教授
研究者番号： 20134893

清田 陽司（KIYOTA YOJI）
東京大学・情報基盤センター・特任講師
研究者番号： 10401316