2F4-3

複数の言語で記述されたブログ記事を対象とした言語横断型関心解析システム

A Cross-Lingual Concern Analysis System using Multilingual Weblog Articles

福原知宏[*1]
Tomohiro Fukuhara

宇津呂武仁[*2]
Takehito Utsuro

[*1] 東京大学人工物工学研究センター
RACE (Research into Artifacts, Center for Engineering), The University of Tokyo

[*2] 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba

中川裕志[*3]
Hiroshi Nakagawa

武田英明[*1*4]
Hideaki Takeda

[*3] 東京大学情報基盤センター
Information Technology Center, The University of Tokyo

[*4] 国立情報学研究所実証研究センター
Principles of Informatics Research Division, National Institute of Informatics

A system for analyzing concerns of people from multilingual Weblog articles is proposed. Finding concerns of people around the world is important for various domains such as business, politics, science, and education. We propose a cross-lingual concerns analysis system that collects and analyzes Weblog articles written in multiple languages. The system collects and analyzes Japanese, Chinese, Korean, and English blog articles. Users can find differences of concerns among languages. Analysis results of differences of concerns on a topic are described.

1. はじめに

本研究では複数の言語で記述されたブログ記事を対象とする言語横断型関心解析システムについて述べる. 現在, ブログ空間(blogosphere)は多言語化の状況に向かっている. ブログを対象とした検索サービスを提供している Technorati (http://www.technorati.com/)によれば, 2007 年 4 月の段階でブログ空間に占める言語は上位から日本語, 英語, 中国語, イタリア語, スペイン語となっている[Sifry2007].

ブログ空間はもとより World Wide Web (WWW)における多言語化は今後益々進展することが予想される. こうした多言語下での情報流通では, ある言語圏で発生した情報は速やかに翻訳されて別の言語圏に伝えられるが, 別の情報は伝達に時間が掛かったり, あるいは全く伝えられないということも考えられる. 大向らの提唱する Community Web プラットフォーム[大向 2006]では WWW 上での人々のつながりを積極的に利用することで知識や情報の効果的な流通を目指すが, これに言語横断的な観点を加えることで, 世界規模の知識と情報の共有が可能となる. 本研究では言語横断的な Community Web の実現に向けて, 複数の言語で記述されたブログ記事を対象とした言語横断型関心解析システムを提案する.

本論文の構成は次の通りである. 2.では言語横断型関心解析システムの概要とプロトタイプシステムについて述べる. 3.では共起語分析による日本語, 韓国語, 英語ブログにおける関心の違いについて述べる. 4.では関連研究と今後の課題について述べる. 5.で本論文の議論をまとめる.

連絡先:福原知宏, 東京大学人工物工学研究センター価値創成イニシアティブ(住友商事)寄付研究部門, 千葉県柏市柏の葉 5-1-5, E-mail: fukuhara at race.u-tokyo.ac.jp

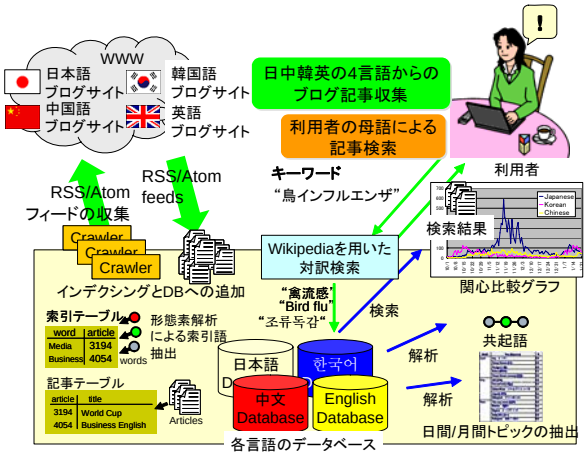

Fig. 1 言語横断型関心解析プロトタイプシステムの概要

2. 言語横断型関心解析システム

言語横断型関心解析システムについて述べる. (1)システム概要, (2)データ収集状況, (3)Wikipedia を用いた対訳表現検索について述べる.

2.1 システム概要

Fig. 1にプロトタイプシステムの全体像を示す. システムは日本語, 中国語, 韓国語, 英語のブログサイトをクローリングして記事を収集し, 各言語のデータベースに格納する. 利用者は母語でキーワード検索を行い, 各言語の記事における関心比較グラフを出力する. システムは各言語における対訳表現を求め, 求めた表現を用いて記事検索を行う.

またシステムは利用者からの要求に応じて各言語における共起語や日間／月間の頻出語(トピック)を出力する.

	日本語	中国語	韓国語*	英語
記事数	148,910,915	7,283,753	25,707,257	8,066,624
サイト数	2,765,470	669,415	461,762	75,832
収集期間	1,123 日	866 日	541 日	156 日
1 日の収集記事数	50 万記事	2 万記事	80 万記事	7 万記事

Table 1　データ収集状況(2007 年 4 月 15 日現在)
(*韓国語については 2007 年 1 月 24 日時点のデータ)

Fig. 2　Wikipedia を用いた対訳検索

2.2 データ収集状況

データ収集状況をTable 1に示す. 日本語記事数は記事収集を開始してから通算 15 億記事, 次いで韓国語, 英語, 中国語の順になる. 記事を収集する元となるサイト数では, 日本語サイトが 275 万サイト, 次いで中国語, 韓国語, 英語の順となる.

2.3 Wikipedia を用いた対訳表現検索

対訳表現検索には Wikipedia を用いた. 対訳表現検索では翻訳元言語で入力されたキーワードを目標言語のキーワードに変換することを目的とする.

Fig. 2に Wikipedia を用いた対訳表現検索の概要を示す. システムは翻訳元言語のキーワードを入力として受け付け, Wikipedia 内に該当するエントリを探す. もしエントリが見つかれば, 他の言語で記述された同一エントリへのリンクを得て目標言語における表記を得る. 翻訳元言語から目標言語への直接的なリンクが無い場合, 他の言語(現在は英語)のエントリを経由することで目標言語の表記を得る. なお, 翻訳元言語にエントリが無い場合, 翻訳元言語のエントリから目標言語の同一エントリにリンクが無い場合は対訳表現を得られなかったものとして処理を中止する.

3. 共起語を用いた言語間関心比較

ここでは日本語, 韓国語, 英語ブログ間での共起語を用いた関心比較について述べる.

3.1 概要

ここでは日本語, 韓国語, 英語語のブログ記事を対象として, クローン技術に関する共起語を用いた言語間の関心比較を行った. 調査期間は 2006 年 12 月 1 日から 2007 年 1 月 27 日までである. 期間中の記事数はそれぞれ 1,593 件(日), 272 件(韓), 2,847 件(英)であった. 分析に用いたキーワードは "クローン"(日), "복제"(韓), "cloning"(英)である. 以下, 各言語の共起語について述べる.

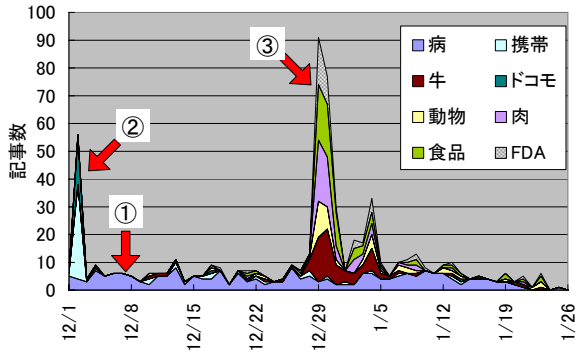

Fig. 3 "クローン" の主要共起語 (日本語ブログ)

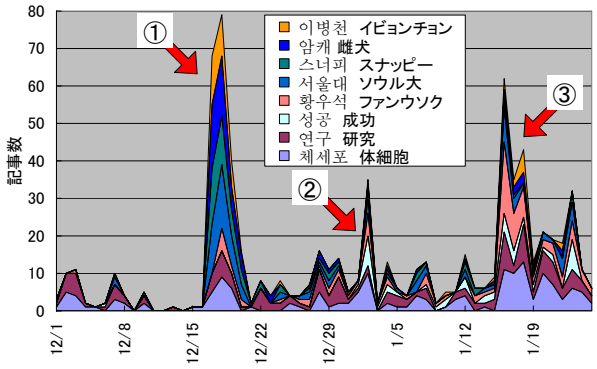

Fig. 4 "복제 (複製)" の主要共起語 (韓国語ブログ)

3.2 日本語ブログにおける関心

日本語ブログではこの時期, 大きく分けて 3 つの対象について関心が持たれていた:(1)クローン病(Crohn's disease)に対する関心, (2)クローン携帯に対する関心, (3)クローン動物の加工食品への利用に対する関心である. Fig. 3に日本語ブログにおける主要共起語の時間推移を示す. X 軸は日付, Y 軸は共起語を含む記事数である. 図は積み上げグラフである.

図中, "病"という共起語が継続して出現している(図中①)が, これは(1)のクローン病に対する関心である. (2)は 12 月 1 日付近にピーク(図中①)として現れている"携帯"と"ドコモ"という共起語である. この時期, クローン携帯が存在するという噂が流れ, この噂を受けての関心が現れた. (3)は 12 月 29 日付近に現れるピーク(図中③)で, "牛", "動物", "肉", "食品", "FDA"からなる関心である. これは米国食品医薬局(FDA)がクローン技術を使った家畜の加工食品に対して安全であると発表したことに起因する.

このように日本語ではクローン病についての関心が定常的に存在し, (2)や(3)のようにその時々の話題がクローンの共起語として結びついていることが分かった.

3.3 韓国語ブログにおける関心

韓国語ブログでは大きく分けて 3 つの話題が見られた:(1) イビョンチョン(李柄千, 이병천)教授のソウル大学復職に関する関心, (2) 狂牛病耐性牛クローン誕生に関する関心, (3) ファンウソク(黄禹錫, 황우석)教授の研究再開に関する関心である. Fig. 4に韓国語ブログにおける主要共起語の推移を示す. 韓国語ブログではクローンが ES 細胞事件関係者の話題と共起して語られていたことが分かる.

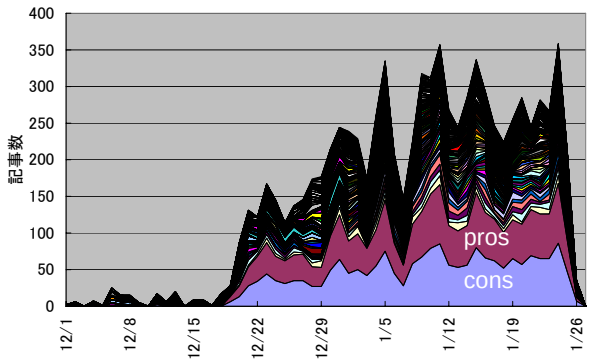

Fig. 5 “Cloning”の主要共起語（英語ブログ）

(1)は ES 細胞事件でファン教授の下で研究していた研究者イ・ビョンチョン教授がソウル大学に復職したと伝えられたことに対する関心である(図中①). (2)は狂牛病に耐性を持つクローン牛が誕生したことに対する関心である(図中②). (3)はファン・ウソク教授が研究活動を再開したとの報道に対する関心である(図中③).

3.4 英語ブログにおける関心

英語ブログでは“pros”と“cons”という語が出現している. Fig. 5に共起語の時間軸上の推移を示す. 12 月 15 日を過ぎた辺りから“pros”と“cons”を含む記事が大量に出現している. これらの記事は定型文を利用したスパム記事(Splog)であり, 人々の関心はスパムに埋もれてしまっている. 今後, Splog を除外する手立てが必要である.

3.5 言語間関心比較:まとめ

以上, 各言語における共起語について述べ, 言語間で関心の対象に違いが見られることが分かった. 日本語ブログでは(1)クローン病, (2)クローン携帯, (3)クローン動物の加工食品への利用に関心が集まっていた. 韓国語ブログでは ES 細胞事件関係者の話題に関心が集まっていた. 英語ブログではスパム記事の影響が強く, これを除外しない限り正しい関心を把握することが出来ないことがわかった.

4. 関連研究と今後の課題

関連研究と今後の課題について述べる.

4.1 関連研究

吉岡は国内外のニュースサイトの報道記事を用いたニュース分析手法を提案している[吉岡 2007]. この手法では国内外で日本語記事を配信しているニュースサイトから記事を収集し, ニュースサイト間に共通する語や, 各サイトに特有の語を抽出する. 本研究では複数の言語で記述されたブログ記事を対象としているが, 吉岡の手法を参考に言語間で共通, あるいは異なる話題を自動抽出する課題についても取り組む必要がある.

米国では DARPA 主催による GALE(Global Autonomous Language Exploitation)プロジェクト[1]がある. GALE プロジェクトでは各国語で記述されたテキストを対象とし, 機械翻訳や情報抽出, 自動要約等の自然言語処理を用いて情報分析者を支援する. 本研究でも各国語のテキストを対象とした分析支援を行うが, 対象がブログであり時々刻々と発信される膨大なデータから分析するに値するデータを取捨選択し, データの質を考慮した分析を今後行う必要がある.

多言語環境下におけるコミュニケーション支援の観点からは, 京都大学の石田らによる言語グリッドプロジェクト[2]がある[石田 2006]. 言語グリッドでは複数の異なる母語使用者間のコミュニケーションを支援するための情報資源とツール開発に重点を置いている. 本研究は現時点ではコミュニケーション支援を想定していないが, 例えば 3 節で示したような言語間での関心の違いを様々な母語利用者に提示することで, 互いの関心を知り, 円滑なコミュニケーションに役立てられるのではないかと考える. 今後, こうしたコミュニケーション支援の側面について検討する.

4.2 今後の課題

本研究の今後の課題は次の通りである.

1. Wikipedia を用いた対訳表現検索の評価
   Wikipedia を用いた対訳表現検索の定量的評価を行う必要がある.
2. Wikipedia 以外の対訳表現の獲得
   対訳表現を自動的に抽出する手法について研究を行う必要がある.
3. 機械翻訳機能の導入
   現在のシステムは検索結果をそれぞれの言語で返戻しているが, 利用者が各自の母語で内容を確認するための機械翻訳を導入する必要がある.
4. Splog 対策
   3.4 節の英語ブログの共起語に見られるように Splog 対策が重要な課題である. Kolari らは英語のブログサイトを対象としたスパムフィルタリング手法を提案している[Kolari2006]. 本研究においても Splog の現状把握と対策を行う必要がある.

5. まとめ

本研究では複数の言語で記述されたブログ記事を対象とした言語横断型関心解析システムについて述べ, 日本語, 韓国語, 英語のブログ記事間での共起語による関心比較を行った. 今後, 言語横断型 Community Web の実現に向けてシステム開発を行う予定である.

参考文献

[Sifry2007] David Sifry: The State of the Live Web, April 2007, 2007. (Available at http://www.sifry.com/alerts/archives/000493.html, accessed 2007-4-15)

[大向 2006] 大向一輝, 松尾豊, 松村真宏, 武田英明: Community Web プラットフォーム, 人工知能学会論文誌, Vol.21, No.3, pp.251-256,2006.

[石田 2006] 石田亨, 内元清貴, 山下直美, 吉野孝: 機械翻訳を用いた異文化コラボレーション , 情報処理 Vol.47,No.3, pp.269-275, 2006

[吉岡 2007] 吉岡真治: 複数のニュース源の差異を考慮したニュース分析の研究, 第 13 回言語処理学会年次大会ワークショップ「大規模 Web 研究基盤上での自然言語処理・情報検索研究」論文集, pp.27-30, 2007.

[Kolari2006] Pranam Kolari, Akshay Java, and Tim Finin: Characterizing the Splogosphere, Proc. of the 3rd Annual Workshop on Weblogging Ecosystem, 15th World Wide Web Conference, 2006.

[1] http://www.darpa.mil/ipto/programs/gale/ (accessed 2007-4-15)

[2] http://langrid.nict.go.jp/ (accessed 2007-4-15)