多言語 Web 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の開発と将来構想

福原 知宏
東京大学人工物工学研究センター
http://www.race.u-tokyo.ac.jp/~fukuhara/

宇津呂 武仁
筑波大学大学院システム情報工学科
http://nlp.iit.tsukuba.ac.jp/member/utsuro/

中川 裕志
東京大学情報基盤センター
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/

武田 英明
国立情報学研究所実証研究センター
http://www-kasm.nii.ac.jp/~takeda/

Abstract

今日，Web の世界は多言語の状況にある．Weblog を対象とした検索サービスを提供している Technorati の報告によれば，今日のブログ空間は日本語，英語，中国語を始めイタリア語，スペイン語，ドイツ語など，さまざまな言語での情報交換が行われている．こうした Web の多言語化 (多言語 Web) は今後ますます進展することが予想されることから，言語の壁を越えた情報共有支援について取り組む必要がある．本論文では多言語 Web の構想について述べ，現在筆者らが研究開発を進めている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る．また人工知能研究が多言語 Web において今後取り組むべき課題について述べる．

1 はじめに

今日，Web 上には様々な言語で記述された文書が増えている．初期の Web では英語が主要な言語であったが，世界的なインターネット利用者の増加とともに現在では様々な言語で記述された文書がやり取りされるようになった．こうした Web の多言語化を本論文では**多言語 Web** と呼ぶ．多言語 Web の具体例として (1)Wikipedia と (2) ブログ空間 (blogosphere) について述べる．

Wikipedia は Web 上で誰もが閲覧・執筆可能な百科事典であり，2007 年 6 月 3 日の時点で 251 の言語で記事が記述されている．記事数の多い順に見ると英語 (1,763,740 記事)，ドイツ語 (577,920 記事)，フランス語 (483,875 記事)，ポーランド語 (373,684 記事)，日本語 (362,751 記事) と続いている．ここで注意すべき点は，同じ主題であっても記述されている内容は言語によって異なる点である．ある事柄を中立的に判断しようとする場合，できるだけ多くの観点 (言語) から記述された情報を確認する必要がある．この際，読者には母語以外の言語を読解する能力が求められるが，全ての人に多くの外国語能力を求めることは容易ではない．このため計算機による読解支援が必要となる．

第 2 の例はブログ空間である．今日のブログ空間は多言語化しており，Technorati のレポート[1]によれば，2006 年第 4 四半期のブログ空間は日本語 (37%)，英語 (36%)，中国語 (8%)，イタリア語 (3%)，スペイン語 (3%)，ロシア語 (2%)，フランス語 (2%) 等から構成されている．ブログ空間においても Wikipedia と同様に言語が変われば観点も変わる．

[1] http://www.sifry.com/alerts/archives/000493.html

このように現在の Web では，必ずしも全ての言語での情報発信は達成されていない [2] ものの，様々な言語による情報発信が行われている．Web 上では有用な情報は言語を超えて共有されるべきであり，このため言語の壁を越えて情報を共有するための技術とツールの開発が必要である．

多言語 Web を支える技術やツールには機械翻訳をはじめ，対訳抽出，複数言語要約，多言語文書クラスタリング，検索ツールやコミュニケーション支援ツールなどが考えられる．本論文では多言語 Web を支えるツールの一例として，複数言語で記述されたブログ記事を対象とした関心解析システムについて述べる．また，人工知能研究が多言語 Web の実現において果たす役割と，取り組むべき課題についても述べる．

本論文の構成は次の通りである． 2 節では多言語 Web と，関連する先行研究について述べる．3 節では多言語 Web を支えるツールの一例とし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について述べる． 4 節では人工知能研究が多言語 Web において今後取り組むべき課題について述べる． 5 節では本論文の議論をまとめる．

2 多言語 Web

本節では多言語 Web の概念と関連する先行研究について述べる．

2.1 多言語 Web とは

本論文では多言語 Web を次のように定義する: (1) 様々な言語による情報発信が可能な情報空間であり，かつ (2) 計算機支援によって他の言語で記述された情報内容に容易にアクセスすることのできる空間である．

現在の Web を見ると，(1) に関してあらゆる言語による情報発信は未だ実現されておらず，また (2) の要件も満たされていない．(1) の要件を満たすためには，現在計算機で扱うことの出来ない言語の調査や，未登録の文字の整備といった地道な作業が必要である．一方，(2) の要件を満たすためには，人工知能技術や自然言語処理技術の導入により効果が期待できる．本論文では (2) の要件に焦点を当てる．

2.2 関連する先行研究

多言語 Web に関連する先行研究について述べる．石田らは言語グリッドプロジェクトにおいて，互いに異なる言語圏の人々が円滑にコミュニケーションを行うための環境作りを進めている [4]．多言語 Web の実現において言語グリッドプロジェクトの取り組みは非常に重要である．言語グリッドプロジェクトでは実世界や計算機上でのコミュニケーションにおける支援を目指している．本研究では様々な言語で記述された文書の分析を支援する側面に焦点を当てる．

自然言語処理の分野では Evans らによる NewsBlaster がある [1]．NewsBlaster は複数の言語のニュース記事を対象とした文書分類・要約システムであり，英語に加え，日本語，ロシア語，フランス語，ドイツ語，スペイン語，イタリア語のニュース記事を収集し分類・要約する．NewsBlaster は各言語のニュース記事を収集し解析するシステムである．多言語 Web ではブログや掲示板など一般の人々の意見が現れやすい文書も対象とする必要がある．

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

本節では多言語 Web 支援ツールの一例として，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について述べる．

筆者ら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ツール: KANSHIN の研究開発を進めている [5]．このシステムは各言語で記述されたブログ記事を収集・解析し，各言語における関心動向を把握したり，言語横断的に関心動向を把握するためのツールである．

Figure 1 にシステムの全体像を示す．システムは日本語，英語，中国語，韓国語の 4 言語で記述されたブログ記事を収集し，各言語の形態素解析器を用いて単語を取り出し索引テーブルを作成し記事を格納する．利用者は各言語のブログ記事に対して (1) 記事検索，(2) 共起語検索，(3) 日間/月間話題検索を行える．これに加えてシステムは Wikipedia を用いて利用者から与えられた検索語をそれぞれの言語の検索語に翻訳し，各言語で記事を検索する．利用者は任意の検索語を用いて 4 言語のブログ記事を対象に検索を行うことができ，集計結果である関心比較グラフによって言語ごとの関心動向を確認できる．Figure 2 に “鳥インフルエンザ” という検索語に対する出力結果を示す．

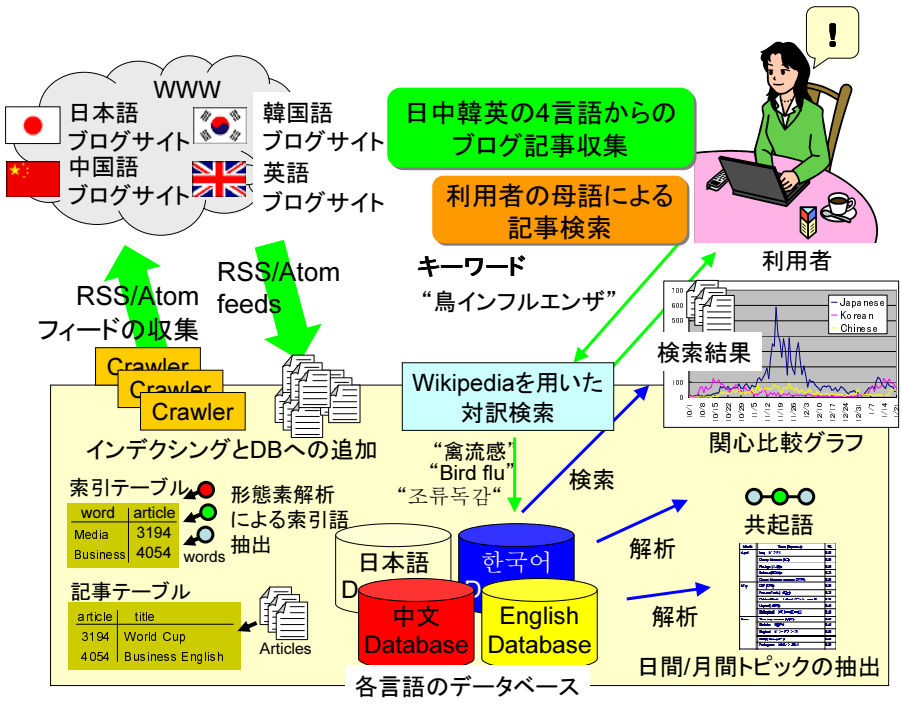

Figure 1: 言語横断型関心解析システムの概要

今後，検索結果の分類や要約，翻訳を行うとともに，上に述べた 4 言語以外にも言語を拡張する予定である．また現在のブログ空間において著者たちは自身の発信する記事が他の言語圏の読者に届くことは強く意識していないと推察する．これは著者たちに他の言語圏の読者からのフィードバックが戻りにくいためであり，例えばある話題についてそれぞれの言語圏から得た意見を一覧できるようになれば他の言語圏を意識した記事につながるであろう．今後の課題として，テーマごとに各言語圏の関心を一覧するためのシステムについても検討する．

4　人工知能研究が今後取り組むべき課題

本節では人工知能研究が多言語 Web において今後取り組むべき課題について述べる．

以下に多言語 Web の実現に向けて今後取り組むべき課題の例を挙げる．

1. 語の一般性を考慮した対訳抽出
2. 多言語 Web を考慮した情報発信ツール

第 1 に語の一般性を考慮した対訳抽出がある．例えば北朝鮮という語の対訳を求める場合，朝鮮民主主義人民共和国という正式名称の対訳より，North Korea や N.Korea や DPRK など，対象となる文書集合で多く使用されている対訳の方がより多くの検索結果が得られる可能性がある．このように対象とする言語や文書に応じて適切な対訳を抽出する手法の開発が必要である．

第 2 に多言語 Web を考慮した情報発信ツールの開発が必要である．今日，様々な情報発信ツールが存在するが，情報の発信段階で多言語 Web を意識してメタデータやアノテーションを自動的に付与するツールが必要である．長尾は情報内に意味情報を埋め込む semantic transcoding を提案している [3]．semantic transcoding により文書はもとより映像や音声といったマルチメディア情報の要約や分類といったことも可能となる．文書中に計算機処理可能なアノテーションを付与することで多言語 Web における情報アクセスを効率化できる可能性がある．

Figure 2: “鳥インフルエンザ” に対する出力結果

5 まとめ

本論文では多言語 Web について述べ，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た．また人工知能研究が多言語 Web において果たす役割と今後の課題について述べた．人工知能研究が全ての人にとってより良い多言語 Web をもたらすよう研究を進めて行きたい．

参考文献

[1] David Kirk Evans, Judith L. Klavans, and Kathleen R. McKeown. Columbia newsblaster: Multilingual news summarization on the web. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Demonstration Papers*, pages 1–4. Association for Computational Linguistics, 2004.

[2] Yoshiki Mikami, Pavol Zavarsky, Mohd Zaidi Abd Rozan, Izumi Suzuki, Masayuki Takahashi, Tomohide Maki, Irwan Nizan Ayob, Paolo Boldi, Massimo Santini, and Sebastiano Vigna. The language observatory project (lop). In *WWW '05: Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 990–991. ACM Press, 2005.

[3] 長尾 確. セマンティック・トランスコーディング: Semantic web のために今やるべきこと. In **情報処理学会第** *65* **回全国大会予稿集**, 2003.

[4] 石田 亨, 内元 清貴, 山下 直美, and 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**, 47(3):269–275, 2006.

[5] 福原 知宏, 宇津呂 武仁, 中川 裕志, and 武田 英明. 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. In **第** *21* **回人工知能学会全国大会予稿集**, 2007. CD-ROM(2F4-3).