URL:http://www.nii.ac.jp/ National Institute of Informatics

大規模ブログデータの研究基盤の構築

― 「Yahoo!ブログ」の研究利用による言語研究の新展開 ―

平成 20 年 4 月 23 日

国立情報学研究所
独立行政法人 国立国語研究所

（要点）

国立情報学研究所（所長：坂内正夫（さかうち まさお）以下、NII）と独立行政法人 国立国語研究所(所長：杉戸 清樹(すぎと せいじゅ)以下、国語研)は、Web上の利用者発信情報を対象とした情報検索、情報分析、情報活用などの研究および現代日本語のコーパス言語学的研究のため、ヤフー株式会社と「Yahoo!ブログ」データの研究利用について合意し、ブログデータを情報関連技術研究コミュニティに対して無償提供するとともに、日本語コーパスとして、一定の条件を満たせばどなたでも利用できるように整備します。

Web2.0の普及にともない、飛躍的な増大を続けるブログなどの利用者発信情報の研究利用は、現代日本語の研究や、情報爆発から新しい価値を汲み出すために必須のデータです。昨年４月の「Yahoo!知恵袋」データに引き続き、今回、「Yahoo!ブログ」データの研究利用が実現することは、情報爆発研究のさらなる加速を可能とするものです。

（具体的内容）

NIIでは、情報の未来価値として、情報爆発から「情報や情報分析結果から新しい価値を発見したり、汲み出したりする」ことの重要性に着目し、従来の書き言葉を主体としたテキスト情報に対して、「書き言葉と話し言葉」の中間的存在として、急速に増大しているブログやSNSなどの利用者発信テキスト情報の研究と、これに必要なデータ利用に関する準備を進めてきました。ここで新しい価値を汲み出す情報源の一つとして注目したのがブログやSNSなどのCGM - Consumer Generated Media（消費者生成メディア）です。

また、国語研では、特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21世紀の日本語研究の基盤整備」（略称「日本語コーパス」）と協力して、現代日本語の全体像をありのままに把握しようとする『現代日本語書き言葉均衡コーパス』の構築を進めてきました。

これらの研究の素材として、ヤフー株式会社からは既に「Yahoo!知恵袋」の投稿データが提供されています。「Yahoo!知恵袋」は、知識検索サービスという性質上、Q&Aという一定の型に従った投稿が多いのが特徴でした。CGMデータの研究利用の第一歩として画期的な試みでしたが、研究者からは、その表現の自由度の高さから、ブログの投稿データの収集が強く望まれてきました。

そして今回、ヤフー株式会社の協力により、これまで困難とされていた体系的なブログ

データの提供が実現するに到りました。

ブログデータは、ヤフー株式会社によりインターネット全体に公開された「Yahoo!ブログ」の記事を対象に収集され、NIIと国語研は固有名詞など個人の特定に結びつく可能性のある表現を除外するなどの作業を行った上で、研究利用に供されます。

（今後の課題）

しかし､ブログは「Yahoo!知恵袋」に比べ個人の意見や個人の特定に結びつく可能性のある表現が多く含まれており，その適切な取り扱いがより重要となるので､ヤフー株式会社、NII、国語研および東京大学情報理工学系研究科などの関連する研究コミュニティの関係者をメンバーとして、そのガイドライン作成を共同研究として行います。

ブログは、誰でも自由に情報発信を行うことができ、論説から独り言までを含む、世の中を反映する鏡のようなものです。ブログが社会に及ぼす影響も日々増大しています。「Yahoo!ブログ」データを使うことによって、言語学、国語教育、日本語教育、辞書編集、自然言語処理､さらにはブログの実態把握や意見分析、話題分析を通じたマーケティングや世論調査などへの利用の可能性も広がっています。このブログデータを最大限に利用することによって、新しい研究価値を生み出すと同時に、多様な利用者発信情報（ブログ、質問・回答文など）の研究利用の促進に向けて活動を続けます。

ヤフー株式会社からのコメント：

日々多くのお客様から投稿される日記やコメントは、時代を反映する“生きた日本語”そのものです。そのような書き言葉を研究素材にしたいという研究者の方々の強い要望に応えるのも Yahoo! JAPAN に期待されている役割の一つです。Yahoo! JAPAN は学術コミュニティへの協力と連携を積極的に行っています。インターネットの世界のみならず、さまざまな分野で研究成果が生かされることを期待しております。

ヤフー株式会社　ソーシャルネット事業部
事業部長　殿村　英嗣

東京大学 石塚満教授からのコメント：

Web2.0 の大きな特徴に「ユーザ参加型情報発信」あるいは CGM がありますが、ブログはその代表です｡新聞，TV 等のマスコミから発せられる情報とは異なり、多数の人々の日常生活を反映した生の声が直接的に現れます。特異な記述も混じっていますが、多数のブログからの統計処理やマイニングにより有意な情報を抽出することができます。これまで製品などの評判情報、流行語や関心事の把握などが行われていますが、今後も新しい利用が開拓されることになります。例えば、政治や経済問題についての意見分布の測定といっ

た人々の反応の詳細な把握などが考えられます。提供されるブログデータはこのような技術開発を促進するものであり、Web からの新価値を創造する機能の開発に活用したいと思います。日本は特にブログ量が多いので、世界をリードする日本発の独自技術を産み出したいと思います。

東京大学 大学院情報理工学系研究科
教授　石塚　満

（用語解説）

【特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」】

略称「日本語コーパス」は、平成 18 年度から 22 年度まで 5 年間のプロジェクト。

領域代表者： 前川喜久雄 （国立国語研究所研究開発部門）

本研究には次のふたつの目標があります。

（１）現代日本語のコーパス言語学的研究の基盤を整備するために、大規模な現代日本語書き言葉コーパスを構築すること

（２）構築途上のコーパスを様々な領域で利用することによってコーパス日本語学の可能性を探り、同時に構築中のコーパスを評価すること

【コーパス】

言語分析を行うための基礎資料として、書き言葉や話し言葉の資料を組織的に収集し、研究用の情報を付与したうえで電子的に保存したもの。偏りのない形で対象の全体像を反映したデータとなっていることが望まれる（均衡コーパス）。

【Yahoo! JAPAN】http://www.yahoo.co.jp/

ヤフー株式会社（市場名：東証 1 部/JASDAQ、銘柄コード：4689、本社：東京都港区、設立年月日：1996 年 1 月 31 日、代表取締役：井上雅博）が運営する Yahoo! JAPAN は、1 か月あたり約 5122 万人のユニークカスタマー数※と、1 日 16 億ページビューのアクセスを誇るインターネットの総合情報サイトで、検索、コンテンツ、コミュニティー、コマース、モバイルなど多くのサービスを提供しています。

※2008 年 2 月の Nielsen Online「NetView AMS JP」における家庭からの視聴率 88.2%、職場からの視聴率 92%というデータをもとに、家庭、または職場からのインターネットユーザーを約 5798 万人（Nielsen Online「インターネット基礎調査」より）として Yahoo! JAPAN のユニークカスタマー数を算出。

【Yahoo!知恵袋】http://chiebukuro.yahoo.co.jp/

「Yahoo!知恵袋」は、質問したい人と回答したい人とを結び、知恵や知識を参加者同士で

共有する、日本最大の知識検索サービス（Q&A サイト）です。2004 年 4 月のサービス開始以来、質問総数 1500 万件以上、回答総数は 4800 万件を数え、現在は約 180 万人が登録するナレッジコミュニティです)。

【Yahoo!ブログ】http://blogs.yahoo.co.jp/
「Yahoo!ブログ」はブログを簡単に作成、更新できるサービスで、Yahoo! JAPAN ID を取得していれば、誰でも無料で利用できます。2005 年 1 月にサービスを開始し、現在のブログ開設数は約 150 万件です（2008 年 3 月現在)。

■本件問い合わせ先

国立情報学研究所
東倉 洋一 教授・副所長
大山 敬三 コンテンツ科学研究系教授・研究主幹
国立国語研究所
前川喜久雄 研究開発部門 言語資源グループ グループ長

取材窓口／その他問合せ
国立情報学研究所 （NII：エヌアイアイ）
企画推進本部広報普及チーム 担当： 佐久間・小野
〒101-8430 東京都千代田区一ツ橋 2-1-2（学術総合ｾﾝﾀｰ 18 階）
TEL:03-4212-2135(直通) FAX:03-4212-2150
e-mail: kouhou@nii.ac.jp
URL：http://www.nii.ac.jp/