オンライン語彙獲得を用いたリアルタイムウェブの言語処理

村脇 有吾　　黒橋 禎夫
京都大学大学院情報学研究科
murawaki@nlp.kuee.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

1　はじめに

日本語の分かち書きしないという特徴は、テキスト処理を利用したアプリケーション構築の敷居を高くしている。簡単な例として、テキストからのキーワード抽出を考える。英語と異なり、分かち書きの単位を抽出するという単純な手法が使えない。そこで必要となるのが形態素解析である。近年は、高精度なオープンソースの解析器の登場により、形態素解析が共有化されている。自然言語処理を専門としないプログラマでも、形態素解析を要素技術として用いたアプリケーションを気軽に構築できるようになりつつある。

しかし現在の形態素解析には様々な問題がある。中でも深刻なのが未知語の問題である。形態素解析ではあらかじめ人手により整備された形態素辞書が重要な役割を果たし、辞書登録から漏れた形態素 (未知語) の解析を誤りやすい。例えば、「ついったー」といった新語を正しく解析できない。

未知語問題を解決するために、我々はオンライン語彙獲得という枠組みとその具体的な実装手法 [3, 4] を提案している。この枠組では、語彙獲得器がテキスト中の未知語を同定し、人手の介在なしに解析用の辞書に追加する。これにより、形態素解析器にとって未知語が既知となり、正しく解析できるようになる。

我々の手法の特徴の一つは、従来手法がテキストをバッチ的に処理すると違い、逐次的に入力されるテキストから未知語を獲得することである。我々は、このオンラインという性質は、リアルタイムウェブの言語解析に応用出来ると考えている。ウェブ上では刻々とテキストが産出されており、こうしたデータをできるだけ早く組織化したいという需要がある。その際に問題となるのが、やはり未知語である。そこで、本稿ではリアルタイムウェブの言語処理を試みる。具体的には、提案手法の共有化を将来の目標として、リアルタイムウェブの代表たるツイッターを対象に、オンライン語彙獲得を行うプロトタイプを実装したので、現状と課題を報告する。

2　リアルタイムウェブ

リアルタイムウェブとは、ウェブ上で日々生み出される情報をできるだけ早く組織化する技術の総称である。その背景には、Web 2.0 やソーシャルウェブなどと呼ばれるウェブの利用形態がある。従来、ウェブの利用者は情報を一方的に受け取っていたのに対し、利用者自身も情報を発信するようになったとされる。特に、ソーシャル・ネットワーキング・サービスなどを通じて、利用者が相互の交流の中でコンテンツを生み出すという特徴がある。こうして生み出されるコンテンツは、潜在的に商業的な利用価値を持っている。例えば、ブログ上の製品の評判を分析すれば、マーケティングに役立つかもしれない。

こうしたコンテンツのもう一つの特徴は、データが刻々と生み出されることである。そうすると、最新の情報をいち早く利用可能にしたいという要求がうまれる。このような特徴を持つウェブそのものや、この要求を実現するための技術を総称してリアルタイムウェブとよぶ。特に注目されているのが、リアルタイム検索[1]である。

本稿は、リアルタイムウェブの代表例であるツイッター[2]に着目する。ツイッターはマイクロブログのサービスに分類される。利用者は「つぶやき」とよばれる投稿を行うが、その長さは一度に最大 140 文字に制限されている。もう一つの主要機能は、別の利用者のフォローである。フォローによって、別の利用者のつぶやきが、図 2 のようにタイムラインとよばれる時系列順のつぶやき一覧に表示されるようになる。タイムラインはチャットのように見えるが相違点もある。一つは、必ずしも双方向的でないことである。自分が相手をフォローしていても、相手が自分をフォローしていなければ、相手のタイムラインに自分のつぶやきは表示されない。もう一つは、個々の独立したコミュニティーが存在するのではなく、サービス全体が緩やか

[1] http://googleblog.blogspot.com/2009/12/relevance-meets-real-time-web.html
[2] http://twitter.com/

につながった一つの系をなしていることである。これらの特徴があわさって利用者が気軽に投稿できる環境ができている。

ツイッターの別の特徴は、サービスを利用するためのAPIが整備されていることである。このAPIを利用して、様々な専用クライアントが開発されたり、豊富なメタデータを使った利用実態の調査 [1, 2] が行われたりしている。また、後述のようにツイッターのデータを利用した別のウェブサービスが開発されている。

3　リアルタイムウェブの言語処理

リアルタイムウェブを利用して新たなウェブサービスを開発する際、注目するデータには2種類ある。一つはメタデータであり、もう一つはテキストそのものである。本稿では後者に着目する。

テキストに注目したサービスとしては、ツイッターを対象に、その瞬間に話題となっているキーワードを抽出するbuzztter[3]がある。また、ブログを対象とした同様のサービスにkizasi.jp[4]がある。

こうしたサービスを実現するには、要素技術としてキーワード抽出が必要となる。キーワードを抽出を行うために、分かち書きされない日本語では、まず形態素解析を行うことが多い。しかし、形態素解析には、辞書にない形態素 (未知語) の解析を誤りやすいという問題がある。形態素解析の辞書は主に新聞記事の解析を想定して作られている。そのため、対象テキストに新聞であまり使われない形態素が頻出する場合、解析誤りが目立つ。例えば、形態素解析器JUMAN[5]は「ついったー」を「つ + いった + ー」と誤解析する。

未知語問題への解決策は、辞書を拡張して未知語を既知に変えることである。未知語の知識源として、ウィキペディア日本語版[6]の見出しや、はてなキーワード[7]などが考えられる。しかし、こうしたデータには問題が少なくない。例えば、「2010年」、「ϕ」、「一覧の一覧」など、キーワードとしてふさわしくない語が含まれている。また、こうしたキーワードの大半が名詞(句)であり、動詞や形容詞の未知語は通常登録されていない。そもそも、ウィキペディアのスナップショット[8]の更新は頻繁ではなく、リアルタイム性に問題がある。

[3] http://buzztter.com/ja
[4] http://kizasi.jp/
[5] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[6] http://ja.wikipedia.org/
[7] http://d.hatena.ne.jp/keyword/
[8] http://download.wikimedia.org/

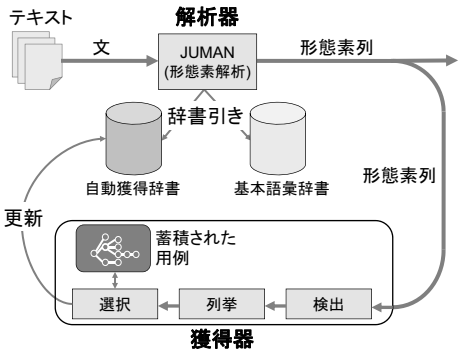

図 1: オンライン語彙獲得システム

4　オンライン語彙獲得

形態素解析における未知語の問題に対して我々が提案する解法は、オンライン語彙獲得 [3, 4] である。オンライン語彙獲得では、図1のように、バッチ処理ではなく、逐次的に入力されるテキストから未知語を獲得する。形態素解析器自体は、通常通りテキストを文単位で解析し、形態素列を出力する。異なる点は、解析の裏で語彙獲得器が動作することである。語彙獲得器は、解析された文から未知語を抽出し、適当な時点で形態素解析器の辞書を更新する。これにより、獲得された未知語が形態素解析に反映される。

獲得の過程を「ついったー」を例に説明する。ある時点で未知語「ついったー」を含む文「ついったーを使う。」が入力されたとする。この入力を形態素解析器は「つ + いった + ー + を +···」と誤解析する。語彙獲得器はこの解析結果に未知語が含まれると判断し、その解釈の候補を列挙する。候補の中には、名詞「ついったー」の他にラ行の動詞「ついる」(「取る」が「取った」と活用するのと同じ) なども含まれる。この時点ではどの解釈が正しいかの判断を保留し、記憶に蓄えておく。さらにテキストを読み進め、別の文「ついったーだ。」や「ついったーで···」が入力されると、同様に解釈の候補を列挙する。このとき、過去に入力されたデータと見比べると、名詞「ついったー」という解釈が、後続要素のバリエーションから、もっともらしいと推測できる。ある時点て、解釈の曖昧性が十分に解消されたと判断すると、形態素解析器の辞書を更新する。

この手法の特徴として次の4個が挙げられる。第一に、入力テキストをオンラインで処理するため、獲得開始時に対象テキストが決まっている必要がない。第二に、複数の入力を見比べて獲得を決めるため、ウェブテキストに頻出する誤字に対してもある程度頑健に動作する。第三に、刻々と辞書が更新されるため、同

じ入力文であっても、処理時点によっては同じ解析結果が返るとは限らない。第四に、オンライン化の副産物として、人間らしい振る舞いをする。少なくとも、巨大なテキストを一括で処理するのではなく、テキストを少しずつ読みながら言葉を覚えると言う点で人間らしい。

5 言葉を覚えるボット

我々は、オンライン語彙獲得のオンラインという性質を生かして、リアルタイムウェブにおける未知語の問題の解決を試みる。そのためのプロトタイプとして、ツイッター上に言葉を覚えるボットを実装した。

このボットは、自分のタイムラインに表示されるつぶやきを定期的に読み、オンライン語彙獲得を行う。ある時点で未知語が獲得されると、

「ふぁぼる:子音動詞ラ行」を覚えた。

のように、覚えた形態素をつぶやく。

つぶやきをオンライン語彙獲得システムに与える際には、ツイッターの慣習にあわせて若干の整形を行う。具体的には、URL や検索用の「ハッシュタグ」を除去する。また、他人のつぶやきを引用しているつぶやきは、「RT」あるいは「QT」というキーワードを手がかりに分割する。

タイムラインに表示されるつぶやきは、フォローしている利用者のものである。したがって、ボットが誰をフォローするかを決める必要がある。現在は以下の基準でフォローを決めている。まず、ボットをフォローした人をフォローし返す。新たにフォローする際には、追加で過去のつぶやきを読みに行く。それとは別に、投稿数が多く被フォローの多い日本語の書き手 (有名人) を我々が選んで手動でフォローしている。こうした利用者には、仮に開発中のボットが何らかの誤作動を起こしても、あまり迷惑が掛からないと考えてのことである。現在は、タイムラインに表示されるつぶやきの大半が、25 名の有名人によるものとなっている。

ツイッターからテキストを得る別の戦略として、大規模なクローリングが考えられる。しかし、以下の理由から現在のところ実装していない。既に大規模化したツイッターのクローリングは容易でない[9]。また、無作為にクロールされたデータは大半が日本語ではないため、つぶやきの言語判定も必要となる。

[9] http://d.hatena.ne.jp/code46/20090329/p1

6 結果と考察

図 2 にボットのタイムラインを示す。ここではボットが動詞「ばずる」 (言葉が buzztter に載ること) を覚えている。オンライン語彙獲得では、獲得を決心するには未知語が何回か出てきた時点だが、獲得の決め手となったつぶやきが、ボットのつぶやきの 3 個下に表示されている。

品詞を考慮した獲得を行うため、「ふぁぼる」のような用言は用言として認識される。そのため、一度獲得されると、「ふぁぼられた」のような活用変化形も正しく解析できるようになる。

図 3 にボットのつぶやきを示す。現状でボットが覚えている言葉は既に存在するものである。実験期間中にはその機会に恵まれなかったものの、生まれたばかりの言葉を覚えることも原理的には可能である。

人名などの固有名詞は、現在は便宜的に普通名詞扱いしている。普通名詞と固有名詞の区別は、品詞識別の手がかりとして用いている形態論的制約だけでは難しいからである。今後は、語彙統語的手がかりをもとに、自動獲得した名詞の細分類を行う予定である [5]。

オンライン語彙獲得の特性上、ある未知語を獲得するまで、その未知語を含んだ入力の解析を誤る。この特性は、オンライン語彙獲得の上にサービスを構築する上で問題となりうる。もっとも、新しい語が生まれた瞬間であれば、最初の何回かのつぶやきの解析を誤ることはやむを得ないと考えている。しかし、現状でボットが覚えているのは、人口に膾炙した言葉である。そもそも、そうした言葉が未知語となっている原因は、獲得開始時の辞書として、形態素解析器 JUMAN のデフォルトの辞書を用いていることに求められる。実際のサービスを構築する場合には、「予習」、つまり、あらかじめ対象と関連するテキストを読み、頻出語彙を獲得した状態を出発点にする必要があるかもしれない。

オンライン語彙獲得システムがボットから受け取るテキストは、現状では規模が小さい。今後大規模化する上での課題の一つは並列化である。オンライン語彙獲得は逐次獲得なので、並列化手法が自明でない。

より重要な課題は、形態素解析における超大語彙の扱いである。従来の形態素解析は、人手で整備した高頻度語の辞書を固定で用い、低頻度の語彙を無視してきた。また、特殊な分野のテキストに解析を適応させる場合には、全体としてはマイナーでも該当分野で頻出の語彙を人手で整備してきた。現在のオンライン語彙獲得は、この分野適応の自動化と位置づけられる。

この分野適応は、ツイッターにはそのままでは適用できない。ツイッターは一つの系であり、明確な分野

図 2: ツイッターのタイムライン

図 3: ボットのつぶやき

境界が存在しないからである。一つの割り切りとして、現在の手法のまま解析の個人化を進めることも考えられる。そうではなく、様々な分野が混在するテキストを一度に扱うと、例えば「サー」の獲得によって、「サーバー」が「サー」と「バー」に過分割されるといった、思わぬ副作用が生じる。大規模テキストにおいて低頻度の語彙も逃さずに正確に扱うことはオープンな問題である。

本稿では語彙に関する問題を扱ったが、ウェブテキストに関しては未対応の問題として、表記や口語的な表現 (文法) に関わる部分が残っている。「楽しく通えますよーー（微笑)」、「あえええーーーーーー・。」といったテキストを形態素解析で正しく処理できない。そのため、オンライン語彙獲得でも、獲得には至らないものの、誤って未知語候補として検出されてしまう。このような現象への対応も今後の課題である。

7 おわりに

本稿では、ツイッターを対象にオンライン語彙獲得を行うプロトタイプを実装し、将来の提案手法の共有化のための考察材料とした。

リアルタイムウェブのもう一つの特徴は、利用者の位置情報や利用者間のネットワークなどの非言語的なメタデータを持つこと。これらを言語情報と組み合わせれば新たなサービスが生み出せるかもしれない。

参考文献

[1] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: An analysis of a microblogging community. In *Advances in Web Mining and Web Usage Analysis*, pp. 118–138, 2009.

[2] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about Twitter. In *Proc. of the 1st Workshop on Online Social Networks*, pp. 19–24, 2008.

[3] Yugo Murawaki and Sadao Kurohashi. Online acquisition of Japanese unknown morphemes using morphological constraints. In *Proc. of EMNLP 2008*, pp. 429–437, 2008.

[4] 村脇有吾, 黒橋禎夫. 語彙獲得のための過分割未知語の検出. 言語処理学会第 15 回年次大会 発表論文集, pp. 324–327, 2009.

[5] 村脇有吾, 黒橋禎夫. 自動獲得された名詞の分類. 言語処理学会第 16 回年次大会 発表論文集, 2010. (to appear).