実テキスト解析をささえる語彙知識の自動獲得

柴田 知秀　村脇 有吾　黒橋 禎夫　河原 大輔
京都大学大学院情報学研究科
{shibata, murawaki, kuro, dk}@i.kyoto-u.ac.jp

1　はじめに

ブログや twitter などの実テキストを用いたアプリケーションを構築するためには、それらのテキストを頑健に解析することが必要不可欠となる。特に、形態素解析での誤りは構文解析や省略解析などの後続する解析での誤りを引き起こすため、形態素解析を高精度に行うことが重要となる。

実テキストの解析では新語や専門用語、固有名詞などの未知語が問題となるが、解析対象の文での出現のみから未知語を認識することやその品詞などを推定することは困難である場合が多い。そこで、Wikipedia や大規模 Web テキストからあらかじめ語彙知識を獲得しておき、それを解析時に利用することで頑健に解析できるようにする。

例えば、以下の文の解析時に語彙知識なく、「爽健美茶」を形態素として認識し、かつ、それが名詞で、上位語が清涼飲料水であることを認識することは困難である。

(1)　ワタシ、爽健美茶派です。

しかし、Wikipedia には「爽健美茶」というエントリがあり、またその記事から上位語が「清涼飲料水」であることを獲得できるので、獲得した情報を形態素解析器の辞書に登録することにより、正しく解析できるようになる。

また、同様に、以下の文の解析時に「カサつく」という動詞を認識するのは困難であるが、Web テキスト上での出現からあらかじめ「カサつく」という動詞を獲得し [1]、それを形態素解析器の辞書に登録することにより、正しく解析できるようになる。

(2)　皮膚がカサついてガサガサする。

本論文では、Wikipedia と Web テキストから語彙知識を自動獲得し、それを形態素解析・構文解析辞書として用いることにより、実テキストを頑健に解析する手法を提案する。図 1 に概要を示す。本論文では形態素解析器として JUMAN、構文解析器として KNP を用いる。形態素解析器 JUMAN では基本語 3 万語は人手で選定し、それらに意味情報を付与しており、残りの語彙に関しては自動獲得するという方針をとっている。

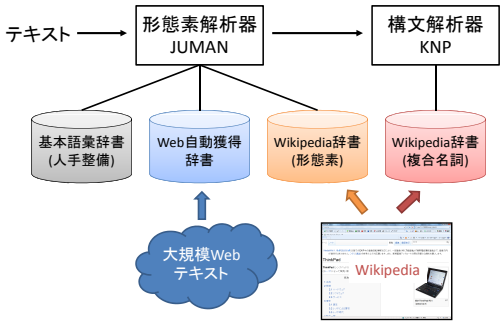

図 1: Wikipedia と Web テキストからの語彙獲得と形態素解析器・構文解析器での利用

Wikipedia のエントリには一形態素のものと複数形態素からなるものが混在しているため、それらの判断を行い、「爽健美茶」「Thinkpad」のような一形態素と判断された語は形態素解析器の辞書として用い、「自然言語処理」「国土交通省」のような複数形態素と判断された語は構文解析器の辞書として登録しておき、解析文中にその語が出現した時に上位語などの情報を付与する。また、大規模 Web テキストからは「カサつく」「ビミョーだ」のような語彙を獲得し、形態素解析器で利用する。

また、編集距離や分布類似度を手がかりとし、基本語と、Web テキストから獲得した語、Wikipedia から獲得した語の間の対応をとり、異表記関係を認識する。例えば、Web テキストから獲得した語である「カサつく」は基本語の「かさつく」と異表記関係であることを認識する。

2　Wikipedia からの語彙獲得

Wikipedia の記事のタイトルから語彙を獲得する。基本語彙における語の基準との整合性をとるために、

すべての語を形態素解析器の辞書に登録するのではなく、一形態素と判断した語を形態素解析器の辞書に登録して用い、それ以外は構文解析時に利用する。

2.1 形態素解析辞書に登録する語の選定

語が一形態素かそうでないかの判断は自明なタスクではない。本論文では、現在の形態素解析器で誤った解析とみなせるものを一形態素であると判断する。具体的には、形態素解析器 JUMAN の解析結果が以下の条件のいずれかを満たすものを一形態素とみなす[1]。

JUMAN の解析結果が未定義語一語になるもの (例: ThinkPad、ミニストップ)

JUMAN の解析結果が一文字形態素のみからなるもの (例: 爽健美茶[2]、霞ヶ浦)

- ただし、以下にあげるような品詞列にマッチするものは除く。
  - 接頭辞 + 数詞 + 接尾辞 (例: 第一次)
  - 地名 + 地名 (例: 日朝)
  - 名詞 + (の|と) + 名詞 (例: 草の根, 男と女)

複数のカタカナ形態素からなる語のうち、語とその主辞の分布類似度が低いもの

カタカナのみからなる語は特に一形態素かどうかの判断が難しい。例えば、「フットサル」の場合、JUMAN による解析では「フット/サル」となるため、複数形態素からなると判断されてしまう。そこで、「フットサル」と「サル」の分布類似度 [4] を計算し、類似度が閾値より低ければ、これは形態素解析誤りであるとみなし、「フットサル」を形態素解析器の辞書に登録することにより、形態素解析時にカタカナ過分割を防ぐ[3]。一方、「アウトレットセンター」のような語の場合、「アウトレットセンター」と「センター」の類似度は高いため、「アウトレットセンター」は複数形態素からなると判断し、形態素解析辞書に登録しない。

[1]ただし、12 文字以上の語は一形態素である可能性が低いため、除去する。

[2]JUMAN の解析では以下のようになる。このような解析になる場合は解析誤りである可能性が高いので、辞書に登録する。
爽 爽 爽 未定義語 15 その他 1 * 0 * 0 NIL
健 けん 健 名詞 6 人名 5 * 0 * 0 "人名:日本:名:22:0.00134"
美 び 美 名詞 6 普通名詞 1 * 0 * 0 "代表表記:美/び 漢字読み:音 カテゴリ:抽象物"
茶 ちゃ 茶 名詞 6 普通名詞 1 * 0 * 0 "代表表記:茶/ちゃ 漢字読み:音 カテゴリ:人工物-食べ物 ドメイン:料理・食事"
EOS

[3]JUMAN では、「フットサル」「フット」「サル」が辞書にある場合、「フットサル」の解析は「フットサル」一語となるため、「フットサル」を形態素解析辞書に登録する。

表 1: 品詞細分類の決定 (上位語の下線部は主辞を表す)

見出し語	上位語	JUMAN カテゴリ	品詞 細分類
ロナウジーニョ	サッカー 選手	人	人名
兼六園	日本 庭園	場所-施設	地名
ダイソー	会社	組織・団体	組織名
	...		
(上記にマッチしなければ普通名詞)			
インクィジター	アクション 小説	抽象物	普通名詞

2.2 意味情報の付与と品詞細分類の決定

Wikipedia を用いるメリットとして、記事の情報やリダイレクトから語彙情報を獲得できることがある。

定義文 (記事の 1 文目) から Sumida らの手法 [3] と同様の手法で上位語を獲得し、意味情報に Wikipedia 上位語という素性を付与する。

品詞細分類としては普通名詞、人名、地名、組織名を考え、上記で獲得した上位語の主辞の JUMAN カテゴリにより、品詞細分類を決定する。表 1 に例を示す。

3 Web テキストからの語彙獲得

我々は以前、生テキストから未知語を自動獲得する枠組みを提案した [1]。語彙獲得の当初の目的は、中規模の生テキストから低頻度語やドメイン固有の語を獲得することであった。本論文では、これに加えて、大規模 Web テキストに対して語彙獲得を実行し、比較的高頻度の語を選択する。

語彙獲得システムはテキストを逐次的に読み込みながら獲得を行う。そのままでは大規模ウェブテキストを処理できないため、Web テキストを分割し並列処理を行う。

この段階で獲得された語に付与されている品詞は、本来の品詞体系から名詞の細分類を省いたものである。そこで、次に、語彙統語的選好を手がかりとして、獲得された名詞に対して完全な品詞を割り当てる [2]。また、可能動詞に対して、対応する基本動詞を記述する。可能動詞と基本動詞の対応は規則的であり、記述は自動で行える。

4 異表記関係の認識

JUMAN では表記揺れを解消するために代表表記を与えている。例えば、基本語彙辞書中の「奇麗だ」と「綺麗だ」にはともに代表表記「綺麗だ/きれいだ」が付与されている。自動獲得した語の中には基本語と異表記関係にあるものが含まれるので、異表記関係を認

識し、それに基本語と同じ代表表記を付与することにより、表記揺れを解消する。

4.1 Wikipedia 辞書 (形態素)

Wikipedia のリダイレクトと基本語彙辞書の形態素が、カタカナをひらがなに正規化した上で編集距離が近い場合に異表記関係と認識し、基本語の代表表記を Wikipedia 辞書の語に付与する。例えば、「マツゲ」は「まつげ」にリダイクレトされているため、「マツゲ」は基本語の「まつげ」の異表記と認識し、「マツゲ」に代表表記「まつ毛/まつげ」を付与する。

また、Wikipedia から獲得した語同士で異表記関係にあるものがある。Wikipedia において、語 A から語 B にリダイレクトがあり、カタカナをひらがなに正規化し編集距離が小さい場合に、語 B の「表記/読み」を代表表記として語 A、B に付与する。例えば、「スパゲッティ」「スパゲティ」「スパゲティー」に対してすべて代表表記「スパゲッティ/スパゲッティ」が付与される。

4.2 Web 自動獲得辞書

Web 自動獲得辞書と基本語彙との異表記関係の認識を行う[4]。本論文では 2 種類の異表記関係について認識を行う。

1 つ目は漢字の異体字である。例えば、異体字を含む自動獲得語「大學」は基本語「大学」の異表記である。異体字の認識には、Unihan データベース[5]を用い、データベースに登録されている異体字について総当りで探索を行う。

2 つ目は、Web テキストによく見られる非規範的な表記である。例えば、自動獲得語「カサつく」は基本語「かさつく」に対応する[6]。このような異表記関係を認識するために、まず候補となる形態素のペアを列挙する。候補の列挙には、重みを人手で調整した編集距離を用いる。非規範的な表記は漢字の異表記ほど自明ではない。例えば、編集距離だけを手がかりとすると、自動獲得語「アワー」と基本語「アワ」のペアが候補に挙がる。このような候補を取り除くため、候補のペアがテキスト中で似たような振る舞いをしているかを調べる。具体的には、分布類似度 [4] が閾値以下の候補を取り除く。名詞ペアについて、係り受け関係に

[4]5 節でも述べるように、Wikipedia 辞書と自動獲得語の重複を取り除くため、Wikipedia から獲得された語との関係は無視できる。

[5]http://www.unicode.org/reports/tr38/

[6]長音挿入 (例:「すごい」と「すごーい」) や小文字挿入 (例:「行きたい」と「行きたぁい」) のような非規範的な表記は JUMAN が解析時に動的に認識する。

表 2: Wikipedia 辞書 (形態素) の例 ([上] は上位語、[代] は代表表記を表す)

見出し語	品詞	品詞細分類	意味情報
爽健美茶	名詞	普通名詞	[上] 清涼飲料水
イチロー	名詞	人名	[上] プロ野球選手
祇園	名詞	地名	[上] 歓楽街
ＧＬＡＹ	名詞	組織名	[上] ロックバンド
マツゲ	名詞	普通名詞	[代] まつ毛/まつげ
スパゲティ	名詞	普通名詞	[代] スパゲッティ/スパゲッティ
スパゲティー	名詞	普通名詞	[代] スパゲッティ/スパゲッティ

表 3: Wikipedia 辞書 (複合名詞) の例 (複合名詞中の「+」は形態素区切りを表す)

複合名詞	付与する情報
湯川+秀樹	[上] 理論物理学者
ラファエル+・+ナダル	[上] 男子プロテニス選手
スーパー+カミオカンデ	[上] ニュートリノ検出装置
ＴｈｉｎｋＰａｄ+６００	[上] ノートパソコン

ある用言の重複を用い、同様に、用言については係り受け関係にある名詞の重複を用いる。自動獲得語と基本語の分布類似度を計算するためにはまず自動獲得した語を形態素解析器の辞書として登録し、Web コーパスに対して形態素解析・構文解析を行う。そして構文解析結果から分布類似度を計算することにより異表記関係を認識し、その情報を自動獲得辞書に付与する。

5 獲得された辞書と解析例

5.1 獲得された辞書

日本語 Wikipedia のダンプ[7]から語彙を獲得した。2011 年 12 月時点で約 205 万記事あり、そのうち約 14 万語を JUMAN の辞書に登録し、KNP の辞書に 80 万語登録した[8]。それぞれ表 2、表 3 に例を示す。また、Wikipedia 辞書 (形態素) 中の 136 個の語が基本語との異表記と認識され、Wikipedia 辞書 (形態素) 中の語約 1 万語おいて異表記関係が認識された。Wikipedia のダンプは月に 1 回程度更新されており、システムは自動的にダウンロードし、語彙を獲得し、辞書に登録する。

また、Web テキスト 1 億ページから語彙獲得を実行し、12,952 語が獲得された。Wikipedia 辞書との重複を取り除いた結果、6,692 語が残った。また、140 個の異表記関係が認識された。辞書の例を表 4 に示す。

[7]http://download.wikimedia.org/jawiki/

[8]残りの記事は例えば「～の一覧」(例:国の一覧) や「～年」(例:2011 年) などであり、ルールでフィルタリングしている。

矢印 を 伸ばして る と アンカー 位置 が 勝手に ズレる ・ ・ ・ 。
ゼウス の 陰謀 だ 。
...
携帯 電話 の QR コード 読取 機能 を 利用 して 、 スタンプ ラリー を 開催 する システム です 。
...
インフォカート で は 売り 切り の 情報 商材 の 販売 だけ で は なく 、 有料 メルマガ など の 継続 型 の 商品 販売 も 可能 。
そして その すべて の 商品 に アフィリエイト プログラム が 用意 さ れて いる 。
インフォカート で は 情報 起業 家 、 アフィリエイター の 双方 に メリット が ある アフィリエイト が できる 。
...

図 2: Web テキストの解析例 (下線をひいた語は Wikipedia から獲得した形態素、二重下線をひいた語は Web から獲得した形態素、四角で囲った語は Wikipedia から獲得した複合名詞を表わす)

表 4: Web 自動獲得辞書の例

見出し語	品詞	品詞 細分類	意味情報
がんがる	動詞	-	
カサつく	動詞	-	[代] かさつく/かさつく
アジャイルだ	形容詞	-	
ビミョーだ	形容詞	-	[代] 微妙だ/びみょうだ
待受	名詞	普通名詞	
大學	名詞	普通名詞	[代] 大学/だいがく

これらの辞書は形態素解析器 JUMAN7.0[9]および構文解析器 KNP4.0[10]に同梱し配布している。

5.2 解析例

自動獲得した語彙を用いて Web テキストを解析した例を図 2 に示す。また、Web テキストに対して自動辞書を用いない場合と用いた場合の解析変化を表 5 に示す。形態素解析誤りが改善している、または、形態素区切りは変化していないが辞書として獲得されていることにより未定義語ではなくなり、情報が付与されていることがわかる。「ファイルサーバ」は自動獲得した辞書を用いることにより、一語となったが、これは「ファイル/サーバ」の方が望ましく、カタカナに関する誤りは今後の課題である。

6 おわりに

本論文では、Wikipedia と Web テキストから語彙知識を自動獲得し、それを形態素解析辞書・構文解析辞書として用いることにより、実テキストを頑健に解析する手法を提案した。今後の課題としては構文解析や省略解析などの高次の解析での語彙情報の利用、語彙の意味知識を利用した形態素解析などがあげられる。

[9] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[10] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

表 5: Web コーパスでの形態素解析変化の例 (括弧内は品詞を表す)

自動獲得辞書なし	自動獲得辞書あり
Wikipedia 辞書によるもの	
トラック (名詞) バック (名詞)	トラックバック (名詞)
ＴＥＬ (未定義語)	ＴＥＬ (名詞)
ＤＶＤ (未定義語)	ＤＶＤ (名詞)
粉 (名詞) 引 (未定義語)	粉引 (名詞)
販 (未定義語) 促 (未定義語)	販促 (名詞)
琉 (未定義語) 球 (名詞)	琉球 (名詞)
ＴＯＹＯＴＡ (未定義語)	ＴＯＹＯＴＡ (名詞)
ファイル (名詞) サーバ (名詞)	ファイルサーバ (名詞)
Web 自動獲得辞書によるもの	
オススメ (未定義語)	オススメ (名詞)
釣 (名詞) 果 (名詞)	釣果 (名詞)
魅 (未定義語) せる (動詞)	魅せる (動詞)
ロハス (未定義語) な (判定詞)	ロハスな (形容詞)

参考文献

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

[2] Yugo Murawaki and Sadao Kurohashi. Semantic classification of automatically acquired nouns using lexico-syntactic clues. In *COLING: Poster Volume*, pp. 876–884, 2010.

[3] Asuka Sumida and Kentaro Torisawa. Hacking Wikipedia for hyponymy relation acquisition. In *Proc. of IJCNLP*, pp. 883–888, 2008.

[4] 柴田知秀, 黒橋禎夫. 超大規模ウェブコーパスを用いた分布類似度計算. 言語処理学会 第 15 回年次大会, pp. 705–708, 2009.3.