機械学習を用いたカタカナ用言の獲得

福島健一　　鍛治伸裕　　喜連川優
東京大学 生産技術研究所
{ken, kaji, kitsure}@tkl.iis.u-tokyo.ac.jp

1　はじめに

日本語形態素解析は単語辞書の知識に強く依存している。形態素解析器の典型的な実装では、辞書に登録されている単語を使って入力文をラティス構造で表現し、そのラティス中の経路から適切な解析結果を探索する。このとき入力中に未知語（辞書に登録されていない単語）が含まれていると正しい解析結果を導くラティスを生成することができず、解析に失敗してしまう。このような未知語の問題は日本語形態素解析における解析誤りの主要な原因のひとつである。

未知語問題への取り組みとして2つのアプローチが考えられる。ひとつは、解析時に未知語に出会ってもロバストに処理を進められる未知語処理ルーチンの設計である。現在の形態素解析器の実装では字種を手がかりにした簡単なルールで未知語を処理している。そしてもうひとつのアプローチは、地道に辞書に登録されている単語を増やしていくというものである。辞書に多くの単語が登録されているほど、解析時に未知語に出会う確率は少なくなり、解析精度の向上が期待できる。世の中で使われている言葉は非常に多様であり、また日々新しい言葉が生み出されていることを考えると、人手で未知語を発見して登録していく方法には限界があり、できるだけ自動で未知語を発見できることが望ましい。

我々は辞書の整備という観点から、未知語の中でも特にカタカナ用言に注目しその自動獲得を試みた。カタカナ用言とは「ググる」「ウザい」「イケテる」などのようにカタカナの語幹とひらがなの活用語尾からなる用言のことである。

カタカナ用言は新語やくだけた口語表現に多いが、こうした語彙が頻繁に現れる典型的なシチュエーションとしてウェブがある。ウェブ上のテキストを解析するうえで、これらの語彙を正しく取り扱えるようになることの意義は大きい。

2　カタカナ用言の自動獲得

カタカナ用言は、自動獲得しやすい性質をいくつか持っている。まず、字種の違いを手がかりにして容易に語幹の位置を同定することができる。さらに用言は規則的な活用語尾という顕著な特徴を持つので、カタカナ文字列に後続する文字を見ればそれが用言であるか否かを判断できる。

我々が提案する手法では、テキストコーパス中に出現するカタカナ文字列が用言の語幹であるか否かを、その「使われ方」から推定する。動詞、形容詞などの用言の品詞ごとに、各文字列がその品詞の活用語幹であるかそうでないかを判断する。したがって、文字列によっては二つ以上の品詞と判断されることもありうる。例えば「エグ」という文字列は同時に動詞「エグる」とも形容詞「エグい」とも見なされる可能性がある。

文字列の「使われ方」とは、その文字列の周辺に現れる文字のことである。例えば、用言は活用語尾を、また用言に限らず内容語は機能語をともなって文中に出現するが、この活用語尾や機能語には内容語の品詞や活用型ごとに固有のパターンがある。動詞の場合はさらに細かい活用型に分かれる。典型的な例として、

- 形容詞の活用語尾「かっ」
- 動詞・ラ行五段活用の活用語尾「る」
- 動詞・サ行変格活用の活用語尾「する」
- 名詞に後続する助詞「を」

などがある。この性質を利用して、ある文字列が用言か否かを、それが実際にコーパス中で使われているパターンに基づいて判断することができる。

3　教師付き学習による定式化

前節の考察を踏まえた上で、カタカナ文字列の品詞推定問題を教師付き学習による分類問題として定式化

する。任意の文字列がある品詞の用言の語幹であるかどうかを判断する分類器を品詞の数だけ作成する。

3.1 素性設計

まず文字列の性質を素性ベクトルで量的に表現する。原理的には文字列の周辺のあらゆる文字 n-gram を素性として用いることができるが、今回は計算量の節約のために直後のひらがな 1-gram〜5-gram のみとした。2 節の例のように、活用語尾や後続する機能語は内容語の品詞に強い制約を受けるが、活用語尾や機能語はほぼすべてがひらがなである。従って後続するひらがな n-gram は他の情報よりも有用だと判断した。ある文字列に対して、各 n-gram が後続する回数を文字列の出現回数で割り、その対数値を各素性の値とする。

3.2 分類先の品詞

自動獲得を行う用言として、今回は以下の品詞・活用型を対象とした[1] 。

- 形容詞
- 形容動詞
- 動詞–五段・ラ行
- 動詞–一段
- 動詞–カ行イ音便

これに加えて、教師データ作成の際には負例として次の品詞も考慮する。

- 名詞
- 副詞

現実には感動詞や助動詞などとしてカタカナ語が使われる場合もあるが、本研究ではすべて内容語だと仮定して議論する。

ある品詞・活用型の用言の獲得を試みる場合、教師データに出現するその品詞・活用型の全単語を正例、それ以外の品詞・活用型のすべての単語を負例として分類器を学習させる。

[1]品詞及び活用型の分類は IPADIC 品詞体系に基づく。

表 1: 教師データ

品詞	異なり単語数	延べ出現回数
形容詞	1,398	43,603,917
形容動詞	2,792	37,625,714
動詞–ラ行五段	2,210	69,451,686
動詞–一段	4,632	74,560,319
動詞–カ行イ音便	788	16,300,137
名詞	103,968	621,827,116
副詞	2,671	64,990,874

4 実験と獲得した用言

4.1 実験結果の概要

実験は東京大学喜連川研究室が収集しているウェブページから抽出した日本語文約 1.6 億文を対象に行った。これを MeCab[2]で形態素解析し、10 回以上出現した単語、10 回以上出現した素性のみを選択し教師データとした。素性の数は 1,193,436 個であった。表 1 に教師データ中の各品詞の統計を示す。

表 1 の各用言について、その品詞を正例、それ以外の用言および名詞と副詞を負例として分類器を作成する。分類モデルには SVM を使い、SVM の実装は TinySVM[3]を用いた。

教師データを作成したのと同じテキストデータからカタカナ文字列を抽出し、各文字列と各品詞の組み合わせについて、その文字列がその品詞であるか否かを分類器を用いて判断する。統計的な信頼性を担保するために出現頻度が 10 回以上のカタカナ文字列のみを対象とした。異なり単語数は 250,832、延べ出現回数は 110,316,818 である。表 2 の獲得数の欄に分類器が正例と判断した文字列の数を示す。

提案手法がどの程度の精度で用言を獲得できるのかを見積もるために、分類器が正例と判断した文字列から各品詞ごとに 100 個をランダムサンプリングし（獲得数の少ない動詞–カ行イ音便はすべて）、それらが本当にその品詞として使われるのかを人手で確認した。文字列の用法のチェックには、実験に用いたコーパスでの出現箇所を見るだけでなく、ウェブ検索エンジンも用いた。獲得した用言のうち、確かにその品詞としての用法を持つ文字列の割合を表 2 の適合率の欄に示す。

また図 1 は、獲得した用言を SVM 分類器が付与したスコアの降順に並べ、その上位 α% 点を閾値とした

[2]http://mecab.sourceforge.jp/
[3]http://chasen.org/ taku/software/TinySVM/

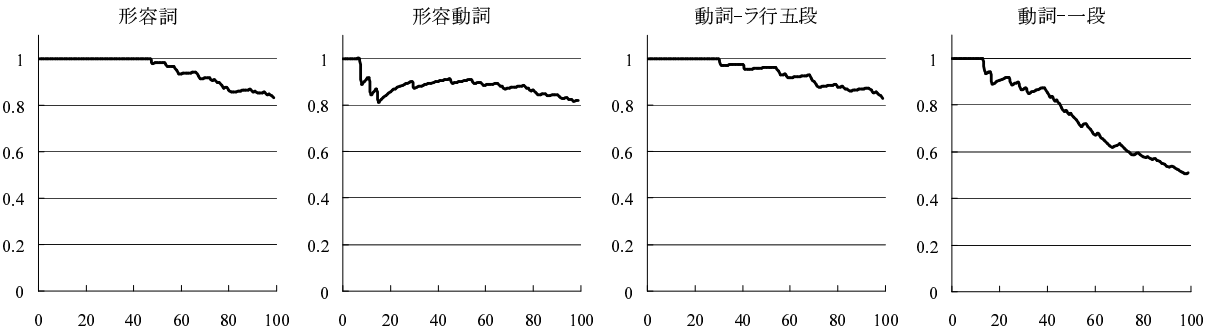

図 1: 適合率の変化（横軸が上位 α% 点、縦軸が適合率）

ときの適合率のグラフである（動詞– カ行イ音便を除く）。形容詞、形容動詞、動詞–ラ行五段では適合率が緩やかに減少しているが、動詞–一段ではスコアの高いところでは適合率が 1 に等しい極めて高い水準で推移し、ある点から急降下していることが読み取れる。これはあらかたの用言はその点までに採り尽くされてしまっていることを示唆する。従って、用言か否かの判断の基準を SVM のスコアが 0 の点よりも高く設定すれば獲得数をあまり犠牲にせずに適合率を高めることが可能である。

分類器が誤って正例と判断してしまう文字列には、品詞ごとに固有の傾向が見られた。動詞–ラ行五段および動詞–一段では擬態語・擬声語の類の副詞が、形容詞では「デカイ」などの形容詞終止形と「イヤラ（しい）」などの語幹にひらがなを含む形容詞が誤判定の大きな部分を占めていた。形容詞の誤判定は当初想定していなかった形態のカタカナ用言に因るものであり、これらを正しく扱うことによる精度の向上の余地がある。

表 2: 獲得したカタカナ用言の数

品詞	獲得数	適合率
形容詞	265	0.83
形容動詞	2173	0.82
動詞–ラ行五段	224	0.83
動詞–カ行イ音便	11	0.82
動詞–一段	515	0.51

4.2 獲得した用言の例

表 3 は実際に獲得した用言の一例である。「ググる」のようなウェブで生まれた典型的な新語や「チャチい」などの評価表現のほかにもさまざまなカテゴリの用言を獲得することができた。趣味などを共有する特定のコミュニティ内のみで使われる表現（「ハラシマる」「バモる」「イナタい」）や、本来は表記ミスや文法的に誤りであっても現実には多くの人が使っている表現（「ガンガる」「スゲい」）が多く観察された。

5 関連研究

辞書整備を目的とするコーパスからの語彙の自動獲得について、これまでにいくつかの研究が報告されている。

森ら [3] は品詞の出現パターンをベクトルで表現した。品詞タグ付けコーパスを使って品詞ごとにその特性を表現するベクトルを計算し、ある文字列のベクトルと各品詞のベクトルのユークリッド距離を使ってその文字列の品詞らしさを算出する。ベクトルの各成分は文字列の前後に隣接する文字 n-gram である。品詞ごとに固有の出現パターンがあるという着想は本研究と同じだが、森らの手法は品詞の特質をひとつのベクトルのみで表現している点が本研究と異なる。文法体系では同一の品詞として定義されている単語でも、おのおのの単語によって出現パターンは実はかなりばらつきがある。教師データにおけるばらつきを残したまま推定を行える点は SVM などの分類アルゴリズムの強みである。

中澤ら [2] は複合カタカナ名詞の抽出と分割を行った。形態素解析器が利用する辞書では単語を単に増やすだけでなく複合語などの冗長性を排除することも重要である。中澤らは単語の分割位置を決定するために辞書などの言語リソースを活用している。カタカナ用言においても「マジヤバい」（副詞＋形容詞）や「エ

表 3: 獲得した用言の例

形容詞	形容動詞	動詞–ラ行五段	動詞–一段	動詞–カ行イ音便
チャチい	オッサレーだ	ググる	シビレる	シゴく
ヤヴァい	エイジレスだ	ガンガる	ノレる	ボヤく
イナタい	イタリアーンだ	ハウる	ホレる	トキメく
カバエい	レピッシュだ	ハラシマる	モエる	ホザく
ウマーーい	ピーピーだ	バモる	イケテる	ドツく
ヘコい	ママモードだ	ヒキコモる	ハッチャケる	シバく
ギトい	エコロジックだ	ストロベリる	トロケる	ムカツく
カッコカワイい	グッチョイだ	バザる	シャガレる	イタダく
マンドクサい	ナツラルだ	ホバる	コジャレる	ワメく
スゲい	グレイッシュだ	シャシャる	デケる	

ロカワイい」（形容詞＋形容詞）といった複合パターンがあり、取り組むべき課題だと考えている。

また、未知語問題へのもうひとつのアプローチであるロバストな未知語処理ルーチンの設計では、形態素ラティスに未知語を認識する機構を取り入れる手法を中川 [1] や東ら [4] が提案している。

6 おわりに

本稿では機械学習を用いてテキストコーパスからカタカナ用言を獲得する手法を提案し、実際に大規模なコーパスに対して行った実験の結果を報告した。実験の結果、品詞によっては数百の用言が獲得でき、精度も 8 割程度と現実的であることが確かめられた。今後はさらに獲得数と精度の向上を目指すと同時に、用言以外の単語も扱えるようなより洗練されたモデルを追求していく。

参考文献

[1] Tetsuji Nakagawa. Chinese and Japanese word segmentation using word-level and character-level information. In *Proceedings of COLING–2004*, pp. 466–472, 2004.

[2] Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. Automatic acquisition of basic katakana lexicon from a given corpus. In *Proceedings of IJCNLP–2005*, pp. 682–693, 2005.

[3] 森信介, 長尾眞. n グラム統計によるコーパスからの未知語抽出. 情報処理学会論文誌, Vol. 39, No. 7, pp. 2093–2100, 1998.

[4] 東藍, 浅原正幸, 松本裕治. 条件付確率場による日本語未知語処理. 情報処理学会研究報告 2006-NL-173, pp. 67–74, 2006.