最長後続ひらがな列に基づく品詞・活用型の自動推定

桑江 常則 † 藤田 篤 † 佐藤 理史 †
† 名古屋大学大学院工学研究科
kuwae@sslab.nuee.nagoya-u.ac.jp,{fujita,ssato}@nuee.nagoya-u.ac.jp

1 はじめに

言語の経時変化の一つに、新語の出現がある。例えば、「ハズい」、「キモい」、「コピる」などのいわゆるカタカナ用言は、若者を中心にかなり定着してきている。一方で、「ちゃち（な）」のような形容動詞が、「ちゃちい」、「ちゃちくない」のような形容詞として使用される例も見られるようになってきた。これは、誤用とみなすこともできるが、そのような使用が定着したのであれば、転用あるいは派生とみなさざるをえない。このような変化も、言語の経時変化の一形態である。

上記のような新語や転用・派生語を収集して人間向けあるいは形態素解析用の辞書を作成する場合、品詞や活用型を認定する必要がある。このような背景から、我々は、与えられた語幹の品詞・活用型を自動的に推定する方法について検討している。これまでに、次のような特徴をもつ手法を提案した [3]。

- 既存の福島らの手法 [1] と同様、用言であれば活用語尾、名詞であれば「が」や「を」などの助詞を後続するように、品詞・活用型ごとに後続しやすいひらがな列が異なる点に着目する。
- 語幹「通」に対する「通る」（五段・ラ行）、「通う」（五段・ワ行促音便）のように、同じ語幹に対する複数の品詞・活用型を認定するために、福島らとは異なり、各品詞・活用型に特徴的な後続ひらがな列の頻度のみを参照する。

しかしながら、この手法では、複数の品詞・活用型が同じ後続ひらがな列を共有する場合に、品詞・活用型を誤って認定してしまう問題があった。これは、例えば、五段・マ行と五段・ナ行がともに「んだ」を後続しやすいため、「飲（む）」という五段・マ行の動詞の語幹を、誤って五段・ナ行と認定してしまうというものである。

本稿では、この問題に対する対処法を示す。そして、既知語を用いた定量的評価の結果およびカタカナ文字列に対する品詞・活用型の推定結果について述べる。

2 提案手法

本節では、まず 2.1 節と 2.2 節で、ベースとなる品詞・活用型推定手法 [3] について説明する。次に、2.3 節で問題点とその解決策について述べる。

2.1 各品詞・活用型に特徴的な最長後続ひらがな列

各品詞・活用型の直後に出現する活用語尾や助詞は、品詞・活用型を推定する強力な手がかりとなる。ただし、実際には、活用語尾の終了点が明確には定まらない場合がある。そこで、活用語尾を厳格に認定するのではなく、語幹に対する最長後続ひらがな列を利用する。

さらに、各品詞・活用型 c に対して、他の品詞・活用型よりも出現しやすい最長後続ひらがな列を **c の特徴ひらがな列**とする。そして、その集合 $H(c)$ のみに基づいて品詞・活用型を推定する。推定に先立ち、各 c に対する $H(c)$ を、形態素解析済みのコーパス中の既知語の情報に基づいて、次の手順で決定しておく。

1. 各品詞・活用型 c とその最長後続ひらがな列 h の出現頻度をそれぞれ集計する。その際、以下の 2 つもそれぞれ最長後続ひらがな列の一種とみなす。
   (a) 語幹の直後に読点が続く場合 $h =$“、”
   (b) ひらがな列も読点も後続しない場合 $h =$“ϕ”
2. 各 c の特徴ひらがな列の集合 $H(c)$ を次式で決定する。

$$H(c) = \{h \mid P(h|c) - P(h|\bar{c}) \geq 0.001\} \quad (1)$$

ここで、$P(h|c)$、$P(h|\bar{c})$ はそれぞれ次のことを表す。

$P(h|c)$ = “h が品詞・活用型 c の最長後続ひらがな列として現れる確率”

$P(h|\bar{c})$ = “h が c 以外の品詞・活用型の最長後続ひらがな列として現れる確率”

このようにして得られる $H(c)$、$P(h|c)$、$P(h|\bar{c})$ を用いて品詞・活用型を推定する。

2.2 品詞・活用型の認定手順

与えられた語幹 x が各品詞・活用型 c をとるか否かを、$H(c)$ に含まれる特徴ひらがな列 h のみに基づいて判定する。具体的な手順を以下に示す。

まず、c が N 回出現したときに期待される h の出現回数 $u(c,h,N)$ と c 以外の品詞・活用型 $\bar{c}$ が N 回出現したときに期待される h の出現回数 $l(c,h,N)$ を求める。

$$u(c,h,N) = \lfloor N \cdot P(h|c) \rfloor$$
$$l(c,h,N) = \lceil N \cdot P(h|\bar{c}) \rceil$$

これらを用いて、語幹 x が c をとることのスコアを次式で求める。

$$score(x,c,N) = \frac{\sum_{h\in H_N(c)} e\big(c,h,N,freq(x,h)\big)}{\sum_{h\in H_N(c)}\big(u\,(c,h,N) - l\,(c,h,N)\big)} \quad (2)$$

ただし、$freq(x,h)$ はひらがな列 h が推定対象の語幹 x の最長後続ひらがな列として出現する回数であり、$H_N(c)$、e は次式で与える。

$$\begin{aligned} H_N(c) &= \{h \in H(c) \mid u(c,h,N) - l(c,h,N) \geq 1\} \\ e(c,h,N,v) &= \max\Big(\min\big(u(c,h,N),v\big) - l(c,h,N), 0\Big) \end{aligned}$$

$score(x,c,N)$ の値域は [0,1] である。この値が閾値 th_s 以上ならば、語幹 x が品詞・活用型 c をとると認定する。

提案手法は、次の 2 つのパラメタで品詞・活用型を認定する際の厳密さを制御する。

- 仮定する出現数 N
- 特徴ひらがな列の頻度に対する充足度 th_s

2.3 特徴ひらがな列を共有する品詞・活用型の再判定

前稿 [3] における調査において、2.2 節で述べた手法は、同じ最長後続ひらがな列が複数の品詞・活用型の特徴ひらがな列となる場合に品詞・活用型を誤って認定してしまうことが明らかになった。例えば、五段・ナ行（「死ぬ」など）、五段・バ行（「飛ぶ」など）、五段・マ行（「飲む」など）は特徴ひらがな列に「んだ」を共有するため、「飛」や「飲」などの語幹を誤って五段・ナ行と認定してしまった。

この問題を避けるため、2.1 節、2.2 節で述べた手法に以下の処理を加えた。

前処理: 他の品詞・活用型と共有し、かつそれのみでスコアが閾値以上となる特徴ひらがな列を持つ品詞・活用型群 C を特定しておく。ただし、そのような品詞・活用型群は、N と th_s に応じて変わる。そして、その品詞・活用型群のみを見た場合の各品詞・活用型の特徴ひらがな列の集合 $H^C(c)$ を求めておく。これには、式 (1) を用いる。

例 品詞・活用型群 C_1 = { 五段・ナ行, 五段・バ行, 五段・マ行 } に対する再判定を考える場合、H^{C_1}(五段・ナ行) は、五段・ナ行を c、C_1 中の五段・ナ行以外の品詞・活用型を $\bar{c}$ として求める。

品詞・活用型の再判定: ある語幹が $H(c)$ に基づいて $c \in C$ をとると認定された場合は、$H^C(c)$ に基づいて各品詞・活用型 $c \in C$ をとるか否かを再度判定する。これには、式 (2) を用いる。

3 既知語を用いた性能評価

コーパス中の各既知語に対して実際に使われた品詞・活用型を推定することで、提案手法の性能を評価した。

3.1 設定

本実験では、次の 15 種類の品詞・活用型を推定対象とする。

動詞・一段、動詞・五段・各行 11 種[1]、
形容詞、形容動詞、サ変名詞

前稿 [3] では N を固定して品詞・活用型を推定した。これに対して今回は、仮定する出現数 N を動かしながら高頻度の品詞・活用型から順に認定する。具体的にはまず、$N = 1024$ とする。このとき認定されなかった品詞・活用型に対しては N を半分にして判定する。これを $N = 64$ になるまで繰り返す。

スコアの閾値 th_s は、前稿 [3] と同様 0.7 とした。

3.2 データセット

3.2.1 モデルの学習

毎日新聞 2005 年版を形態素解析器 ChaSen[2] によって解析した結果から、15 種類の各品詞・活用型および一般名詞、固有名詞に対して最長後続ひらがな列を抽出した[3]。そして、全体で 100 回以上出現した最長後続ひらがな列から各品詞・活用型の特徴ひらがな列の集合 $H(c)$ を定めた。$N = 64$ のとき、各品詞・活用型の認定に用いられる特徴ひらがな列 $H_{64}(c)$ の一覧を表 1 に示す。

表 1 に示したように、他の品詞・活用型と共有し、かつそれのみでスコアが閾値以上となる特徴ひらがな列が存在する。N が小さいほど、このような特徴ひらがな列による誤認定の可能性が高くなる。そこで、本実験における N の最小値 $N = 64$ のときに特徴ひらがな列を共有する品詞・活用型を、再推定が必要な品詞・活用型群とした（表 2)。そして、各品詞・活用型群 C において、$N = 64$ のときに各品詞・活用型の再推定に用いる $H^C(c)$ を決定した。この結果、表 3 に示すように他の品詞・活用型と共有していた特徴ひらがな列が各品詞・活用型において特徴的、もしくは支配的でなくなった。ここで、C_3 は C_2 の部分集合となっている。このような場合、C_2 内で $c \in C_3$ と認定された語幹のみに対して、C_3 内の品詞・活用型を判定するという手順をふむ。

[1] カ行イ音便、カ行促音便、ガ行、サ行、タ行、ナ行、バ行、マ行、ラ行、ワ行ウ音便、ワ行促音便。

[2] http://chasen-legacy.sourceforge.jp/

[3] 語幹「ない」、「ある」、「する」、「なる」は特殊な活用語尾を持つため除外した。

表 1: 全品詞・活用型の $H_{64}(c)$
各ひらがな列のうち、太字は複数の品詞・活用型に共通、
下線はそれのみでスコア ≥ 0.7 になりうる。

品詞・活用型	特徴ひらがな列（$u-l$ の降順）
一段	た，**る**，て，、（読点），ている，ません
五段・カ行イ音便	**く**，いた，**き**，いて，かれた
五段・カ行促音便	**った**，**く**，かない，**き**
五段・ガ行	ぐ，ぎ，いで，いだ，いでいる，ぐため
五段・サ行	す，**した**，し，して，している
五段・タ行	つ，ち，**った**，**って**，っている，たない
五段・ナ行	**んだ**，ぬ
五段・バ行	ぶ，**んだ**，び，ばれた，ばれる，**んで**
五段・マ行	む，**んだ**，み，**んで**，んでいる，まれた まれる
五段・ラ行	**った**，**る**，り，ると，**って**
五段・ワ行ウ音便	**う**，いの，**い**
五段・ワ行促音便	**う**，**った**，**い**，**って**，われた
形容詞	**い**，**く**
形容動詞	な，に
サ変名詞	**した**，を，する

表 2: $N = 64$、$th_s = 0.7$ の場合にそれのみでスコア ≥ 0.7 になりうる特徴ひらがな列と、それを共有する品詞・活用型群

h	h を特徴ひらがな列に持つ品詞・活用型
んだ	五段・ナ行、五段・バ行、五段・マ行　(C_1)
い	形容詞、五段・ワ行ウ音便、五段・ワ行促音便　(C_2)
う	五段・ワ行ウ音便、五段・ワ行促音便　(C_3)

3.2.2 評価データ

品詞・活用型推定の対象となる語幹は、毎日新聞 2004 年版を ChaSen によって形態素解析した結果において、以下の条件をすべて満たす 15,359 件とした。

- 先頭の文字が漢字である
- 50 回以上出現する

本実験では、コーパス中で使われた品詞・活用型を認定すべき品詞・活用型とする。この際、推定対象とする語幹の品詞・活用型の出現回数 M も集計しておく。

3.3 評価結果

実際にコーパス中で使われた品詞・活用型を認定するといっても、どの程度出現した品詞・活用型を認定すべきかは自明ではない。そこで、実際の出現回数 M に対する閾値 th_M を設け、$M \geq th_M$ である品詞・活用型を認定すべき品詞・活用型とみなすことにした。

th_M を変化させたときの各 th_M における再現率と精度を表 4 に示す。th_M と再現率の関係から、提案手法は、高頻度の品詞・活用型ほど正しく認定できるということが分かる。この際、$M < th_M$ である品詞・活用型を認定した場合は誤認定とみなすため、精度は見かけ上低くなっている。

3.4 考察

3.4.1 認定もれの分析

語幹や品詞・活用型によっては最長後続ひらがな列に偏りがあり、その品詞・活用型の特徴ひらがな列が出現しにくい場合がある。それでも我々は、仮定する出現数 N の 2 倍以上出現していれば、それを認定できるだろうと考えた。そして、th_M が N の最小値 64 の 2 倍である 128 のときの認定もれ（703 件）を分析した。

表 3: 再判定する品詞・活用型群内の $H_{64}(c)$
各ひらがな列のうち、太字は複数の品詞・活用型に共通、
下線はそれのみでスコア ≥ 0.7 になりうる。

	品詞・活用型	特徴ひらがな列（$u-l$ の降順）
C_1	五段・ナ行	ぬ，んだ，に
	五段・バ行	ぶ，び，ばれた，ばれる，ばれ
	五段・マ行	む，み，まれた，まれる
C_2	形容詞	い，く，かった，さを
	五段・ワ行ウ音便	**う**，いの
	五段・ワ行促音便	**う**，った，って，われた，われる われ，っている
C_3	五段・ワ行ウ音便	う，いの
	五段・ワ行促音便	った，い，って，われた，われる われ，っている

表 4: 各 th_M における再現率と精度
T_H は認定すべき品詞・活用型の集合、T_S は認定した品詞・活用型の集合、T_{HS} は正しく認定できた品詞・活用型の集合。

th_M	$\lvert T_S\rvert$	$\lvert T_H\rvert$	$\lvert T_{HS}\rvert$	再現率	精度
1,024	3,607	890	815	91.6%	22.6%
512	3,607	1,496	1,333	89.1%	37.0%
256	3,607	2,325	1,993	85.7%	55.3%
128	3,607	3,413	2,710	79.4%	75.1%
64	3,607	4,606	3,293	71.5%	91.3%
32	3,607	5,167	3,450	66.8%	95.6%
1	3,607	5,590	3,471	62.1%	96.2%

認定もれの大部分を占めたのが、サ変名詞（535 件）と形容動詞（52 件）であった。これらは、そもそも名詞であるが、サ変動詞や形容動詞としても用いられる可能性があるために、一般名詞と区別されているものである。したがって、名詞としての用例しか観察されなければ、認定できなくても当然である。

3.4.2 認定誤りの分析

一度もその用例が出現しなかったにも関わらず、誤って認定された品詞・活用型が 136 件あった。これらの認定誤りの大部分は次の 2 種類のいずれかであった。

辞書未登録語による誤り（96 件）： 本実験では、認定すべき品詞・活用型は形態素解析結果に基づいて決定した。このため、辞書に未登録の語や、品詞・活用型を認定した場合は誤りとみなすことになった。

- 品詞・活用型が未登録（48 件）：認定した品詞・活用型 c とみられる用例は存在するが、その c が形態素解析用辞書には登録されていない。
  例 「返信」（サ変名詞）、「一生懸命」（形容動詞）
- 語幹が未登録（48 件）：形態素解析用辞書に登録されていない語が分かち書きされ、他の語幹として認定されてしまった。
  例 「退部」が「退」と「部」に分かち書きされ、語幹を正しく認定できず、「部」をサ変名詞と認定してしまった。

複数の品詞・活用型を持つ語幹による誤り（26 件）：
複数の品詞・活用型の特徴ひらがな列が混在する場合に、提案手法が誤って本来持たないはずの品詞・活用型を認定してしまう場合があった。

- 名詞の「を」、五段・サ行「した」
  → サ変名詞と誤認定（21 件）
- 形容詞の「く」、五段・ラ行の「った」
  → 五段・カ行促音便と誤認定（2 件）
- 五段・カ行イ音便の「く」、五段・ラ行の「った」
  → 五段・カ行促音便と誤認定（2 件）
- カ行促音便の「く」、五段・ワ行促音便の「い」
  → 形容詞と誤認定（1 件）

4 カタカナ文字列に対する品詞・活用型推定

提案手法が未知の語に対しても既知語と同様の精度で品詞・活用型を推定できるかを確認するために、カタカナ文字列に対する品詞・活用型推定実験を行った。

4.1 データセット

カタカナ文字列およびそれらの最長後続ひらがな列を、河原ら [2] のウェブコーパス[4]から正規表現によって抽出した。抽出した全 6,711,001 種類のカタカナ文字列のうち、50 回以上出現し、かつ 2 字以上の 218,415 種類を品詞・活用型推定の対象とした。

品詞・活用型の推定にあたり、各カタカナ文字列に対して参照する用例数の上限を 3 万件とした。理由は次の通りである。

ノイズの影響を減らす: ウェブコーパスには、書き間違いなどの誤った用例が少なからず存在する。提案手法は、最長後続ひらがな列の出現頻度を用いるため、ノイズの影響を減らすには、判定に用いる用例数を制限する必要がある。

稀な品詞・活用型も認定する: 語幹の頻度に対して 1% しか使用されない品詞・活用型を 85%以上の再現率で獲得するためには、表 4 から、256 件の 100 倍以上の用例を観察する必要がある。

4.2 推定結果と評価

2 つのパラメタ N、th_s の設定を 3 節と同様にしたところ、提案手法によって 9,563 語幹に対する 10,763 件の品詞・活用型が認定された。品詞・活用型ごとの認定数および例を表 5 に示す。

既知語と異なり、カタカナを語幹とする語の大部分に対しては品詞・活用型の正否を判断するための信頼できるリファレンスが存在しない。そこで、検索エンジンを用いて対象のカタカナ文字列の用例を 30 件観察し、以下の 2 つの条件を満たす品詞・活用型のみ正しいと判断した。

- 付与した品詞・活用型の活用語尾が 3 回以上出現
- 基本形の他に少なくとも 1 種類の活用語尾が出現

[4]日本語のウェブページから収集された約 4.6 億文。

表 5: カタカナ文字列の品詞・活用型推定の評価結果
正しい例がなかった品詞・活用型には“×”を記す。

品詞・活用型	認定数	例	評価数	正解数	精度
形容動詞	5,574	マルチな	200	184	92.0%
サ変名詞	4,666	ダビする	200	184	92.0%
五段・ラ行	132	モニョる	132	132	100%
形容詞	128	イナタい	128	125	97.7%
一段	119	イケテる	119	115	96.6%
五段・サ行	104	ゴマカす	104	33	31.7%
五段・カ行イ音便	18	ワメく	18	18	100%
五段・マ行	12	タレコむ	12	12	100%
五段・カ行促音便	5	×	5	0	0%
五段・ワ行促音便	4	ワラう	4	4	100%
五段・ガ行	1	ハシャぐ	1	1	100%

各品詞・活用型ごとに最大 200 件の正否を評価した結果を表 5 に示す。五段・サ行、五段・カ行促音便以外に対しては、90%以上の高い精度を達成した。

五段・カ行促音便の誤りに関しては、3.4 節で述べた複数の品詞･活用型を持つ語幹に対する誤認定であった。

一方、五段・サ行の 71 件の誤りのうち 69 件は、最長後続ひらがな列の抽出誤りが原因であった。我々は、コーパス中の各行は 1 文に対応すると仮定して語幹と最長後続ひらがな列を抽出した。しかし、実際は、「アピールす/る方向で」(“/”は改行）のように文の途中にも関わらず改行されている場合があり、これにより頻度情報にノイズが混入し、品詞・活用型の誤認定につながった。

改行位置が原因で語幹そのものの抽出を誤っている例も 19 件あった。例えば、「ハガキを簡単にダウ/ンロード」(“/”は改行）から「ンロード」を語幹として抽出し、これに対してサ変名詞と推定してしまった。

5 おわりに

本稿では、与えられた語幹の品詞・活用型を推定する手法を提案した。既知語を用いた性能評価を通じて提案手法の性能を確認した。また、カタカナ文字列に対する品詞・活用型推定を通じて、未知語に対しても高い精度を得られることを確認した。

参 考 文 献

[1] 福島健一, 鍛治伸裕, 喜連川優. 機械学習を用いたカタカナ用言の獲得. 言語処理学会第 13 回年次大会発表論文集, pp. 815–818, 2007.

[2] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. 情報処理学会研究報告, NL-171-12, pp. 67–73, 2006.

[3] 桑江常則, 佐藤理史, 藤田篤. 後続ひらがな列に基づく語の活用型推定. 情報処理学会研究報告, NL-186, pp. 7–12, 2008.