# 形態素解析誤りの多い助詞・助動詞の再解析

中村　純平*　　伝　康晴†
*東京農工大学大学院工学府　　†千葉大学文学部

## １．はじめに

我々は，多様な目的に適した形態素解析用電子化辞書 UniDic[1]を開発しており，より柔軟な解析のために解析エンジンに MeCab[2]を用いた実装について検討している[3]．現状で，単語境界認定で 99.5%以上，品詞同定で 99%前後の精度を得ているが，一部の品詞に誤りが集中している．特に，助詞・助動詞間の誤りが多く，品詞同定での誤り数の 30～40%を占めている．したがって，形態素解析の解析精度の向上には，これらの誤りを減少させる必要がある．

品詞同定での誤りを減らす方法として，N-gram を利用することが考えられるが，高次の N-gram を形態素解析全体に用いることは現実的ではない．しかし，コーパスの誤りなどを発見する手法には両方向 *N-gram* を用いて効果を得ているため[4][5]，形態素解析誤りの多い品詞に対し部分的に N-gram を用いることは有効であると考えられる．

本稿では，先行・後続単語からの両方向 *N-gram* を用いて，助詞・助動詞に特化した品詞の判別を行い，形態素解析結果の修正を試みる．助詞・助動詞の誤りのうち，「に」（助動詞「だ」の連用形と格助詞），「で」（助動詞「だ」の連用形と格助詞），「と」（格助詞と接続助詞）に本手法を適用した結果，書き言葉では最大で 54.4%，話し言葉では最大で 38.8%の解析誤りを修正した．

## ２．品詞同定における解析誤り

MeCab における品詞同定の形態素解析誤りの割合を調べた（表 1）．使用したコーパスは，『現代日本語書き言葉均衡コーパス』[6]に含まれる（人手修正済み）白書データ（以下，白書），『RWCP テキストコーパス』[7]（以下，RWCP），『日本語話し言葉コーパス（CSJ）』[8]（以下，CSJ）である．

各コーパスとも，助詞・助動詞間の品詞誤り（例えば，助動詞「だ」の連用形であると解析すべきところを，格助詞であると解析してしまう誤り）が，品詞同定誤りの中の 1/3 程度を占めていることがわかる．その他に多いのは，名詞と接尾辞間の品詞誤りや，連体形と終止形の誤りなどで，この３つで品詞同定の誤り数の約半分を占めていることが分かる．したがって，助詞・助動詞間の誤りを減らすことは非常に有意であると言える．

**表１　品詞同定解析誤りの割合の一例**

| コーパス | 形態素解析誤り | 誤り数 | 割合 |
|---|---|---|---|
| 白書 | 助詞・助動詞間の品詞誤り | 174 | 26.7% |
| | 名詞と接尾辞間の品詞誤り | 103 | 15.8% |
| | 連体形と終止形の誤り | 104 | 16.0% |
| | 記号と名詞間の品詞誤り | 35 | 5.4% |
| | 接続詞と副詞間の品詞誤り | 35 | 5.4% |
| RWCP | 助詞・助動詞間の品詞誤り | 300 | 36.3% |
| | 名詞と接尾辞間の品詞誤り | 73 | 8.8% |
| | 連体形と終止形の誤り | 94 | 11.4% |
| CSJ | 助詞・助動詞間の品詞誤り | 242 | 38.6% |
| | 名詞と接尾辞間の品詞誤り | 41 | 6.5% |
| | 連体形と終止形の誤り | 46 | 7.3% |

各コーパスの助詞・助動詞間の品詞誤りでは，「に」（助動詞「だ」の連用形と格助詞），「で」（助動詞「だ」の連用形と格助詞），「と」（格助詞と接続助詞）に対する誤りが目立った（表 2）．したがって，本稿では，これらの助詞・助動詞を再解析対象とする．

**表２　助詞・助動詞間の解析誤りにおける「に」，「で」，「と」の誤りの割合**

| | 白書 | RWCP | CSJ |
|---|---|---|---|
| に | 67.8% | 27.7% | 18.2% |
| で | 28.7% | 53.3% | 55.0% |
| と | 2.3% | 6.7% | 12.0% |

表 3　MeCab の助詞・助動詞解析の精度

| | | 助詞・助動詞総数 | 正解数 | 誤り数 | 精度 |
|---|---|---|---|---|---|
| に | 白書 | 4353 | 4235 | 118 | 97.3% |
| | RWCP | 2888 | 2805 | 83 | 97.1% |
| | CSJ | 1595 | 1551 | 44 | 97.2% |
| で | 白書 | 1279 | 1229 | 50 | 96.1% |
| | RWCP | 1637 | 1477 | 160 | 90.2% |
| | CSJ | 1015 | 882 | 133 | 86.9% |
| と | 白書 | 1876 | 1872 | 4 | 99.8% |
| | RWCP | 1833 | 1813 | 20 | 98.9% |
| | CSJ | 1468 | 1439 | 29 | 98.0% |
| Total | 白書 | 7508 | 7336 | 172 | 97.7% |
| | RWCP | 6358 | 6095 | 263 | 95.9% |
| | CSJ | 4078 | 3872 | 206 | 94.9% |

なお，MeCab における上記助詞・助動詞の品詞同定の精度は表 3 のようになっている．

## 3．先行・後続単語からの助詞・助動詞の推定

再解析対象の助詞・助動詞の先行・後続単語を調べてみると，品詞ごとにある程度異なる傾向を持つことがわかった．例えば，格助詞「に」は，普通名詞やサ変可能名詞，接尾辞などが先行語となる割合が多い．対して助動詞「だ」の連用形では，形状詞が先行語となることが多い．また，格助詞「で」と助動詞「だ」の連用形は，先行語には共に名詞や接尾辞がくることが多いが，後続語には，格助詞「で」では格助詞，係助詞，名詞が，助動詞「だ」では，非自立可能動詞が多く見られ，明確な違いがあった．

MeCab では,前後の bi-gram を用いてパラメータを学習していが，tri-gram 以上の情報を用いることで，解析誤りの減少を期待できる．

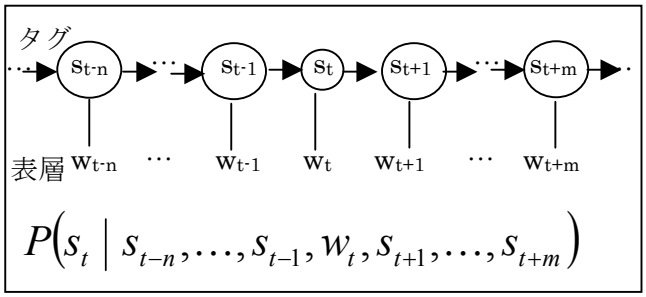


**図 1　両方 *N-gram* による助詞・助動詞の品詞推定**

助詞・助動詞の両方向 *N-gram* を用いて，助詞・助動詞の品詞を推定する手法を考える（図 1）．対象語 $w_t$ に付与する形態素タグ $s_t$ を推定するために，前接する $n$ 語のタグ $s_{t-n}, \ldots, s_{t-1}$ と後続する $m$ 語のタグ $s_{t+1}, \ldots, s_{t+m}$ を利用する．本稿では，$n=1$～3，$m=1$～3 として考える．また，推定に用いる形態素タグとして，品詞，活用型，活用形を利用し，品詞体系は UniDic に準拠する．ただし，対象語に隣接する $s_{t-1}$ と $s_{t+1}$ では，語彙素（辞書的見出し）も利用する．例えば，「昨日公園に行った。」という文があったとき，「に」の品詞（が助動「だ」であるか，格助「に」であるか）に対する両方向 *N-gram* の確率を求めることになる（図 2）．

$P(s_t$[助動詞ダ] $\mid$ $s_{t-2}$[名詞,普通名詞,副詞可能],
$s_{t-1}$[名詞,普通名詞,一般,公園],
$w_t$[に],
$s_{t+1}$[動詞,非自立可能,五段 - カ行 - イク,
連用形 - 促音便,行く]$)$

**図 2　「昨日公園に行った。」に対する「に」が助動詞「だ」である両方向 *N-gram* の例（n=2,m=1）**

## 4．助詞・助動詞品詞推定の両方向 *N-gram* の確率値の学習

助詞・助動詞品詞推定の両方向 *N-gram* の確率値を，CRF[9]を用いて学習した．CRF は，条件付確率をそのまま近似することができるため有効であると思われる．また，本稿の手法では，「に」や「で」など助詞・助動詞の表層が分かっている場合を対象としており，2 値分類の問題と捉えることができるため，2 値分類に強い SVM[10]を用いて助詞・助動詞の品詞推定を行い，CRF と比較する．

学習及び評価用のコーパスには，MeCab の形態素解析辞書作成用に作られた各コーパスの学習用・評価用データ[3]を利用する．MeCab 用学習・評価データから，修正対象の助詞・助動詞と，その前後 3 単語及び形態素タグを 1 事例として抜き出し，両方向 *N-gram* 用の学習・評価データとする．学習・評価データにおける各品詞の事例数は表 4 のようになる．

**表 4　学習・評価データにおける各品詞の事例数**

| | | 事例数 | | | |
|---|---|---|---|---|---|
| | | 学習用 | 評価用 | | |
| 表層 | 品詞 | 全体 | 白書 | RWCP | CSJ |
| に | 助動詞 | 2817 | 444 | 125 | 327 |
| | 格助詞 | 31109 | 3937 | 2793 | 1284 |
| で | 助動詞 | 4664 | 557 | 420 | 440 |
| | 格助詞 | 13219 | 727 | 1246 | 597 |
| と | 格助詞 | 16152 | 1789 | 1707 | 1336 |
| | 接続助詞 | 1450 | 99 | 141 | 153 |
| 計 | | 69411 | 7553 | 6432 | 4137 |

### 4.1　CRF による助詞・助動詞の品詞推定

CRF++[1]を用いて，先行・後続単語から助詞・助動詞への条件付き確率を学習した．結果の一部を表 5 に示す．白書では $n = 1$，$m = 3$ のとき，RWCP では $n = m = 2$ のとき，CSJ では $n = 2$，$m = 3$ のときに精度の最高値を得た．また，$n = m = 2$ のときが，どのコーパスに対しても高い精度を得，MeCab の解析結果より高かった．

**表 5　CRF による助詞・助動詞の品詞推定結果**

| 先行単語数 | 後続単語数 | コーパス | 正解数 | 誤り数 | 精度 |
|---|---|---|---|---|---|
| 1 | 1 | 白書 | 7415 | 138 | 98.2% |
| | | RWCP | 6269 | 163 | 97.5% |
| | | CSJ | 3973 | 164 | 96.0% |
| 1 | 3 | 白書 | 7430 | 123 | 98.4% |
| | | RWCP | 6288 | 144 | 97.8% |
| | | CSJ | 3995 | 142 | 96.6% |
| 2 | 2 | 白書 | 7427 | 126 | 98.3% |
| | | RWCP | 6307 | 125 | 98.1% |
| | | CSJ | 4007 | 130 | 96.9% |
| 2 | 3 | 白書 | 7414 | 139 | 98.2% |
| | | RWCP | 6295 | 137 | 97.9% |
| | | CSJ | 4008 | 129 | 96.9% |

### 4.2　SVM による助詞・助動詞の品詞推定

YamCha[11]を用いて SVM による助詞・助動詞の品詞同定を行った．SVM のモデルは，それぞれの表層語（「に」,「で」,「と」）別に作成した．結果の一部を表 6 に示す. 白書では $n = 1, m = 2$, RWCP では $n = m = 2$，CSJ では $n = 2$，$m = 3$ のときが精度が高かった．CRF と同様に，$n = m = 2$ のときがどのコーパスに対しても高い精度を得ており，CRF の精度を上回った.

**表 6　SVM による助詞・助動詞の品詞推定結果**

| 先行単語数 | 後続単語数 | コーパス | 正解数 | 誤り数 | 精度 |
|---|---|---|---|---|---|
| 1 | 1 | 白書 | 7437 | 116 | 98.5% |
| | | RWCP | 6299 | 133 | 97.9% |
| | | CSJ | 3962 | 175 | 95.8% |
| 1 | 2 | 白書 | 7456 | 97 | 98.7% |
| | | RWCP | 6321 | 111 | 98.3% |
| | | CSJ | 4030 | 107 | 97.4% |
| 2 | 2 | 白書 | 7442 | 111 | 98.5% |
| | | RWCP | 6337 | 95 | 98.5% |
| | | CSJ | 4042 | 95 | 97.7% |
| 2 | 3 | 白書 | 7435 | 118 | 98.4% |
| | | RWCP | 6323 | 109 | 98.3% |
| | | CSJ | 4045 | 92 | 97.8% |

## 5．助詞・助動詞の再解析

作成した CRF，SVM のモデルを用いて，MeCab による形態素解析結果から助詞・助動詞の部分を再解析し，解析誤りがどの程度減少するかを調べた. MeCab で解析した MeCab 用評価データ[3]から同様に助詞・助動詞とその周辺単語を抜き出し，比較した．助詞・助動詞の前後の単語に単語境界誤りがあるものに関しては扱わない（各評価用データの事例数から境界誤りを引いた数は，表 8 の総数の欄を参照）.

CRF による助詞・助動詞再解析結果の一部を表 7 に，SVM による助詞・助動詞再解析結果の一部を表 8 に示す.

表 7 と，表 8 では，それぞれのコーパスにおける正修正数が最大のものが示されている．したがって, 白書では, SVM で n = 1, m = 2 のとき (+50) が最高で，白書における助詞・助動詞解析誤り（172 個）の内 29.1%を修正できたことになる．RWCP では，CRF で n = 2, m = 2 のとき (+143) が最高で，助詞・助動詞解析誤り（263 個）の内 54.4%を修正できた．CSJ では，SVM で n = 2, m = 3 のとき（+80）が最高で，助詞・助動詞解析誤り（206 個）の内 38.8%を修正できた．$n = m = 2$ の場合においては，CRF の方が多く修正できていた.

なお，表中の誤適用数とは，MeCab の出力結

[1] http://crfpp.sourceforge.net/

果では正解であったものを修正した結果，不正解になった数を表わしている．誤適用数に関しては，CRF の方が少なく，特に n = m = 2 のときは，SVM の半分程度であった．

**表 7　CRF による助詞・助動詞修正結果**

| 先行数 | 後続数 | コーパス | MeCab正解数 | 再解析正解数 | 正解数差 | 誤適用数 |
|---|---|---|---|---|---|---|
| 1 | 3 | 白書 | 7337 | 7386 | +49 | 45 |
| | | RWCP | 6096 | 6222 | +126 | 41 |
| | | CSJ | 3872 | 3932 | +60 | 52 |
| 2 | 2 | 白書 | 7337 | 7384 | +47 | 48 |
| | | RWCP | 6096 | 6239 | +143 | 34 |
| | | CSJ | 3872 | 3948 | +76 | 38 |
| 2 | 3 | 白書 | 7337 | 7369 | +32 | 51 |
| | | RWCP | 6096 | 6228 | +132 | 36 |
| | | CSJ | 3872 | 3949 | +77 | 41 |

**表 8　SVM による助詞・助動詞修正結果**

| 先行数 | 後続数 | コーパス | 総数 | MeCab正解数 | 再解析正解数 | 正解数差 | 誤適用数 |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 白書 | 7508 | 7336 | 7386 | +50 | 59 |
| | | RWCP | 6358 | 6095 | 6209 | +114 | 66 |
| | | CSJ | 4078 | 3872 | 3928 | +56 | 74 |
| 2 | 2 | 白書 | 7508 | 7336 | 7370 | +34 | 67 |
| | | RWCP | 6358 | 6095 | 6220 | +125 | 61 |
| | | CSJ | 4078 | 3872 | 3946 | +74 | 62 |
| 2 | 3 | 白書 | 7508 | 7336 | 7360 | +24 | 68 |
| | | RWCP | 6358 | 6095 | 6206 | +111 | 68 |
| | | CSJ | 4078 | 3872 | 3952 | +80 | 62 |

## 6．考察

それぞれの結果から，CRF や SVM による先行・後続単語からの助詞・助動詞の品詞推定は有効であることがわかる．特に，先行・後続単語ともに 2 つずつ使うと，どのコーパスも高い精度で助詞・助動詞の品詞を推定できた．

表 5･6 を見ると，SVM の方が高精度であるが，形態素解析結果に対し助詞・助動詞を再解析した表 7･8 では，CRF が総合的に良い結果になった．また，表 7・8 からわかるように，誤適用が非常に多く見られた．これらの理由として，助詞・助動詞の周辺単語の形態素解析結果に誤りが含まれていることが考えられる．しかし、形態素解析に誤りがある個数は誤適用数の 1/6 程度しかなかった．そのため，形態素解析が間違う箇所と助詞・助動詞の推定が間違える箇所が異なるのだと考えられる．お互いに得手不得手な部分が分かれば，誤適用を避け，適切な修正も可能になる．

## 7．おわりに

形態素解析誤りの多い助詞・助動詞に対して，両方向 *N-gram* を用いて品詞推定を行い，形態素解析された結果を再解析した．その結果，白書では最大で 29.1%，RWCP では最大で 54.4%，CSJ では最大で 38.8%の解析誤りを修正した．

## 参考文献


[1] 伝康晴，小木曽智信，小椋秀樹，山田篤，峯松信明，内元清貴，小磯花絵．(2007)．コーパス日本語学のための言語資源: 形態素解析用電子辞化辞書の開発とその応用．日本語科学, *22*, 101-123.

[2] Kudo, T., Yamamoto, K., & Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 230–237). Barcelona, Spain (pp.148-153).

[3] 伝康晴，中村純平，小木曽智信，小椋秀樹．(2008)．語種情報を用いた同表記異音語の解消．言語処理学会第 14 回年次大会発表論文集.

[4] Eleazar Eskin. (2000). Detecting errors within a corpus using anomaly detection. In *Proceedings of the 1st Meeting of the NAACL.*

[5] 河田岳大，工藤峰一，外山淳，中村篤祥．(2005)．両方向 N-gram 確率を用いた誤り文字検出，自然言語処理，vol.J88-D2, No.3 (pp.629-635)

[6] 山崎誠，前川喜久雄，田中牧郎，小椋秀樹，柏野和佳子，小磯花絵，間淵洋子，丸山岳彦，山口昌也，秋元祐哉，稲益佐知子，吉田谷幸宏．(2006)．代表性を有する現代日本語書き言葉コーパスの設計．言語処理学会第 12 回年次大会発表論文集 (pp.440-443)．東京.

[7] 新情報処理開発機構（RWCP）テキスト・サブ・ワーキンググループ．(1998)．研究開発用知的資源: タグ付きコーパス報告書.

[8] 前川喜久雄．(2004)．『日本語話し言葉コーパス』の概要．日本語科学, *15*, 111-133.

[9] Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282–289). Williamstown, MA.

[10] V.N.Vapnik. (1995). The Nature of Statistical Learning Theory, *Springer*.

[11] 工藤 拓，松本 裕治．(2002)．Support Vector Machine を用いた Chunk 同定．*自然言語処理, Vol.9, No, 5*（pp 3-22）