Vol.2009-CH-84 No.3
2009/10/24

MeCab を用いた古典中国語形態素解析器の改良

守　岡　知　彦[†1]

CH 79 で発表した MeCab を用いた古典中国語形態素解析器のその後の改良について述べる。ここでは、特に、品詞（素性）階層の設計の問題と異体字処理の問題に関して焦点を当てる。

Improvement of Morphological Analyzer for Classical Chinese based on MeCab

MORIOKA Tomohiko[†1]

This paper explains improvement techniques of an experimental Morphological Analyzer for Classical Chinese based on MeCab. Especially it focuses hierarchy of parts of speech (features) and handling of character variants.

1. はじめに

MeCab[1] を用いた古典中国語形態素解析器の改良の試みについて述べる。CH 79 で発表したように[2]、IPA 辞書をベースにヒューリスティックスで古典中国語的な語彙を抽出することで、古典中国語のための辞書や形態素コーパス[⋆1]（以下、単に「コーパス」と呼ぶ）が十分に存在しない状況において、極めて少ない労力である程度の実用性を有する古典中国語形態素解析器のプロトタイプを実現することができたが、それでもさまざまな古典中国語の文献を扱う上ではまだまだ不十分な点が多い。そもそもこの手法はちゃんとした辞書やコーパスが存在しない段階での代替物なのであり、ちゃんとした辞書やコーパスを作るための作業の土台を提供するためのものなのである。[⋆2]

しかしながら、実際にコーパスを蓄積するためにはその形式を十分に検討する必要があり、具体的には品詞体系をどうするかが問題となる。妥当な品詞体系を検討するためには多数のコーパスから検討を重ねる必要があるが、コーパスを蓄積するためには品詞体系が必要である、という鶏が先か卵が先か的な問題があるが、まずは古典中国語の MeCab 用品詞体系の定性的な性質や傾向性を見出し、暫定的な品詞体系を決める必要がある。実際には品詞体系は漸進的に改良していく他ないといえ、そのことを鑑みれば、コーパスの品詞体系をメンテナンスするための手法も必要である。

また、実際にさまざまなテキストを形態素解析し、その結果を基にコーパスを作っていくためには、異体字の問題に対処する必要があるといえる。異体字の問題は文字処理の分野で古くからさまざまに扱われてきた問題であるが、ここでは文字レイヤーではなく語彙レイヤーの問題として扱う必要があるといえる。漢字の異体字関係は時代や地域、分野、テキスト、文脈等に依存するということが知られているが、ここではこうした問題を語彙の表記の揺れという観点から陽に扱う必要があるといえる。しかしながら、この問題に関しても、文字単位のデータはある程度蓄積されていても、形態素・語彙という視点で分節化されたテキストやコーパスはあまり蓄積されてないといえ、これは伝統的な文字・テキスト・自然言語処理等が同様な問題を扱うのにもレイヤーで分断されてきたという問題のひとつの例であるといえる。理想的には文字処理と自然言語処理が連係するのが望ましく、そのためには、文字オントロジーと自然言語処理のための辞書やコーパス、その他データベースを連係して運用することが望ましいといえるが、現状では、必ずしも簡単なことではない。しかしながら、そうした方向に向かうための指針を見出す必要がある。

ここではこうした観点に基づいて行った試みとそこで得られた知見について報告するとともに、今後の改良のための方向性について議論する。

2. 品詞（素性）階層

MeCab では 4 階層の品詞（素性）階層を使って bigram を用いたパラメータ推定が行えるが、ここでどのような品詞（素性）階層を用いるかが問題となる。

そこで、品詞（素性）階層を評価するために、それぞれの体系基づき、コーパスと辞書を記述し、各コーパスを互いに学習用コーパスとテストデータとして認識実験を行い、全組合

†1 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University

⋆1 ここでは MeCab の学習用コーパスを指すものとする。

⋆2 コーパスがあれば辞書はできるので、作業として必要なのはコーパスを作ることである。

Vol.2009-CH-84 No.3
2009/10/24

せの認識精度を計算した。

この実験には、M（69 語）、K（68 語）、R（320 語）、W（248 語）、J（363 語）というコーパスを用いた。なお、M は雑多な文例、K は典型的な構文例、R は三国志呉書列伝よりの抜粋、W と J は旧唐書東夷伝の項目であり、地理的な記述が多く、特に J には外来語が多い。

2.1 大 品 詞

学習データ ＼ 入力	M	K	R	W	J
M（大品詞/品詞）	100	90/82	90/88	92/89	94/91
M（大品詞なし）	100	87	90	91	93
K（大品詞/品詞）	91/90	100	92/89	94/90	95/93
K（大品詞なし）	90	100	95	88	91
R（大品詞/品詞）	100/99	90/85	100	94/91	97/94
R（大品詞なし）	94	87	100	90	92
W（大品詞/品詞）	100/100	93/90	92/90	100	98/97
W（大品詞なし）	96	90	94	100	93
J（大品詞/品詞）	97/94	94/88	93/91	97/95	100
J（大品詞なし）	91	88	93	94	100

表 1 大品詞がある場合とない場合の大品詞と品詞の F 値（意味素性あり）

学習データ ＼ 入力	M	K	R	W	J
M（大品詞/品詞）	100	89/84	87/84	90/83	88/85
M（大品詞なし）	100	83	85	83	85
K（大品詞/品詞）	85/81	100	84/80	88/86	84/82
K（大品詞なし）	81	100	80	84	77
R（大品詞/品詞）	96/92	84/76	99/99	91/85	89/84
R（大品詞なし）	87	76	100	82	82
W（大品詞/品詞）	94/93	87/79	87/84	99/99	94/93
W（大品詞なし）	91	78	84	100	93
J（大品詞/品詞）	93/90	87/79	93/90	92/88	99/99
J（大品詞なし）	91	76	87	87	100

表 2 大品詞がある場合とない場合の大品詞と品詞の F 値（意味素性なし）

大品詞を設けない場合の品詞の認識精度は、意味素性（サブ品詞）を付ける場合、概ね、大品詞を設けた場合の大品詞の認識精度と品詞の認識精度の中間ぐらいの値を取るといえるが、学習用コーパスと入力データの文体の差が大きい場合、大品詞を設けた方がロバスト性が高まると考えられる。

一方、意味素性を付けない場合、大品詞を設けない場合の品詞の認識精度は、大品詞を設けた場合に比べて下がりやすいといえる。このことからも、大品詞を設けた場合のロバスト性の向上効果がうかがえる。

よって、意味素性を付与したコーパス・辞書を用意でき、適切な学習用コーパスを用意できる場合においては、大品詞は必ずしも必要ないといえるが、開発途上の段階では大品詞を設けた方が安全であるかもしれない。

2.2 意 味 素 性

学習データ ＼ 入力	M	K	R	W	J
M	100	89/84	87/84	90/83	88/85
K	85/81	100	84/80	88/86	84/82
R	96/92	84/76	99/99	91/85	89/84
W	94/93	87/79	87/84	99/99	94/93
J	93/90	87/79	93/90	92/88	99/99

表 3 名詞・動詞ともに意味素性がない場合の大品詞と品詞の F 値

学習データ ＼ 入力	M	K	R	W	J
M	100	88/79	87/83	88/83	88/83
K	88/82	100	87/83	90/86	89/87
R	95/89	90/82	99/99	93/89	93/89
W	97/97	93/87	90/86	100	95/92
J	94/86	96/91	94/90	95/90	100

表 4 動詞の意味素性がない場合の大品詞と品詞の F 値

コーパスに意味素性を付与した方が付与しない場合よりも認識精度は向上する。また、部分的に意味素性を付与する場合、動詞に意味素性を付与し名詞の意味素性を省略する方が、名詞に意味素性を付与し動詞の意味素性を省略するよりも効果的であるといえる。

2.3 動詞の品詞体系の問題

動詞の分類としては形態論・統語論的なものと意味論的なものが考えられるが、MeCab は隣接する形態素の素性の bigram を学習に用いているので、そこから導出できるような情報を動詞の品詞（素性）として付けても冗長であり、認識精度はあまり向上しないといえ

Vol.2009-CH-84 No.3
2009/10/24

学習データ ＼ 入力	M	K	R	W	J
M	100	90/84	92/90	92/88	93/91
K	88/86	100	88/86	90/88	93/91
R	96/95	87/81	100	92/87	95/90
W	96/94	87/79	92/89	100	97/96
J	94/94	88/84	96/93	93/92	100

表 5 名詞の意味素性がない場合の大品詞と品詞の F 値

学習データ ＼ 入力	M	K	R	W	J
M	100	90/82	90/88	92/89	94/91
K	91/90	100	92/89	94/90	95/93
R	100/99	90/85	100	94/91	97/94
W	100/100	93/90	92/90	100	98/97
J	97/94	94/88	93/91	97/95	100

表 6 名詞・動詞ともに意味素性がある場合の大品詞と品詞の F 値

る。また、複雑な係り受け構造に関わるような情報は形態素解析のレイヤーでは扱えないので、統語論的な情報もまたあまり有益ではないと考えられる。古典中国語の動詞が屈折しないということも鑑みれば、意味論的な情報を中心に２階層の素性を用いて[⋆1]動詞を分類するのが良いといえる。

動詞の分類としては、アスペクト（相）、意志・無意志、視点などがあるが、古典中国語の動詞には時制やアスペクトによって変化せず（形態論的に表現されず）、意志・無意志の両方をとることがあり得、能動態・受動態の対立もない。特定のアスペクトや意志・無意志、視点をとるような動詞を見つけることができればそうした観点での分類・素性付けを行う価値があるといえるが、一般にはあまり適切ではなさそうである。同様の理由から概念依存性に基づく分類もまたここでの目的には向いてないと考えられる。

一方、政治・司法・国家制度、軍事、雇用・役職、農業、商業・売買・契約、家族関係、人間関係、文字・書物といった対象領域の分野による分類はある程度有用であると考えられる。しかしながら、この種の分野による分類を用いた場合、コーパスの分野依存性が高くなると考えられる。

現在のところ、アドホックに付けた品詞体系の方がある程度理論的・システマティックに付けたものより良い成績を収めており、適切な品詞体系を設計するためにはコーパスをもっと蓄積して、その係り受け関係や意味論的な分析を含めて検討を行い、現実的な品詞体系の設計を行う必要があるといえる。

2.4 品詞体系のメンテナンス

コーパスに基づいて適切な品詞体系を設計できない段階においては、大まかな指針に基づきつつも、実際の文例に基づきコーパスを作る段階でアドホックに品詞・素性を付けるしかないといえる。この場合、問題となるのは品詞・素性の揺れである。例えば、同じ形態素に対して誤って違う品詞・素性を付けてしまうことがある。この問題の解決には、コーパスから生成した辞書[⋆2] が役立つと考えられる。もしコーパスから生成した辞書の同じ正規形に対して異なる品詞・素性を持つエントリーが存在する場合、品詞・素性の揺れが存在する可能性があるので、それをチェックすれば良い訳である。一方、異なる語彙間の品詞・素性の揺れをチェックするには、辞書・コーパスで用いられている品詞・素性のリストが有用である。しかしながら、品詞・素性のリストだけで品詞体系を検討すると個別の文脈の問題が捨象されてしまいがちであり問題がある。よって、品詞・素性のリストとその語彙の例、そして、その語彙が置かれたコーパス中の文を関連付けるようなツールが有用であると考えられる。

品詞体系を変更した場合、それに従って辞書やコーパスの品詞・素性を修正する必要がある。IPA 辞書から変換した辞書の場合、変換スクリプトを修正し、再度辞書を変換することで容易に品詞・素性を変換可能であるが、人手で作った辞書やコーパスに関しては工夫が必要である。この問題を軽減するためには、品詞体系を変更した時に辞書やコーパスを一貫して変更するようなツールを作るのが良いかも知れない。

また、別のアプローチとしては、形態素解析器の品詞体系に対して独立した品詞体系、いわば情報の蓄積や交換のための標準的な品詞体系（ここでは、交換用品詞体系）を設け、コーパスや辞書はその品詞体系で書くというものである。形態素解析器の品詞体系に対しては交換用品詞体系から変換することにする訳である。交換用品詞体系が形態素解析器の品詞体系に対して十分に表現力があり、一意に変換可能であればこれは容易である。但し、実際には形態素解析に用いない品詞・素性までコーパスに書かなければならず、また、形態素解析の結果をそのままコーパスに用いることができないという点でこのアプローチには問題がある。

⋆1 MeCab の制約による。

⋆2 misc.corpus.csv

一方、MeCab の部分解析（制約付き解析）機能を利用するアプローチも考えられる。これはコーパスの内、変更され得る部分をワイルドカードにして、辞書の情報から補完するという方法である。この場合、元のコーパスの情報を落すだけであるから実現は簡単であるが、コーパスに記述していた情報の幾つかを落すことになるので、詳細な素性を付けたコーパスを蓄積している場合、問題である。

3. IPA 辞書からの変換スクリプトの改良

3.1 取り込む情報の拡大

CH 79[2] で発表したもの（第 0.9 版）では、動詞として基本形かつ語尾が「しめる」「しむ」「る」「う」「く」「ぐ」「す」「ず」「つ」「づ」「ぬ」「ふ」「ぶ」「む」「ゆ」「る」のものを取り込んでいたが、今回報告するもの（第 0.796 版）ではこれらに加えて語尾が「かす」「がす」「さす」「ざす」「たす」「だす」「なす」「はす」「ばす」「ます」「やす」「らす」「わす」「ある」「かる」「がる」「さる」「ざる」「たる」「だる」「なる」「はる」「ばる」「まる」「やる」「らる」「わる」「うる」「くる」「ぐる」「する」「ずる」「つる」「づる」「ぬる」「ふる」「ぶる」「むる」「ゆる」「える」「ける」「げる」「せる」「ぜる」「てる」「でる」「ねる」「へる」「べる」「める」「れる」「ゑる」のものも取り込むようにした。[⋆1]これにより、動詞 (Verb.i.csv) の実エントリー数[⋆2]は 1683 から 3357 に増加した。形容詞に関しても、従来では文語基本形でかつ語尾が「なし」「ない」「しい」「し」「い」のものをとっていたが、今回のものではこれらに加えて語尾が「ばし」のものも取り込むようにした。これにより、形容詞の実エントリー数は 221 から 226 に増加している。

また、サ変名詞に関して、従来、1 文字のもののみを取り込んでいたが、今回のものでは２文字のものも取り込むようにした。これにより、サ変名詞由来の動詞 (Verb.s.csv) の実エントリー数は 129 から 7945 に増加している。この結果、複合動詞がうまく解析できるようになった半面、動詞+名詞のパターンが 1 形態素として解析されてしまう可能性があるが、その場合も動詞句を構成することには違いない訳で、現時点では実用上は弊害よりも利点が上回ると判断した。

3.2 記　　号

記号をサポートするようにした。これにより、句読点、括弧、下駄（〓）等の記号が混ざったテキストを認識する際の利便性が向上した。

3.3 除外リスト

取り込みたくないものを記述する除外リストをどう管理するかという問題がある。現状では、除外リストは変換スクリプトに埋め込んだ形で管理しているが、複数人で作業することを考えれば、何らかのデータベース化が望ましいといえる。

4. 異体字処理

CH 79 で発表したように[2]、現在のところ、名詞・動詞を中心とした辞書中の語彙の多くは IPA コーパスから作成された「IPA 辞書」から機械的に抽出したものである。IPA コーパスは現代日本語のコーパスであるから、そこでの漢字表記は常用漢字をはじめとする現代日本略字を中心とするものである。一方、古典中国語のテキストはいわゆる『康熙体』を中心とする繁体字や『本字』と呼ばれるような『康熙体』以前の規範に基づくような字体が用いられていることが多く、UCS ではこれらの字体に対して異なる符号位置が与えられていることが少なくないので、「IPA 辞書」から変換したままの辞書では十分に解析できないといえる。また、常用漢字でも『康熙体』でも『本字』でも同じ文字として学習でき、どの字体で表記したコーパスから学習しても別の字体での表記が十分に解析できることが望ましい。こうしたことから、何らかの異体字処理のための工夫をする必要があるといえる。

MeCab では日本語の活用による語尾変化に対応するために、「表層形」と「原形」を対応づけることが可能であり、この際、どの表層形でも原形として学習することができる。この仕組みはハードコーディングされているのではなく、辞書形式の定義（rewrite.def, feature.def 等）によって日本語以外の言語でも利用可能である。ここではこの仕組みを使って、表層形として各種異体字表記を、原形としていわゆる『康熙体』を用いることにする。但し、以下では表層形を「出現形」、原形を「正規形」と呼ぶことにする。

IPA 辞書から変換した辞書の出現形を正規形に変換するために、XEmacs CHISE[3] で動作する Emacs Lisp プログラムを用いた。この変換プログラムでは CHISE 文字オントロジー[4)5)] を用いて、出現形の常用漢字（およびその他現代日本略字）を正規形に変換する他、（生成した）正規形を異体字に変換して生成した出現形のエントリーを付加する機能を持つ。

常用漢字といわゆる『康熙体』は 1 対 n 対応であり、また、語彙固有の表記もあり、機械的に一意に変換することはできないが、ここでは取り得る全組合せを生成することで対処している。この結果、有り得ない正規形・出現形ができてしまうが、出現形に関しては実際にテキストやコーパスに存在しないものは出現頻度が 0（ないしは微小）になり、無駄なだ

⋆1 このうちの幾つかは第 0.9 版のものでも取り込むことを意図していたが、bug のために変換できていなかった。
⋆2 エントリー数の内、生成した異体字出現形を除いたもの。

けで大きな悪影響はないと考えられる。正規形に関してはテキストやコーパスが常用漢字で書かれていた場合に正しく正規形が決定できず、悪影響が考えられるが、実際には『康熙体』や『本字』を中心とした繁体字で書かれたテキストを扱う場合がほとんどであるので、実用上問題がないといえる。

異体字の出現形に関しては、現在の所、試験的にいわゆる『本字』を対象としているが、いわゆる『同字』、『別体』、『古字』などの他のカテゴリーや音通のようなものも扱うべきかどうかは問題である。無駄な出現形エントリーがあっても良いということからすれば単純になるべく多くの異体字・類字をサポートすべきであるが、あまりに多くのものをサポートした結果、複数の語彙が衝突すると認識精度の低下を招くといえ、この点で検討が必要であるといえる。また、そもそも異体字・類字の出典やそこでの文脈を無視して、単純に文字間の関係として扱ったデータから語彙レイヤーの情報を生成しようとすることには問題があり、むしろ、実際のテキストを基に作成したコーパスから異体字・類字関係に関する文字オントロジーを構成するようなアプローチの方が望ましいといえる。異体字の出現形の生成に関しては、コーパス作成作業の手間を下げるためのものと考えた方が良いといえる。

5. おわりに

MeCab[1] と IPA 辞書をベースにした古典中国語形態素解析器の改良の試みについて述べた。CH 79 で報告したように[2]、IPA 辞書をベースにヒューリスティックスで古典中国語的な語彙を抽出することで、極めて少ない労力で古典中国語形態素解析器のプロトタイプを実現することができたが、本格的な形態素解析器を実現するためには実際にさまざまなテキストを対象に形態素解析して、大規模なコーパスを蓄積していくことが必要となる。その作業にはやはり大きな労力がかかり、また、いくつもの試行錯誤を重ねる必要があると思われるが、ここではその労力をなるべく削減するための方法について議論した。

品詞体系に関しては、十分に辞書やコーパスが整備できていない段階、あるいは、適切な分類・整理が行えてない段階では、情報を増やしたがためにかえって認識精度を下げてしまうという問題があるが、CH 79 で提案した『大品詞』はロバスト性を高める上で有効な方策であると考えられる。細分類品詞に関しては形態論・統語論的なものよりも意味的なもの（意味素性）の方が有効であるといえるが、実際にどういうものを付けるべきかということを判断するにはコーパスの蓄積が必要であるといえる。特に、動詞の分類は重要であり、多数の文例を集めた検討が待たれる。

現実のさまざまなテキストを対象にする場合、異体字の問題に対処する必要があるが、既存の異体字データベース等を利用してある程度対処は可能である。但し、ここでは文字レイヤーではなく形態素・語彙レイヤーにおける表記の問題として扱う必要があり、文脈等を無視して単純に異体字・類字関係として整理した既存の一般的な異体字データベースでは問題があると考えられる。文字間の関係として記述したものでも、その関係の種類が書かれたようなものの方がこの用途には適しているといえる。また、複数の語彙間で異体字・類字が衝突して認識精度が下がらないように考慮しなければならない。より抜本的には形態素解析のためのコーパスと文字オントロジーを連係させるようなシステムを実現することが望ましいといえる。

最後に、本研究を行う上で、京都大学人文科学研究所「東アジア古典文献コーパスの研究」共同研究班のメンバー諸氏、特に、山崎直樹氏からさまざまな示唆を受けたことに感謝する。しかしながら、本論文での誤りは著者の責に依るものであることはいうまでもない。

参考文献

1) : MeCab: Yet Another Part-of-Speech and Morphological Analyzer, `http://mecab.sourceforge.net/`.
2) 守岡知彦：MeCab を用いた古典中国語の形態素解析の試み，情処研報， Vol.2008, No.73, pp.17–22 (2008). 2008-CH-79.
3) : XEmacs CHISE, http://kanji.zinbun.kyoto-u.ac.jp/projects/chise/xemacs/.
4) 守岡知彦：文字オントロジーに基づく文字処理について，情処研報， Vol.2006, No.112, pp.25–32 (2006). 2006-CH-72.
5) Tomohiko, M.: CHISE: Character Processing based on Character Ontology, *Large-scale Knowledge Resources (LKR2008)*, LNAI, No.4938 (2008).