古典中国語形態素解析による地名の自動抽出

安岡 孝一　守岡 知彦　ウィッテルン クリスティアン
京都大学人文科学研究所附属東アジア人文情報学研究センター

山崎 直樹　二階堂 善弘　鈴木 慎吾
関西大学外国語学部　関西大学文学部　大阪大学言語文化研究科

MeCab を用いた古典中国語の形態素解析に際し，地名に特化した自動抽出法を提案する．具体的には，形態素解析に用いる古典中国語辞書に，地名を大量に追加する手法を提案する．ただし，そのような手法においては,「1 文字の地名」と同一の漢字が別の意味をも持ちうる場合，それら対抗用例に誤検出が生じる可能性がある．この問題に対し，我々は，地名用例を含む古典中国語形態素コーパスと，その対抗用例コーパスの両方を準備することで，誤検出を低く抑えることに成功した．合わせて，古典中国語辞書に含まれる地名の数と，F 値との関係を調べ，本稿の手法の定量的評価を試みた．

Extraction of Place Names from Classical Chinese Texts Using Morphological Analysis

Koichi Yasuoka　Tomohiko Morioka　Christian Wittern
Center for Informatics in East Asian Studies, Institute for Research in Humanities
Kyoto University

Naoki Yamazaki
Faculty of Foreign Language Studies
Kansai University

Yoshihiro Nikaido
Faculty of Letters
Kansai University

Shingo Suzuki
Graduate School of Language and Culture
Osaka University

In this paper we propose a method to extract place names from classical Chinese texts. In the method, we use our original morphological analyzer based on MeCab with our digital dictionary, in which we especially added many place names. A place name with one character can be easily mistaken for another morpheme. We include many example sentences with place names in our digital corpus, and also include many counter-examples in the corpus, in order to reduce incorrect detection. Additionally, we evaluate the method quantitatively using F-measure, changing the number of place names in the dictionary.

1 はじめに

これまでに我々は，MeCab [1] を用いた古典中国語の形態素解析について，その実際的手法と実用性とを研究し，一定の成果を上げてきた [2–8]. ただ，古典中国語に対する現実的な検索を考えた場合，どうしても，地名，人名，官職など，固有表現の検索が不可欠となり，それらに対する自動抽出手法を考える必要がある．これら固有表現に対しては，我々は当初，形態素解析を超える手法を導入しなければならないのではないか，と予想していたが，あにはからんや地名に関してだけは，形態素解析を「強引」におこなうことで，どうやら自動抽出できそうな目途が立った．あるいは古典中国語だけの特殊事情である可能性もあるのだが，その点も含めて御叱正をいただくべく，ここに報告する．

2 古典中国語形態素解析の概要

我々の古典中国語形態素解析は，3 つの要素技術で構成される．形態素解析に特化した古典中国語の品詞体系 [4,5]，その品詞体系にもとづく古典中国語辞書，および，その品詞体系に基づく古典中国語形態素コーパスである．

形態素解析に特化した古典中国語の品詞体系は，MeCab の 4 階層の品詞体系に合わせており，上位層から順に「大品詞」「品詞」「意味素性」「小素性」と呼んでいる．大品詞は「n」「v」「p」の 3 種類であり，「v」と「n」が，古典中国語の動賓構造の「動」と「賓」に対応している．品詞は「名詞」「代名詞」「数詞」「動詞」「前置詞」「副詞」「助動詞」「助詞」「感嘆詞」の 9 種類であり，従来の漢文文法等で見られた「形容詞」を廃止している．意味素性は 43 種類，小素性は 83 種類を定義しており，形態素解析の結果として得られる各単語を，意味の面からも捉えやすいよう工夫している (図 1).

古典中国語辞書は，IPA 日本語辞書 [1] を基に作成した古典中国語辞書 [2,3] に対し，様々な取捨選択をおこないつつ，我々の品詞体系への移行をおこなったものである．フォーマットは，MeCab の辞書フォーマットに準拠している．現在も単語の追加を続けており，約 6,000 語を収録している．

古典中国語形態素コーパスは，『漢文大系』[9] から『十八史略』†を中心に例文を選び，複数のコーパス入力者が，それらの例文を単語ごとに区切って，我々の品詞体系で分類したものである．フォーマットは，MeCab のコーパスフォーマットに準拠しており，それをさらに Linked Data 化した上で，CHISE-Wiki の一部として WWW 公開している [6]．現在も例文の追加を続けており，約 46,000 文を収録している．

3 地名の自動抽出

古典中国語すなわち漢文の中に現れる地名に，何がしかの特徴が見られないか，もし見られるのならそれらを自動抽出できないか，というのが，本研究の基本スタンスである．特徴は，文法的な特徴であっても構わないし，あるいは文字づらでの特徴であっても構わない．とにかく，どんな汚い手段を使ってもいいから，漢文中の地名を自動抽出する，というのが我々の目標の一断面である．

その目標に添って，我々は，我々が作成してきた古典中国語形態素コーパスをざっと洗い直してみることにした．特に，我々の品詞分類で「n, 名詞, 固定物, 地名」あるいは「n, 名詞, 主体, 国名」に分類されている単語と，その単語を含む例文を見直してみた．この結果，我々が辿りついたのが，「2 文字の地名には地名以外の用例はない」という仮説だった．たとえば「洛陽」という形態素は，それがどこにあった洛陽なのかは別として，地名以外の単語として使われることはない，という仮説である．

この仮説に基づき，我々は「2 文字の地名」の地名以外の用例を，我々の古典中国語形態素コーパスに対して，サンプリング調査してみた．そうしたところ，そのような地名以外の用例は，どの「2 文字の地名」においても 10%未満だった．しかも，それら 10%足らずの用例も「n, 名詞, 固定物, 地形」など，山や川の名前をコーパス入力者が地形とみなしたものが大多数で，これらを仮に地名だとみなしても大した問題は起こらない．「2 文字の地名には地名以外の用例はない」という仮説は，少な

*本研究は，学術研究助成基金助成金/科学研究費補助金基盤研究 (B) 25280122「品詞素性情報つき古典漢文コーパスの発展的応用」および京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの応用研究」の共同成果である．

† 『十八史略』は，地名，人名，官職を，平易簡略な漢文で，かなり多く含んでいる．

図 1: 形態素解析に特化した古典中国語品詞体系

くとも 90%の確率で当たっており，地名の自動抽出という観点からは，採用するに値する．この結論に基づき，我々は，古典中国語形態素コーパスから抽出した「2 文字の地名」を，そのまま，我々の古典中国語辞書に追加した．また，3 文字以上の地名は，その多くが「○○府」や「○○縣」の形を取るものだったが，同様に古典中国語辞書に追加した．

では,「1 文字の地名」は，どうなのか．たとえば「渭」のように，地名用例しかないような「1 文字の地名」に関しては，そのまま古典中国語辞書に追加すればよい．しかし，たとえば「夏」という形態素は，王朝名としての「夏」かもしれないし，季節としての「夏」かもしれない．あるいは「莫」という形態素は，地名用例はむしろ少数で，大多数の用例が「v, 副詞, 否定, 禁止」である．もし,「莫」を無理矢理に地名だとみなすような処理をおこなうと,「v, 副詞, 否定, 禁止」であるべき「莫」を，誤って「n, 名詞, 固定物, 地名」だと処理してしまう危険性がある．その場合，後続の動詞にも悪影響が及ぶので，文法上のミスとしては致命的である．そのようなミスは，絶対に避けなければならない．

この問題に対し，我々は，たとえ「1 文字の地名」を全て古典中国語辞書に追加したとしても，古典中国語形態素コーパスを十分に準備すれば，MeCab による形態素解析において，そのようなミスは発生しないだろう，という希望的観測を持ってみることにした.「2 文字の地名」という巨大な用例による接続確率 (裏を返せば非接続確率) が効いてくるはずで，それによって「1 文字の地名」も正しく認識されるはずだ，という甘い予想を立てたわけである．

もちろん，この予想がうまくいくためには，他の地名用例コーパスも含め，できるだけ多くの地名用例コーパスが必要な上に，対抗用例コーパスも十全に収録しておかねばならない．たとえば「莫」であれば,「n, 名詞, 固定物, 地名」の「莫」も,「v, 副詞, 否定, 禁止」の「莫」も，いずれも古典中国語辞書に含まれている必要があるし,「莫」の副詞用例コーパスも十全に収録しておかねばならない．また，地名用例コーパスや対抗用例コーパスに加え，それら以外のコーパスも，バランスよく収録しておく必要がある．

この目標のために，我々は，我々が既に入力した約 46,000 文のコーパスから，複数の入力者による分析結果が品詞レベルで完全に一致した用例 (約 2,000 文，地名を約 400 語収録) を，本手法の学習用コーパスとして用いることにした．結論を言えば，この手法によって，我々の古典中国語形態素解析システムは，たとえば「莫滅莫」という (かなり人工的な) 漢文を

莫　v, 副詞, 否定, 禁止
滅　v, 動詞, 変化, 制度
莫　n, 名詞, 固定物, 地名

「莫を滅するなかれ」と正しく処理できるようになった．定性的な観点からは，本手法の有効性が示されたことになる．

4　本手法の評価

ただし，工学的な観点から見た場合，本手法の有効性と，本手法によって引き起こされている悪影響とを，可能であれば定量的に評価すべきである．そのような定量的評価の足がかりとして，我々は，以下の 3 種類の古典中国語辞書を準備した．

Ⓐ 従来，我々が使用してきた古典中国語辞書．

Ⓑ 辞書Ⓐに,「1 文字の地名」も含め，知りうる限りの古典中国語地名を追加した辞書．

Ⓒ 辞書Ⓐから，地名を取り除いた辞書．

辞書Ⓐに収録されていた地名の単語数は 111 語，辞書Ⓑに収録されている地名の単語数は 1,240 語，辞書Ⓒは 0 である．

さらに,「1 文字の地名」文例およびその対抗用例を，地名テストデータ P (88 語) として準備した．以下に，地名テストデータ P の具体例「代王これを聞き大いに恐る」と「瓜に及んで而して代わる」を示す．

代　n, 名詞, 主体, 国名
王　n, 名詞, 人, 役割
聞　v, 動詞, 行為, 伝達
之　n, 代名詞, 人称, 止格
大　v, 副詞, 程度, 極度
恐　v, 動詞, 行為, 態度

	テストデータ P	テストデータ M	テストデータ R
辞書Ⓐ	96/86/85/76	93/90/90/77	96/83/81/71
辞書Ⓑ	96/89/88/84	93/90/90/76	96/83/81/71
辞書Ⓒ	96/86/84/73	93/90/90/77	94/81/79/69

図 2: 各辞書に対する各テストデータの F 値 (大品詞/品詞/意味素性/小素性)

及　v, 動詞, 行為, 移動
瓜　n, 名詞, 可搬, 糧食
而　p, 助詞, 接続, 並列
代　v, 動詞, 行為, 交流

この例では,「代」という漢字が，地名 (王朝名) を指している用例と,「代わる」という動詞として使われている対抗用例とを，テストデータとして用いている．

また，地名テストデータ P との比較検討のために，[4] で用いた M (69 語) と R (320 語) も，テストデータとして用いた．なお，比較を容易にするために，辞書ⒶⒷⒸともに，学習用コーパスは約 2,000 文で固定とした．

実験結果として，各辞書に対する各テストデータの F 値 (大品詞/品詞/意味素性/小素性) を図 2 に示す．地名テストデータ P に関しては，辞書Ⓐより辞書Ⓑの方が F 値が上がっており，我々の手法の有効性が，定量的にも評価されたと言えるだろう．また，辞書Ⓐより辞書Ⓒの方が F 値が低いことから，少なくとも地名テストデータ P に関しては，地名は追加すればするほど良い，という結論になると思われる．実際，地名テストデータ P の中で，F 値の良悪を決定づけていたのは，以下のような例文であった．

晉　n, 名詞, 主体, 国名
克　v, 動詞, 行為, 交流
衞　n, 名詞, 固定物, 地名
磁　n, 名詞, 固定物, 地名
洛　n, 名詞, 固定物, 地名
州　n, 名詞, 制度, 場

「晉は衞，磁，洛州に克つ」である．このような「1 文字の地名」が連続している例文において，辞書Ⓐや Ⓒは,「衞」や「磁」や「洛」を，地名以外の名詞だと誤検出してしまうのである．

一方，テストデータ M については，辞書Ⓑで小素性の F 値がわずかに下がっているものの，全体としてほとんど変化が見られない．テストデータ M には地名用例が含まれていないことから，辞書Ⓑにおける地名の「過学習」は，一般的な漢文の形態素解析に対して，ほとんど悪影響を及ぼさない，と結論づけることができる．

テストデータ R については，辞書Ⓐと辞書Ⓑで F 値に変化がなく，辞書Ⓒで大幅に F 値が下がっている．これは，テストデータ R に地名が含まれており，辞書Ⓒにおいてそれらの地名が取り除かれてしまったために，F 値が下がったと考えられる．一方，辞書Ⓑで追加した地名は，テストデータ R の形態素解析に，良い影響も悪い影響も及ぼしていない．

以上，我々のテストデータに関しては，古典中国語地名を知りうる限り追加した辞書Ⓑが，最も良好な結果を得られたと言える．少なくとも地名テストデータ P に関しては，辞書Ⓑが最も良い結果となっているし，M と R に関しては，辞書Ⓑで追加した地名はほとんど悪影響がなかった．

5　おわりに

古典中国語における地名用例を，形態素解析によって自動抽出する手法を示した．端的には，古典中国語辞書に知りうる限りの地名を追加し，さらに，地名用例を含む古典中国語形態素コーパスと，その対抗用例コーパスの両方を準備する手法を提案した．また，定性的定量的な観点において，本手法の有効性を確認した．

ただ，本手法は，あくまで，古典中国語での地名に限定したものである．古典中国語に現れる地名が，そもそも限定的であり，未知語という考え方をほとんど必要としない，という点には注意が必要である．その点を考えると，本手法は，他の固有表現，たとえば人名や官職の自動抽出には，応用できない可能性が高い．それらの固有表現に対

しては，本手法とは別の自動抽出手法が必要となるだろう．今後さらなる研究を進めていきたい．

参考文献

1) 工藤拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer. 入手先 <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html> (参照 2014-11-03)

2) 守岡知彦: MeCab を用いた古典中国語の形態素解析の試み，情報処理学会研究報告，Vol. 2008-CH-79，pp. 17-22 (2008)．

3) 守岡知彦: MeCab を用いた古典中国語形態素解析器の改良，情報処理学会研究報告，Vol. 2009-CH-84，No. 3，pp. 1-5 (2009)．

4) 山崎直樹，守岡知彦，安岡孝一: 古典中国語形態素解析のための品詞体系再構築，人文科学とコンピュータシンポジウム「じんもんこん 2012」論文集，pp. 39-46 (2012)．

5) Morioka, T., Wittern, C., Yasuoka, K. and Yamazaki, N.: A Study of Linguistic Analysis for Classical Chinese Texts, Proc. *2013 International Conference on Culture and Computing*, pp. 143-144 (2013).

6) 守岡知彦: 古典中国語形態素コーパスの Linked Data 化の試み, 人文科学とコンピュータシンポジウム「じんもんこん 2013」論文集, pp. 187-194 (2013).

7) 「東アジア古典文獻コーパスの研究」共同研究班報告, 東方學報 (京都)，第 88 册，pp. 292-287 (2013)．

8) Yasuoka, K., Yamazaki, N., Wittern, C., Nikaido, Y. and Morioka, T.: A Morphological Analysis of Classical Chinese Texts, *Digital Humanities 2014*, pp. 410-412 (2014).

9) 服部宇之吉，三島毅，重野安繹，竹添進一郎，星野恆，小柳司氣太，安井小太郎，島田鈞一，岡田正之，井上哲次郎：漢文大系，冨山房 (1909-1916).