様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ４ 月 ２６ 日現在

機関番号：１４３０１
研究種目：基盤研究(B)（一般）
研究期間：2013～2015
課題番号：２５２８０１２２

研究課題名（和文）品詞素性情報つき古典漢文コーパスの発展的応用

研究課題名（英文）Applied Study on Morphological Analysis of Classical Chinese Texts

研究代表者
安岡　孝一（YASUOKA, Koichi）

京都大学・人文科学研究所・教授

研究者番号：２０２３０２１１

交付決定額（研究期間全体）：（直接経費）　　13,900,000円

研究成果の概要（和文）：われわれは、MeCabを用いた古典漢文の形態素解析について、その実際的手法と実用性をこれまでに研究してきた。本研究においてわれわれは、これまでのわれわれの手法をさらに拡張し、地名、官職、人名など、漢文における固有表現抽出に挑戦した。本研究の成果として、われわれは、漢文における地名の自動抽出に成功し、官職に対しては、われわれの手法の有効性を呈示することができた。一方、人名に対しては、姓氏の自動抽出には成功したものの、名や諱に対しては有効な手法を構築するに至らなかった。

研究成果の概要（英文）：We have been studing on a morphological analysis of classical Chinese texts based on MeCab. In this applied study we have investigated on extracting named entities from classical Chinese texts, especially names of places, names of official titles, and names of persons. As a result we have succeeded in extracting names of places perfectly. For extracting names of official titles, we have proposed a method to build up long titles. We also have succeeded in extracting surnames of persons from classical Chinese texts, but we could not construct effective methods to extract forenames of persons.

研究分野：人文情報学

キーワード：漢文処理　形態素解析

１．研究開始当初の背景

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は、その前身である附属東洋学文献センター時代から、現在に至るまで、約 120,000 タイトルの古典漢籍文献を収集し、その保存と公開につとめてきた。また、1980 年代から、京都大学大型計算機センターとの共同研究で、古典漢籍の全文テキストデータベース化に携わってきた。

これらの膨大な古典漢文テキストをコンピュータで処理するためには、白文(単なる漢字の列)ではなく、テキストを自然言語解析する必要がある。古典漢文のように、単語の間にも文の間にも区切りを持たない書写言語の解析では、まず、単語を認識することが必須であり、そのために形態素解析をおこなわなければならない。

この問題に対し、研究代表者は、平成 20 年度より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研究」を組織し、古典漢文に対する形態素解析の研究を開始した。この共同研究班において、われわれは、言語に依存しない解析エンジンとして MeCab を選び、さらに、古典漢文を形態素解析するための品詞分類を研究した。また、この共同研究班を母体として、平成 22～24 年度には、科学研究費基盤研究(B)『形態素解析のための品詞情報つき古典漢文コーパスの構築』により、古典漢文コーパスの構築と形態素解析の研究をおこなった。これらの研究成果は、古典漢文コーパスや辞書ファイルも含めて WWW 上で全て公開しており、古典漢文を形態素解析する環境は、ほぼ整いつつある。

しかし、上記の研究は同時に、古典漢文の解析において、形態素解析というものの限界を示す研究でもあった。すなわち、同一の単語が複数の用途に用いられている場合、それらの違いは、形態素レベルでは単純には解析できないのである。例を挙げると、「引置左右」という漢文に対し、われわれの形態素解析システムは、「引」は動詞、「置」は動詞、「左右」は名詞、というレベルまでは解析が可能であり、いわゆる単語切りをおこなうことはできる。しかし、この漢文で「左右」が官職を示しているという解析までは、研究開始当初おこなえていなかった。それはすなわち、官職としての「左右」の用例を検索したい研究者に対して、それを他の「左右」から分離して呈示するような検索システム、言いかえるなら、古典漢文の固有表現抽出をおこなうようなシステムは、まだ実現できていなかった。

２．研究の目的

本研究の主眼は、固有表現抽出に特化した MeCab 漢文コーパスの再構築と、形態素解析を応用した単語間共起関係解析システムの設計である。先の「引置左右」の例に即して言えば、「置」と「左右」の接続確率を用いた形態素解析の上に、「引」と「左右」の意味素性を含む共起確率を解析するシステムを構築して、「左右」が官職である確率を導出するシステムを設計する。卑近な言い方をすれば、「引置左右」の「左右」が官職だと当ててみせるシステムを、本研究で実現しようと考えた。

３．研究の方法

本研究の中心をなすのは、品詞・素性情報つき MeCab 漢文コーパスである。品詞・素性情報つき MeCab 漢文コーパスの各例文は、元になる漢文を単語に分解し、それぞれに大品詞・品詞・意味素性・小素性を付加したものである。たとえば「天帝使我長百獣」という例文に対しては、以下のようになる。

天帝	n,名詞,人,役割
使	v,動詞,行為,使役
我	n,代名詞,人称,止格
長	v,動詞,行為,設置
百獣	n,名詞,主体,動物

大品詞は古典漢文の動賓構造に対応しており、「v」(動)、「n」(賓)、「p」(その他)の 3 種類としている。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 種類としており、「形容詞」はない。意味素性は「人」「主体」「行為」「不可譲」「固定物」「可搬」「制度」など 44 種類を、小素性は 88 種類を、本研究開始までに準備したが、研究の進捗にしたがって拡充をおこなう。

本研究における MeCab 漢文コーパスは、『漢文大系』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、われわれの品詞体系で分類する形で作成する。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 公開する。さらに、これらの MeCab 漢文コーパスを元に、地名・官職・人名など固有表現抽出の検討をおこなう。

４．研究成果

(1)地名の自動抽出

漢文での地名を自動抽出する、という目標に向け、われわれは、それまでに作成してきた MeCab 漢文コーパスを洗い直してみた。特に、われわれの品詞体系において「n,名詞,固定物,地名」あるいは「n,名詞,主体,国名」に分類されている形態素オブジェクトと、その形態素オブジェクトを含む文例を見直してみた。この結果、われわれが辿り着いたのが、「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。

この仮説に基づき、われわれは「2 文字の地名」の地名以外の用例を、MeCab 漢文コーパスに対して、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、山や川の名前を例文入力グループが地形だ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、われわれは、MeCab 漢文コーパスから抽出した「2 文字の地名」を、そのまま MeCab 漢文辞書に追加した。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 MeCab 漢文辞書に追加した。

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 MeCab 漢文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、われわれは、たとえ「1 文字の地名」を全て MeCab 漢文辞書に追加したとしても、MeCab 漢文コーパスを十分に準備すれば、そのようなミスは形態素解析において発生しないだろう、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 MeCab 漢文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、われわれは、それまでに入力していた約 46,000 文の MeCab 漢文コーパスから、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。

この手法により、われわれの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
```

「莫を滅すなかれ」と正しく処理できるようになった。また、この手法を定量的に評価すべく、地名を加えない MeCab 漢文辞書との比較をおこなった。この結果として、できる限り多くの地名を MeCab 漢文辞書に追加する手法は、地名を含む漢文の認識精度を高めると同時に、地名を含まない漢文には悪影響がない、という形で、本手法の有効性が確認された。

(2)官職の自動抽出

漢文における官職を自動抽出する際も、文字数の短い官職であれば、地名と同様の手法が効果的だった。実際、MeCab 漢文辞書と MeCab 漢文コーパスを十全に準備することで、たとえば「上下左右」の「左右」と、「引置左右」の「左右」を

```
上下　n,名詞,固定物,関係
左右　n,名詞,固定物,関係

引　　v,動詞,行為,動作
置　　v,動詞,行為,設置
左右　n,名詞,人,役割
```

という形で正しく見分けることは、われわれの形態素解析システムでは既に可能となっている。

その一方、複数の形態素から構成される(ように見える)官職もあり、これがわれわれを悩ませた。以下に、いくつかの典型例を示す。

・丞

「○○丞」の形を取る名詞は、ほぼ全て官職とみなせる。しかしながら、その形態素解析処理は問題を孕んでいる。たとえば「御史中丞」を一つの形態素だとみなしてしまうと、「右御史中丞」や「知御史中丞」をうまく処理できない。「右御史臺中丞」となると、もうどうしていいかわからない。また、「右」は必ずしも最初に付加されるとは限らず、「尚書右丞」「尚書左丞」のような例もある。これらに加え、「湖州長城丞」や「長沙縣丞」のように地名との複合が起こる場合もあって、混沌を極める。

・郎中

「○○郎中」の形を取る名詞は、まず間違

いなく官職である。これらのうち、「兵部郎中」や「司勳郎中」のように、部署名や他の官職との単純な複合は、まだ何とか処理できる。しかしこれが、「兵部左司郎中」や「尚書司勳郎中」という形で複合すると、もはや形態素解析の手に負えない。

・判～事、知～事、～従事

「判○○事」「知○○事」「○○従事」の形を取る名詞は、かなりの確率で官職である。しかしながら、形態素解析の立場からすると、「判」「知」「従」はいずれも動詞とみなすべき形態素であり、これが問題を複雑にしている。たとえば「知民事」は、通常は「民事を知る」という文であって、官職ではない。一方「知政事」は官職である。あるいは「知吏部尚書事」は官職だが、内部に他の官職である「吏部尚書」を含んでしまっている。

複数の形態素から構成される官職は、形態素処理の後に「組み上げ処理」をおこなうことで、抽出可能だと考えられる。しかしながら、この「組み上げ処理」は、たとえば「丞」と「郎中」と「事」とで、全く異なる処理をおこなわざるを得ない。つまり、官職中に使われている文字ごとに異なった処理が必要で、それぞれをバラバラに力業で組み上げるしかない、というのが、われわれの結論である。

(3)人名の自動抽出

漢文における人名の自動抽出に向けて、われわれは、MeCab 漢文コーパスを洗い直し、「n,名詞,人,姓氏」「n,名詞,人,名」に分類されている形態素オブジェクトと、その形態素オブジェクトを含む文例を見直してみた。その結果、「n,名詞,人,姓氏」については、地名抽出と同様の手法が有効だ、との感触が得られた。しかし、「n,名詞,人,名」については、他の用例とのバッティングが、奇妙な方法で回避されていることが判明した。具体例として、『十八史略』巻之二に現れる「李斯」という人名に関して、われわれが得た知見を、以下に述べる。

『十八史略』巻之二には、「斯」という文字が、全部で 16 例、出現する。これらのうち 6 例は、「李斯」という形で出現することから

李　n,名詞,人,姓氏
斯　n,名詞,人,名

であることは確実であり、実際の形態素解析においても、そう処理できる。問題は残る 10 例である。これら 10 例は「斯」が単独で出現するのだが、われわれの判断では、最初の 9 例は全て「n,名詞,人,名」であり、最後の 1 例だけが「n,代名詞,指示,*」なのである。具体的には、「李斯」の話が続いている間は、ずっと「斯」は特定の人名である「李斯」を指しており、その後「李斯」が出てこなくなって、かなり文章が進んでから、やっと代名詞の「斯」がたった 1 例だけ出現する。つまり、「李斯」の話が続いている間は、話がややこしくならないよう、代名詞の「斯」の使用をあえて避けているわけである。

このような形で、「ストーリー」全体における用字の分布に異常が見られる場合、それが人名を指している可能性が高い、ということは推定できた。「斯」の例で言えば、曖昧語となりうる「斯」の曖昧性を下げるために、代名詞としての「斯」を使わない、という形で『十八史略』巻之二における用字分布が変わってしまっている。しかしながら、この推定を、人名の自動抽出にまで結びつけるような手法は、われわれには開発し得なかった。というのも、「ここで斯が出てきたなら、それは李斯であって、代名詞じゃないよな」ということを理解するには、本質的には「ストーリー」の理解が必要だからである。現状のわれわれの力不足を、痛感する限りである。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 8 件）

① 安岡孝一　Christian Wittern　守岡知彦　池田巧　山崎直樹　二階堂善弘　鈴木慎吾　師茂樹、古典中国語(漢文)の形態素解析、東洋学へのコンピュータ利用、査読無、第 27 巻、2016、pp.3-14

② Koichi Yasuoka　Naoki Yamazaki　Christian Wittern　Yoshihiro Nikaido　Tomohiko Morioka、A Morphological Analysis of Classical Chinese Texts、Proceedings of Digital Humanities、査読有、Vol.2014、2014、pp.410-412

③ Christian Wittern 、Kanripo and Mandoku: Tools for Distributed Repositories of Premodern Chinese Texts 、Proceedings of Digital Humanities、査読有、Vol.2014、2014、pp.408-409

④ 安岡孝一　守岡知彦　Christian Wittern　山崎直樹　二階堂善弘　鈴木慎吾、古典中国語形態素解析による地名の自動抽出、人文科学とコンピュータシンポジウム「じんもんこん」論文集、査読有、Vol.2014、2014、pp.63-68

⑤ 守岡知彦、古漢字データベースの要件に関する試論、情報処理学会研究報告、査読無、Vol.2014-CH-103(5)、2014、pp.1-7

⑥ 守岡知彦、比較的最近の CHISE、東洋学へのコンピュータ利用、査読無、第 25 巻、2014、pp.33-46

⑦ 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」論文集、査読有、Vol.2013、2013、pp.187-194

⑧ Tomohiko Morioka Christian Wittern Koichi Yasuoka Naoki Yamazaki、A Study of Linguistic Analysis for Classical Chinese Texts、Proceedings 2013 International Conference on Culture and Computing, IEEE、査読有、Vol.2013、2013、pp.143-144

〔学会発表〕（計 8 件）

① 安岡孝一、古典中国語(漢文)の形態素解析、東洋学へのコンピュータ利用　第 27 回研究セミナー、2016 年 3 月 18 日、京都大学(京都市)

② 安岡孝一、古典中国語形態素解析による地名の自動抽出、じんもんこん 2014、2014 年 12 月 13 日、一橋講堂(東京都千代田区)

③ 守岡知彦、古漢字データベースの要件に関する試論、人文科学とコンピュータ研究会、2014 年 8 月 2 日、兵庫県立歴史博物館(姫路市)

④ Koichi Yasuoka 、A Morphological Analysis of Classical Chinese Texts、Digital Humanities 2014、2014 年 7 月 11 日、ローザンヌ(スイス)

⑤ Christian Wittern 、Kanripo and Mandoku、Digital Humanities 2014、2014 年 7 月 9 日、ローザンヌ(スイス)

⑥ 守岡知彦、比較的最近の CHISE、東洋学へのコンピュータ利用、第 25 回研究セミナー、2014 年 3 月 14 日、京都大学(京都市)

⑦ 守岡知彦、古典中国語形態素コーパスの Linked Data 化の試み、じんもんこん 2013、2013 年 12 月 14 日、京都大学(京都市)

⑧ Koichi Yasuoka、A Study of Linguistic Analysis for Classical Chinese Texts、2013 International Conference on Culture and Computing、2013 年 9 月 16 日、立命館大学(京都市)

〔図書〕（計 1 件）

① 二階堂善弘　師茂樹　他、好文出版、論集：中国学と情報化、2016、137

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
ホームページ等
http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/archive2016.html

６．研究組織

(1)研究代表者
安岡　孝一（YASUOKA, Koichi）
京都大学・人文科学研究所・教授
研究者番号：２０２３０２１１

(2)研究分担者
山崎　直樹（YAMAZAKI, Naoki）
関西大学・外国語学部・教授
研究者番号：３０２３０４０２

二階堂　善弘（NIKAIDO, Yoshihiro）
関西大学・文学部・教授
研究者番号：７０２９２２５８

師　茂樹（MORO, Shigeki）
花園大学・文学部・教授
研究者番号：７０３５１２９４

ウィッテルン　クリスティアン
（WITTERN, Christian）
京都大学・人文科学研究所・教授
研究者番号：２０３３３５６０

池田　巧（IKEDA, Takumi）
京都大学・人文科学研究所・教授
研究者番号：９０２５９２５０

守岡　知彦（MORIOKA, Tomohiko）
京都大学・人文科学研究所・助教
研究者番号：４０３２４７０１

鈴木　慎吾（SUZUKI, Shingo）
大阪大学・言語文化研究科・講師
研究者番号：２０５１３３６０