E-020

古文における形態素解析実用化に関する研究

Research on practical morphological analysis in classical literature

柳川　亮†　　河合　敦夫†　　桝井　文人†　　井須　尚紀†

Ryo Yanagawa　Atsuo Kawai　Fumito Masui　Naoki Isu

1. はじめに

近年のテキスト電子化の流れに伴い，形態素解析や構文解析といった基礎技術の開発が進められてきた．しかし，これらの技術がほぼ実用段階である現代文と比較して，古文における自然言語処理技術は未だ未成熟である．特に古文には，省略や体言止めといった，作者の意図や作為的技法が多く用いられるため，その解析は容易ではない．

安武ら[1]は早くから古文における形態素解析システムの開発を行っており，文節数最小法[2]が古文においても有効であることを示している．しかし，文節数最小法で絞り込んだ回答候補が複数ある場合，その中からの正解の選択については触れておらず，また文節数最小法で取りこぼす可能性のある，複合語を含んだ文についても対応できていない．

松本ら[3]は，形態素解析システム JUMAN，及び茶筌を用いて，古文の形態素解析を行っている．同システムでは，利用者が文法を自由に定義することが可能であるため，古典文法に基づく形態素辞書および文法規則を与えることで，古文の形態素解析を行うことができる．しかし，これらは文法規則の一部である単語間の連接頻度を，品詞タグ付きの正解データから学習しているため，大規模な正解データのない古文においては，一様に精度の高い解析を行う事が難しい．

本研究では，茶筌を用いた古文の形態素解析実用化を目指し，システムの学習に必要な品詞タグ付き正解データ作成と，半自動型の形態素解析システムの構築を行った．

また，作成された時代幅の大きい古典文学では，時代ごとの特徴，あるいは作者による違いが，少なからず解析精度に影響を与えるのではないかという考えから，物語ごとの解析精度や特徴を調査した．

2. タグ付き正解データ作成

品詞タグ付き正解データを全て人手で作成するには，多くの時間と，解析のための専門的な知識が必要になる．本章では，人的負担の軽減を目的とした半自動型形態素解析システムの説明を行う．

図 1 では，入力文からタグ付き正解データを作成するまでの処理の流れを示している．システムには初期の辞書として，古文用茶筌形態素辞書を用いた．

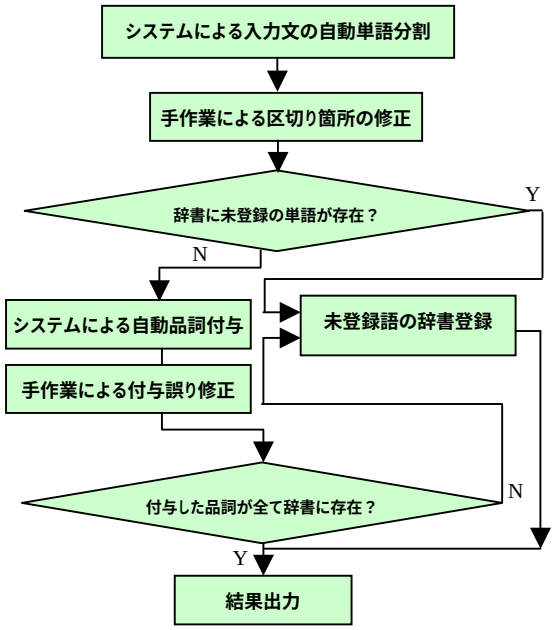

図 1.品詞タグ付き正解データ作成の流れ

まず，入力文を自動で単語単位に分割し，手作業で区切り箇所の修正を行い，誤りのない区切り結果を取得する．分割した単語群に辞書登録されていない単語が存在した場合，品詞情報を与え辞書に登録する．全て辞書登録されていれば，それぞれの単語に対し，単語分割部と同じく品詞連接頻度と単語出現頻度を用いて，自動で品詞付与を行う．また，手作業により品詞付与誤りを修正した結果，修正後の品詞情報が辞書になければ，辞書に追加登録する．

単語分割部と品詞付与部を分けることで，精度を下げやすい単語分割部が，品詞付与部に与える影響を低く抑えた．また，それぞれの自動処理に用いる品詞連接頻度，及び単語出現頻度は，既に作成済みの品詞タグ付き正解データから実際の頻度を算出し用いた．

以上の一連の行程をシステム化し，複数の古典文学作品について解析を行い，正解データを作成した．

3. 解析実験

作られた年代や作者の異なる物語間では，品詞連接や出現単語の差異が，解析精度に影響を与える．本実験では，

† 三重大学工学部情報工学科人工知能講座

これら複数の物語間での精度比較を行い，年代，あるいは作者による違いを調査した．自動による単語分割は解析精度が低く，自動品詞付与に大きく影響を与えるため，物語ごとの正確な精度比較が困難になる．そこで今回は，自動品詞付与部分のみの精度比較とし，単語分割結果は誤りのないものを用いた．

実験対象データとして，源氏物語から『桐壺』と『帚木』，『枕草子』，『伊勢物語』，及び『徒然草』を用いた．源氏物語以外の物語については，冒頭から同程度の単語数の文章について解析を行った．データとして用いた文章の単語分割数を表 1 に示す．

表 1.各データの単語数

	桐壺	帚木	枕草子	伊勢物語	徒然草
単語数	5304	3309	3364	3261	3017

単語出現頻度は全てのデータの合計で算出し，品詞連接頻度に関しては，自身のデータ以外の品詞連接頻度の和を用いて，クロスバリデーションにより解析精度を測定した．結果を表 2 に示す．また，各データごとの特徴を比較するため，それぞれの品詞連接頻度を用いて各データの解析を行い，その精度を調査した．結果を表 3 に示す．

表 2.クロスバリデーションによる各データの解析精度

	桐壺	帚木	枕草子	伊勢物語	徒然草
解析精度(%)	94.8	95.5	94.1	94.8	95.2

表 3.学習データを変えた各物語の解析精度比較

		学習データ				
		桐壺	帚木	枕草子	伊勢物語	徒然草
評価データ	桐壺	97.3	95.2	94.0	94.1	95.0
	帚木	94.4	97.2	94.0	94.0	94.7
	枕草子	92.6	94.2	96.5	93.4	94.0
	伊勢物語	91.7	93.1	92.8	96.7	93.6
	徒然草	92.1	94.7	92.7	92.9	96.5

4. 考察

表 2 では，およそ一様に 95%付近の解析精度となった．対して表 3 では，同じ源氏物語である『桐壺』と『帚木』は比較的近い解析精度であるのに対し，歌物語である『伊勢物語』では，自身以外の学習データを用いた場合，全体的に解析精度が低くなった．学習データを『桐壺』とした場合，同一の物語である『帚木』，同年代に作られた『枕草子』，特殊な歌表現が多く出現する『伊勢物語』の順に解析精度が下がっている．しかし，成立年代が他と異なる『徒然草』では予想した程の大きな精度差は見られなかった．それぞれのデータ間の品詞連接頻度を比較したところ，名詞，形容詞，連体詞など，比較的多義でない単語に関しては大きな違いが見られなかったが，動詞，助詞，助動詞，補助動詞など，多義になる可能性が高く，物語の特徴を左右するような単語に関しては，いくつか明確な違いが見られた．例えば『桐壺』では尊敬表現が助動詞，補助動詞を合わせておよそ 300 箇所以上出現するのに対し，伊勢物語ではほとんど用いられていないことがこれにあたる．

次に，クロスバリデーションを用いて全てのデータを解析した際の，解析誤り 828 箇所のうち，頻出した誤りを表 4 に示す．

表 4.各データで多く見られる解析誤り例

正解	誤り	頻度
断定の助動詞連用形『に』	格助詞『に』	47
断定の助動詞連用形『に』	接続助詞『に』	14
断定の助動詞連用形『なり』	四段型動詞『なり』	32
接続助詞『に』	格助詞『に』	35
接続助詞『を』	格助詞『を』	41
体言止め／終止形終り	終止形終り／体言止め	92

これら，特に断定の助動詞連用形『に』と格助詞『に』の取り違え誤りは，どのデータでも上位 3 件以内に入っており，特に解析の難しい単語である．この断定の助動詞連用形『に』は，本来，後ろ数単語以内に補助動詞『あり』や『はべり』を，あるいは接続助詞『て』を伴うとされているが，『桐壺』の冒頭にもあるように，補助動詞が省略されていて，判断の難しい場合もいくつか見られた．また，その他の解析誤りとして，連体形と終止形が同じである四段型や上一段型活用では，体言止めと終止形終りとの区別が付きづらく，両者の取り違えによる誤りが多かった．

5. おわりに

古文における形態素解析実用化を目的として，正解データ作成と一連のシステム化を行った．今後は，物語の種類を増やすとともに，現在使用している物語に関しても文章数を増やし，物語間での特徴差異がどのように変化するか調べる必要がある．また，今回精度低下の要因となった解析誤り箇所について，現在バイグラムである前後間の品詞連接頻度をトライグラムにするなど，可変長に対応することで，より精度の高い品詞付与を目指す予定である．

文献

[1] 安武 満佐子, 吉村 賢治, 首藤 公昭, “古文の形態素解析システム,” 福岡大学工学集報第 54 号,March 1995.

[2] 吉村 賢治, 日高 達 他 , “文節数最小法を用いたべた書き日本語文の形態素解析,” 情報処理学会論文誌, Vol.24,No.1, 1983.

[3] 山本 靖, 松本 裕治, “日本語形態素解析システム JUMAN による古文の形態素解析とその応用,” 情報処理学会文学研究会第 19 回研究発表大会, July 1996.