ユーザビリティを考慮した古典研究用形態素解析ツールの構築

柳川　亮　　桝井　文人　　河合　敦夫　　井須　尚紀
三重大学大学院　工学研究科

１．　はじめに

古典文学研究に関する分野において，文学資料の電子化[1][2]が進んでいる．そのため，これらの分野においても自然言語処理ツールへの要求が高まっている．

これらの要求に答えるものとして，安武ら[3]や山本ら[4]の研究がある．安武らは，古文から現代文への翻訳の一環として形態素解析システムを構築し，文節数最小法による解析が有効であることを示している．また，山本らは，辞書構築とコスト学習によって，現代文用に構築された形態素解析システム JUMAN[5]が，古文にも利用可能であることを示している．しかしながら，時代や作者が異なる書物を扱う場合には，語彙や統語構造の異なるテキストを処理する事になるため，汎用的な性能だけでは，網羅的に高精度の解析を行うことは困難である．

研究対象となる古典文学テキストは，現代文と比べ記述された年代が幅広く，様々なジャンル（例えば物語や随筆）に及ぶ．そのため，それらの文体は，記述された時代や文化的要因の影響，作者やジャンルの差異などを反映し，形態素や品詞，活用などの出現分布が大きく変化する[6][1]．このような差異を吸収し，かつ安定した精度を確保するためには，汎用性だけでなく，個々のテキストに効率よく最適化できるユーザビリティやメンテナンス性も重要である．

本論文では，このような古典文学における研究支援を目的として，汎用的性能を確保しつつ，適合フィードバックによって解析性能を容易に最適化できる古典研究用形態素解析システムについて述べる．

以下，2 章では提案する古典研究用形態素解析システムの概要について述べ，3 章ではシステムを用いて古文テキストの解析実験を行い，4 章でそれらを考察する．

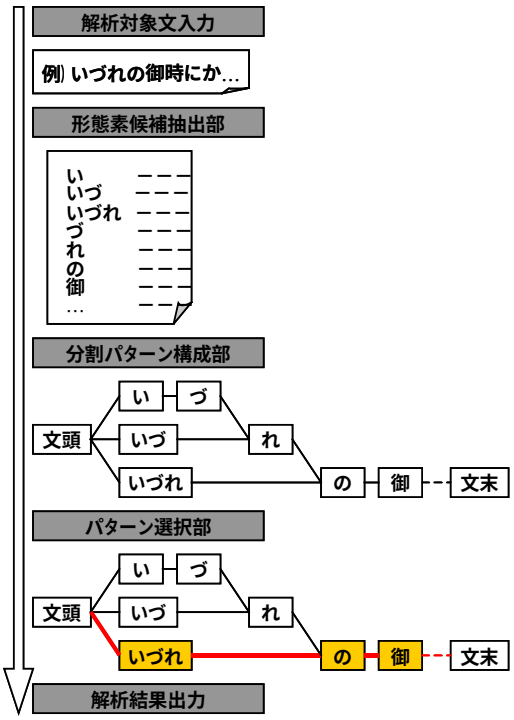

図１．システム構成

２．　システムの概要

本章では，古典研究用形態素解析システムの概要について説明する．本システムの構成を図１に示す．

２．１．システム構成

本システムは，(1) 形態素候補抽出部，(2) 分割パターン構成部，(3) パターン選択部より構成される．以下，各処理部について詳述する．

(1) 形態素候補抽出部

本処理部では，システムへの入力文に対し，一文字ずつずらしながらCommon Prefix Search[2]を行い，辞書に登録された形態素と一致する，文中の全ての部分文字列を抽出する．抽出された文字列は形態素候補としてパターン構成部に渡される．Prefix Searchで用いる形態素辞書は，山本ら[4]のJUMAN用形態素辞書をベースに，さらに自立語を

[1] また，和歌の技法である本歌取りのように，過去の作品に影響を受けることもある．

[2] 検索キーのprefix（接頭文字列）になっている辞書エントリを取り出すアルゴリズム

約 800 語，非自立語を約 30 語拡張し用いている．
(2) 分割パターン構成部
抽出された形態素候補を用いて，入力文を表現可能な，全ての組合せパターンを構成する．
(3) パターン選択部
構成した組合せの中から，コスト最小法を用いて最適経路を決定する．

２．２．コスト最小法

コスト最小法とは，図１にあるようなラティス状のグラフの，全てのノード（形態素）とリンク（形態素同士の連接）に適当なコストを与え，コストの合計値が最小な経路を最適解として選択する手法である．i番目のノードのコストを形態素コストM_i，i番目とi+1 番目のノード間のリンクのコストを連接コスト$R_{i,i+1}$と定義し，以下にコスト合計値の計算式を示す．Nは形態素総数である．

$$Sum\_Cost = \sum_{i=0}^{N} \left| M_i + R_{i,\,i+1} \right|$$

本システムでは，この二つのコストを，ユーザが目的文書に応じて適合フィードバックにより最適化する．各コストの初期値は，後述する実験で用いた古文テキストを人手で解析し，形態素（品詞）の出現傾向，及び形態素間の連接傾向を調査し，さらに文法書の定義を調査した結果を総合して与えた．

３．　実験と評価

３．１．実験内容

本実験では，源氏物語の桐壺を用いて適合フィードバックを行い，解析精度の変動について検証した．適合フィードバックには，桐壺中の 3216 形態素分のテキストを用い，残りの 969 形態素分のテキストを評価データとして用いた．フィードバック回数は 3 回とし，各フィードバック時にそれぞれのコストを調整した回数を表 1 に示す．

表1．各フィードバック時のコスト調整回数

	一回目	二回目	三回目
連接コスト	23	17	14
形態素コスト	51	34	23

さらに，桐壺における適合フィードバックが，桐壺以外の古文に対しどのような影響を与えるか比較するために，同じ源氏物語である帚木と，枕草子，徒然草，伊勢物語について，各適合フィードバック時点での解析性能を調査した．評価に用いた各データ量は表２に示す．

表2．評価に用いた形態素数

	桐壺	帚木	枕草子	伊勢物語	徒然草
形態素数	969	981	1034	1091	998

３．２．評価結果

評価基準には以下の式を用いて F 値を算出し用いる．正解データは人手で形態素解析を行い作成した．

$$Precision = \frac{\text{システムの正解数}}{\text{システムの出力数}}$$

$$Recall = \frac{\text{システムの正解数}}{\text{正解データ数}}$$

$$F\text{-}measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

図 2 は，複数回適合フィードバックを行った際の，各評価データにおける形態素分割のＦ値の推移を示している．また，形態素の分割に加え，品詞ラベルの付与を行った際のＦ値の推移を図 3 に示す．

さらに，図 2，3 について，ベースラインからフィードバック後までのＦ値の向上率を算出し，図 4，5 に示した．

４．　考察

本章では評価結果について考察する．

適合フィードバックの結果，全てのデータで形態素分割，品詞ラベル付与共に性能向上が見られた．特に適合フィードバックに用いた桐壺が，最も高い向上率を示したことから，古文の形態素解析における提案手法の有効性が確認できた．

桐壺と同作品である帚木が，二番目の向上率を示していることから，作品内での類似性が確認できる．これは，作品の一部に対する適合フィードバックが作品全体に有効に働いた結果である．また，Ｆ値の向上率の上がり幅は，フィードバック回数が増すにつれて，減少している．今回の実験

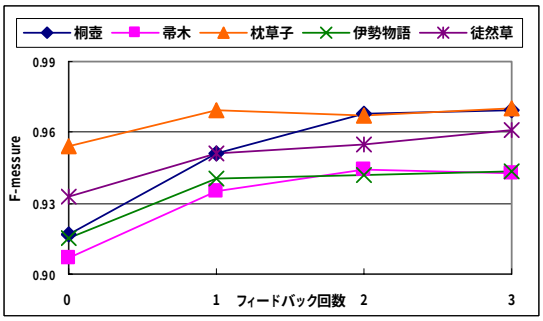

図２．形態素分割のＦ値の推移

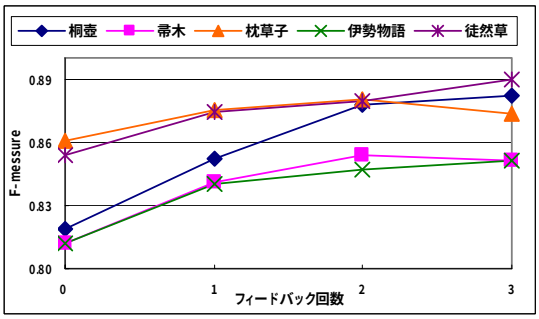

図３．品詞ラベル付与のＦ値の推移

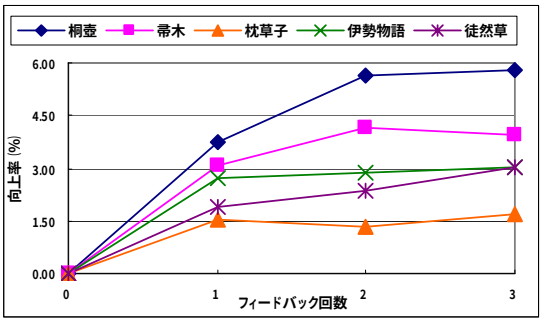

図４．形態素分割のＦ値の向上率

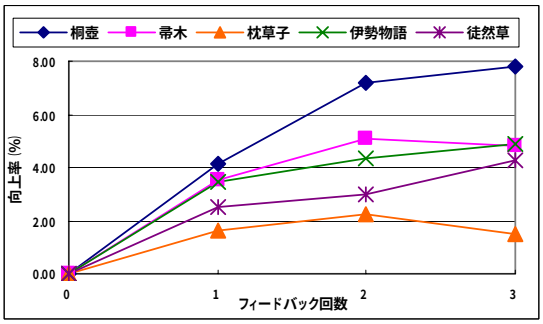

図５．品詞ラベル付与のＦ値の向上率

では，フィードバック回数を 3 回までとしたが，3 回目のフィードバック時には，Ｆ値の向上率はほぼ頭打ちになっている．以上の二点から，作品全体を網羅するような適合フィードバックを行う必要はなく，また少量のフィードバックにより解析性能を早期に向上させることができると考えられる．これらは，フィードバック作業にかかるユーザへの負担が小さく，ユーザビリティが高いことを示している．

次に，解析誤りと向上率の変化について考察する．フィードバック回数が少ない段階では，出現頻度の小さい品詞に関する誤りが多く見られた．これは，初期コストを与える際に，人手で解析した古文テキストから，品詞の出現頻度や品詞間の連接頻度を調査し参考にしたため，出現頻度が十分でない品詞について，連接コストが大きく設定されていたことが原因である．図 4，5 の 1 回目のフィードバック時に，全てのデータで見られた性能向上は，以上の部分が改善された事を示唆している．

図 4，5 で 2 回目以降の向上率は，桐壺と帚木以外では，明確な向上はない．これは，適合フィードバックに用いた桐壺が宮廷物語であり，『たまひ』などの尊敬の補助動詞や，『せ』などの尊敬の助動詞が多く出現したため，フィードバック調整が尊敬表現を中心に行われたことを示唆している．このことは，尊敬表現が使われ難い，随筆である枕草子と徒然草が，共に低い向上率を示している事とも一致する．

以下の例は，3 回目のフィードバック後にも見られた解析誤りである．これらは作品を問わず共通して現れている．

- 『なむ』『にて』などの助詞,助動詞に絡む誤り
- 『え・・・ず』などの係り受け表現での誤り
- 定型的に用いられる特殊表現

助詞，助動詞に絡む表記は図 6 のように多義である場合が多い．これらは，同一作品内でも複数の意味で用いられることが多く，どちらかに重きを置くことは難しい．例えば，『なむ』を強意表現と捕らえるか，完了表現と捕らえるかは文脈などに依存し，コストによる調整では対応が困難である．また，係り受け表現や定型表現なども，品詞連接のケースが特殊であることが多く，通常のコストによる判定だけでは対応が困難である．

このような，コストによる処理では判断が困難な解析誤りに対しては，ルールベースを作成し，

『なむ』	⇒ 係助詞 ⇒ 終助詞 ⇒ 完了助動詞『ぬ』未然形＋推量助動詞『む』 ⇒ ナ変動詞未然形語尾＋推量助動詞『む』

図6. 表記『なむ』の多義解釈

処理を行うことが有効と考える．図 7 はルールベースを構築する際の，基準となる文法規則の一例である．

i. 体言･助詞＋なむ
⇒ 係助詞
ii. 未然形＋なむ
⇒ 終助詞
iii. 連用形＋な＋む
⇒ 完了助動詞『ぬ』未然形＋推量助動詞『む』
iv.『死ぬ･往ぬ』未然形＋む
⇒ ナ変動詞未然形語尾＋推量助動詞『む』

図7. 多義判断ルールベース一例

現状で見られる誤りの多くは，文法的なルールベースを組み込むことで, 対処可能である. また, 今回の実験でも見られたように，助動詞は文書固有の特徴となりやすく，係り受けや定型表現も同じ傾向にある．ユーザの観点からも，作品の特徴として理解しやすいため，これらのルールに重みを与え，目的文書に応じた調整をすることは比較的容易である．加えて，ユーザが個々でルールベースを追加することで,よりユーザビリティに富んだ柔軟な解析が可能になる.

５．　おわりに

本論文では,様々な性質を持つ古文作品に対し，柔軟な解析を可能にするため，ユーザビリティを考慮した古典研究用形態素解析システムの構築を行った．実験の結果，ジャンル，作成年代の異なる複数の古文作品において，およそ 87%の解析精度となった．また，目的文書に合った調整を行うことで，より高精度の解析精度が得られることを確認した.

今後は，考察で述べたルールベース処理を加え，より高精度の解析を目指すと共に，視覚的に分かりやすいツールとして確立する予定である.

参考文献

[1] 国文学研究資料館, 国文学データベース研究集会法報第 4 号，1994

[2] 安永尚志：“ 日本古典文学の本文データベース ”，情報処理，Vol.35，No.7，pp.642-650，1994

[3] 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，1995

[4] 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 ”，情報処理語学文学研究会第 19 回研究発表大会，1996

[5] 松本裕治，黒橋禎夫，宇津呂武仁，妙木裕，長尾真:“ 日本語形態素解析システム JUMAN 使用説明書 version2.0 ”，Information Science Technical Report NAIST-IS-TR94025，Graduate School of Information Science，Nara Institute of Science and Technology，1944

[6] 伊藤雅光：“ 計量言語学入門 ”，大修館書店，2002