ISS-A-035

古文書における形態素解析を用いた単語検出

Word detection using morphological analysis in ancient documents

片山歩希 [1]　　松尾賢一 [1]
Ayuki Katayama　　Ken-ichi Matsuo

奈良工業高等専門学校 専攻科 [1]
National Institute of Technology, Nara College, Faculty of Advanced Engineering

1　はじめに

初めて古文書を解読する（以降，翻刻）には，まず古文書から一文字を見つけ出し，その文字を現代の文字に置き換える必要がある．置き換えの次に，各文字同士のつながりを考えて，単語を作ることで文章の理解ができるようになる．

しかし，この一文字の置き換えから単語を作るまでのステップには，多くの経験を積まないと進むことができず，翻刻初学者にとって大きな壁となる．

今村の研究 [1] により，言語初学者が単語を自動的に把握することは内容理解に認知資源を使えるため，読解に重要であることが分かっている．そこで古文書でも同様に，翻刻初学者が単語を自動的に把握する方法として形態素解析を用いて古文書に書かれる文章から単語を検出し，検出した単語部分に傍線を引くことで翻刻初学者に強調して提示する．

前述の方法を実現するために本研究の目的は，古文書における形態素解析を用いた単語検出である．検出された単語から成功例と失敗例を比較し，提案手法が単語として成立するかを確認する．

2　単語検出の提案手法

古文書内の単語を検出する方法の一つとして，形態素解析を用いる．この形態素解析を用いることで，数字や地名などの辞書に載らない単語も提示が可能である．

現代語で使われる係り受け規則 [2] を参考に，本提案手法である次の表 1 合成語のルールを作成する．合成語は取得した品詞同士における係り受けの処理により単語を検出できる．

表 1　**合成語のルール**

ルール	用法
1	数詞同士を足し合わせる
2	連体詞と名詞を足し合わせる
3	接頭辞と名詞/動詞を足し合わせる
4	数詞と助数詞を足し合わせる
5	名詞と接尾辞を足し合わせる
6	人名と一般を足し合わせる

3　単語検出に関する実験

3.1　実験目的

本実験では，形態素解析・合成語のルールを用いたとき，古文書に対する単語検出の成功例と失敗例の結果を明らかにする．単語検出の結果を比較することで更なる合成語のルールが必要であるかを判断する．

3.2　実験方法

本実験で取り扱う古文書は，歴史資料がデジタルデータとして保存されている「みんなで翻刻」から鳴門教育大学が所蔵する「諸御趣意書并御下知向之写」,「公辺御役人為御廻浦御立越諸御用一巻」である．形態素解析エンジンは，MeCab を利用し，参照するデータベースは UniDic が公開する近世文語 UniDic[3] を扱う ．

本実験では単一の形態素から成る単語を単純語，複数の形態素から成る単語を合成語として扱うことで，形態素と単語を区別する．

3.3　実験結果

次に単語の検出結果の一例を示す.

成功した単語：御茶屋，十四日，鶴林寺，村役人
失敗した単語：五 尺，西 矢野 村，右 之 通

図 1　**検出した単語（正解と失敗)**

図 1 から提案手法による合成語のルールによって,「御」や「日」などの接頭辞・接尾辞が単語として検出された．しかし，連続した地域の名前・「之」などの代名詞は表 1 では単語として検出できないことが分かる．

4　考察

結果から本提案手法である合成語のルールでは，検出できる単語は接頭辞・接尾辞や数詞の連結などと限りがある．そこで失敗した単語を検出するために連続した地名や「之」などの代名詞に対応する更なるルールの追加が必要である．

5　おわりに

本実験の結果では失敗の単語を含んでいた．しかし，翻刻初学者に提示するためには失敗を含まない単語検出を目指す必要性がある．

参考文献

[1] 今村一博．初級英語学習者と中級英語学習者の読解中の眼球運動はどのように異なるか？LET 関西支部研究集録, Vol. 20, pp. 19–32, 2022.

[2] 宮崎正弘ほか．係り受け解析を用いた複合語の自動分割法．情報処理学会論文誌, Vol. 25, No. 6, pp. 970–979, 1984.

[3] 松本 裕治小木曽 智信．歴史的日本語資料を対象とした形態素解析．自然言語処理, Vol. 20, No. 5, pp. 727–748, 2013.