ひらがな語の追加と形態素解析の精度についての考察

情報科学科　林 聖人　　指導教員：山村 毅

1　はじめに

言葉を機械で取り扱う分野を自然言語処理という. その歴史は, 古くコンピュータの出現とほぼ同時に始まっている. これまでに数多くの研究がなされ, 多くの自然言語処理システムが開発されてきたが, 特に誤りなどの雑音に弱いことが指摘されている.

本研究では, 堅牢な形態素解析システムの実現を目標にひらがな語の追加が形態素解析システムの精度にどう影響するかを調べる.

2　形態素解析

自然言語処理を行う場合, 基本的な処理として文章を単語ごとに区切るというものがある. これは文字列を自然言語処理の階層的構造における形態素の列に変換するということであり, これに「各形態素の品詞を同定すること」と「活用語処理を行うこと」を加えた処理全体のことを形態素解析 (morphological analysis) という [?].

たとえば,「彼は大学で勉強します」を形態素解析すると, 表 1 のようになる.

表 1　形態素解析

形態素	彼	は	大学	で	勉強	し	ます
品詞	代名詞	副助詞	普通名詞	格助詞	さ変名詞	動詞	助動詞

一般に形態素解析は, 入力文のすべての部分文字列を辞書引きし, ノードを作成し, 文頭から文末までを過不足なく被覆し, ある評価関数値が最大となるようなノードで選択することで行われる. 辞書引きの結果, 各要素の品詞も求められる. 評価関数値は人手によるコスト, 隠れマルコフモデルや条件付き確率場による確率が用いられる [?].

3　ひらがな語の追加

形態素解析をする基本的な手法については, 広く研究され, 自然言語処理システムを開発するためのツールがいくつか公開されているが, いずれもひらがな語が多く入った文の解析精度が悪いという問題がある. これは,

- 形態素解析システムの学習に, 通常の新聞記事を用いている.
- 通常, 漢字を用いることが多い言葉は, ひらがなでは、辞書に登録されていない.

ということに帰因するものと考えられる.

そこで, 辞書中に漢字を含んで登録されている単語を, ひらがなでも登録することによって, ひらがな語を含む, 文章の解析精度がどう変化するかを調べる.

4　評価実験

辞書中の漢字を含んだ単語を, ひらがなに変換して, 新たに単語として登録して辞書を拡張する. 追加前の単語数は 392126 個でひらがな語の追加後は 712593 個となった. この拡張した辞書を用いて, 形態素解析を行い, どの程度正しく, 形態素の区切りを抽出できているか（形態素区切り適合率・再現率）, どの程度正しく, 形態素の品詞を同定しているか（品詞適合率・再現率）を調べた.

評価に用いた文は毎日新聞, あおぞら文庫の小説, 小学校国語教科書（1 年〜4 年）から収集した 297 文である. あるいは, これらの文をそのまま（オリジナル）/すべてをひらがなに変換したもの（ひらがな）を用いて適合率, 再現率を求めた. また比較のために, 辞書を拡張しなかったものについても, 同様に適合率, 再現率を求めた. 結果得られた総計 1188 個 (297 × 4) の解析結果とあらかじめ作成しておいた正解と比較し, 形態素の区切りと品詞の数を手作業で比較し数えた.

形態素解析器には MeCab を, 辞書には IPADIC を用いた.

解析結果を表 2,3 に示す.

表 2　全体統計

	オリジナル	ひらがな	オリジナル辞書追加	ひらがな辞書追加
形態素区切り適合率	97.8	90.7	99.3	97.7
形態素区切り再現率	99.6	98.2	99.8	99.1
品詞適合率	95.4	81.2	97.9	94.5
品詞再現率	97.2	88.0	98.4	95.9

表 3　対象別品詞適合率

	オリジナル	ひらがな	オリジナル辞書追加	ひらがな辞書追加
小説	97.9	80.0	96.8	92.5
毎日新聞	99.9	63.6	99.7	94.7
小学 1 年生	86.7	85.8	99.0	98.3
小学 2 年生	93.9	88.2	97.3	96.0
小学 3 年生	96.6	93.4	97.0	93.6
小学 4 年生	94.7	86.0	97.7	94.0

表 2 より全体として文節適合率, 再現率また品詞適合率, 再現率がすべてひらがな語を追加した方が精度が高いことがわかる. しかし表 3 の対象別で見ると, 小学 3 年生の結果のようにひらがな語を追加してもあまり変化のない部分があることがわかる.

表には現れていないが, ひらがな語を追加したことで, 逆に上手く解析できなくなったものが「オリジナル」で 19 例,「ひらがな」で 31 例あった.

5　まとめ

本研究によりひらがな語を辞書に追加することによって形態素解析の精度が上がることがわかった. 但し, 逆に悪くなる場合もあった. 今後はひらがな語を含めて, 形態素解析システム全体を学習し直すことで, 精度がどう変化するか調べたい.

参考文献

[1] 松本 裕治:"形態素解析システム「茶筅」", 情報処理学会誌情報処理 VOL.41 NO.11, 2000

[2] 森信介:"形態素解析", 情報処理学会誌情報処理 VOL.57 NO.1, 2016