MeCabによる平仮名のみの形態素解析

井筒 順　　明石 陸　加藤 涼　岸野 望叶　小林 汰一郎　金野 佑太　古宮 嘉那子
茨城大学工学部情報工学科

{17t4013g, 17t4002x, 17t4025h, 17t4029g, 17t4036s,
17t4039x, kanako.komiya.nlp}@.vc.ibaraki.ac.jp

1　はじめに

日本語の形態素解析は高性能であり，実用レベルで使用されている。

一方で，主に日本語学習者に，平仮名だけの形態素解析を行いたいというニーズが存在する。全て平仮名で書かれた文を形態素解析する場合，既存のツールでは十分に形態素解析をすることが難しい。なぜならば，既存のツールは，漢字かな交じり文の形態素解析を解析することを想定して作成されているからである。

例えば,「市役所の市民課の職員の私生活」という文を MeCab [1] で形態素解析することを考える。デフォルトの設定で形態素解析を行えば，正しい解析結果が得られる。しかし，この文を全て平仮名にした文である「しやくしょのしみんかのしょくいんのしせいかつ」を同様に形態素解析すると，以下のような誤った結果が得られる。

```
$ mecab
しやくしょのしみんかのしょくいんのしせいかつ
し 動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
や 助詞,並立助詞,*,*,*,*,や,ヤ,ヤ
くし 名詞,一般,*,*,*,*,くし,クシ,クシ
ょのしみんかのしょくいんのしせいかつ 名詞,一般,*,*,*,*,*
EOS
```

さらに，日本語を全て平仮名で表すと，その文の曖昧性が非常に高まる。例えば,「帰る時にバイバイする」という文を考える。これを全て平仮名に変換すると「かえるときにばいばいする」となる。以下にこの例文において考えられる意味を列挙する。

- 帰る時にバイバイする
- 帰る時に売買する
- 帰る朱鷺にバイバイする
- 孵る朱鷺にバイバイする
- 買える時に売買する
- 蛙と木に売買する

上記の文から，平仮名のみの文は，意味的な曖昧性が解消されていないため，単語分割が特に難しいことことが分かる。

本研究では，上記のような全て平仮名で書かれた文を高い精度で形態素解析するため，形態素解析ツールの MeCab を利用し，平仮名のみに特化した形態素解析を実行して分析する。

2　関連研究

日本語の形態素解析は，本来平仮名と漢字が混合している文を入力として想定しているため，平仮名文のみの文に対する精度は低い。これは，藤田ら [1] も指摘しているように，一般的な形態素解析モデルを構築する際に用いられる学習データと，解析対象であるデータの構成が大きく異なっているからである。このような課題を解決するために，様々な手法が提案されている。

工藤ら [2] は，平仮名混じり文が生成される過程を生成モデルでモデル化し，そのパラメータを大規模 Web コーパス及び EM アルゴリズムで推定することで，平仮名交じり文の解析精度を向上させる手法を提案している。大崎ら [3] は，文中の特徴的な表現を新たな名称として扱うことで，コーパスの構築を行っている。また，藤田ら [1] は既存の辞書やラベルありデータを，

[1] https://taku910.github.io/mecab/

対象分野の特徴に合わせて自動変換し，それを使って形態素解析モデルを構築する教師なし分野適応手法を提案している。林ら [4] は, 平仮名語の単語を辞書に追加することによって, 形態素解析の精度が向上することを報告している。

本論文では，平仮名に特化した形態素解析を行うために，平仮名のみの辞書とコーパスを用いて MeCab の学習を行い，その効果を調べた。

3 提案手法

関連研究の節でも述べたように，学習データとテストデータの性質が異なる場合，形態素解析の精度は低下してしまう。そのため，平仮名文のみの形態素解析するには，テストデータにあった辞書とコーパスを作成する必要がある。つまり本研究でいえば，全て平仮名で構成された辞書とコーパスの作成が必要となる。

このためには，平仮名文のみで構成された膨大な数の辞書とコーパスを集める，もしくは一から作成する必要があるが，既存の辞書と文に対し何かあるアクションを加え，その結果として平仮名文のみで構成された辞書と文を得ることができたならば，膨大な数の辞書とコーパスを集め，一から作成する必要はなくなる。

そこで本稿では，平仮名文のみで書かれた文を既存の設定よりも高精度に形態素解析するために，既存の辞書を平仮名文に変換したものを辞書として用い，かつ，平仮名のみで構成されたコーパスを学習データとして用いる。

4 実験

本実験では，MeCab を利用する際に，利用する辞書内に存在する全ての csv ファイルの表層系を平仮名に変換したものを平仮名のみの形態素解析用の辞書として使用する。また，Wikipedia のデータを MeCab を使って形態素解析し，それを平仮名に変換したものを平仮名のみで構成されたコーパスとして用いる。

本研究では，既存の辞書として，mecab-ipadic-2.7.0-20070801(以下 ipadic) を利用する。コーパスとしては，Wikipedia のダンプデータ（後述）を用いた。

4.1 コーパスの変換

まず，もともとの ipadic 辞書を使った MeCab により，Wikipedia のデータの形態素解析を行い，結果をファイルに保存する。このファイルを A とする。

次に，ファイル A の形態素解析された結果の内，分ち書きされている部分を平仮名に変換する。これにより，変換された例を以下に示す。

変換前:

郵便 名詞, 一般,*,*,*,*, 郵便, ユウビン, ユービン

変換後:

ゆうびん 名詞, 一般,*,*,*,*, 郵便, ユウビン, ユービン

この平仮名に変換する操作をファイル A の全ての行に対して行い，結果をファイルに保存する。このファイルを B とする。このファイル B を本実験における学習用コーパスとして用いる。

4.2 辞書の変換

ipadic に存在する全ての csv ファイルに対し，ファイル中の各行の表層系を平仮名に変換し，そのディレクトリ上に上書き保存をする。変換した後，コマンド mecab-dict-index を使用し，そのディレクトリに学習用バイナリ辞書を作成する。このディレクトリを C とする。ディレクトリ C に作成したこれらの辞書を，平仮名用の形態素解析のための辞書として利用した。

次に，ファイル B とディレクトリ C に対し，コマンド mecab-cost-train を使用して CRF パラメータの学習を行い，モデルを生成した。ここで，CRF のハイパーパラメータは 1.0 と設定した。作成したモデルはディレクトリ C に保存する。

上記で作成したモデルを使用し，コマンド mecab-dict-gen を使用して評価用辞書の作成を行った。これにより生成された評価用の辞書が平仮名文のみを形態素解析可能な辞書となる。また，コマンド mecab-dict-index を使用し，解析用バイナリ辞書の作成を行なった。

作成した評価用辞書の評価はコマンド mecab-system-eval を使用し形態素解析の精度を測った。5 分割交差検定により形態素解析結果の妥当性を判定する。

4.3 実験データ

本実験で使用した，Wikipedia のデータは，次の Web サイト

https://dumps.wikimedia.org/jawiki/latest/

において公開されている

jawiki-latest-pages-articles.xml.bz2

を展開し，その一部である 3,623 文をコーパスとして使用した。コーパス中の単語の出現数は 96,402 である。

また既存の辞書として ipadic 辞書内に存在する csv ファイルおよび設定ファイルを使用した。

次に，5 分割交差検定により学習用コーパスを 5 つに分けた際の各学習用コーパスを CRF パラメータの学習の際に表示された文の数と素性 (単語分割，品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3，活用型，活用形，基本形，読み，発音) の数の数を以下に示す。

表 1: CRF パラメータの学習中のデータ

分割	文の数	素性の数
1/5	2,956	910,811
2/5	2,875	926,114
3/5	2,852	921,346
4/5	2,941	921,037
5/5	2,868	920,996
合計	14,042	4,600,304

また本稿において，「N 番目の辞書」とは，学習用コーパスを 5 分割した際，5 分割したコーパスの内，N 番目のものをテストに使用し，それ以外の 4 つを学習用コーパスとして作成した辞書を指す。

5 実験結果

表 2 は作成した辞書を用いて 5 分割交差検定を行った結果である。

分割においてそれぞれに場合における，分ち書きの精度と，素性全ての評価値（全素性が正しく推定できたときにだけ正解とする）の結果を表 2 として以下に示す。

次に，もともとの ipadic 辞書を用いたテスト結果を表 3 として以下に示す。

表 2: 作成した辞書での評価

分割	精度	再現率	F 値
1/5	70.43	73.49	71.93
2/5	62.41	63.52	62.96
3/5	77.83	78.75	78.29
4/5	65.07	68.29	66.63
5/5	68.85	70.29	69.56
平均	68.92	70.87	69.88

表 3: もともとの ipdadic 辞書での評価

分割	精度	再現率	F 値
1/5	44.73	46.40	45.55
2/5	45.78	45.93	45.85
3/5	39.75	39.95	39.85
4/5	49.07	49.96	49.51
5/5	45.98	46.70	46.33
平均	45.06	45.79	45.42

表 2，表 3 から既存の辞書よりも我々が作成した辞書の方が平仮名文の形態素解析の精度が高いことが分かる。

6 考察

本実験により，平仮名文を形態素解析した際の正答率が既存の辞書の正答率よりも高いことを確認できた。

また，1 節で形態素解析をした文を作成した辞書で形態素解析すると以下のようになった。ここで使用した辞書は 5 分割交差検定で作成した 1 番目の辞書である。

```
$ mecab -d final1_5
しやくしょのしみんかのしょくいんのしせいかつ
しやくしょ 名詞,一般,*,*,*,*,市役所,シヤクショ,シヤクショ
の 助詞,連体化,*,*,*,*,の,ノ,ノ
しみん 名詞,一般,*,*,*,*,市民,シミン,シミン
か 名詞,接尾,一般,*,*,*,科,カ,カ
の 助詞,連体化,*,*,*,*,の,ノ,ノ
しょくいん 名詞,一般,*,*,*,*,職員,ショクイン,ショクイン
```

```
の 助詞,連体化,*,*,*,*,の,ノ,ノ
しせいかつ 名詞,一般,*,*,*,*,私生活,シセイカツ,シセイカツ
EOS
```

この評価から形態素解析を既存の辞書よりも高い精度行うことができたことが分かる。

一方で，形態素解析された単語が意味する漢字が何であるかを推測することは難しく，市民課の課を科と推測している。

上記と同様に「かえるときにばいばいする」という文をを 1 番目の辞書を用いて形態素解析すると以下のようになる。

```
$ mecab -d work1_5/final1_5/
かえるときにばいばいする
かえる 動詞,自立,*,*,一段,基本形,変える,カエル,カエル
とき 名詞,非自立,副詞可能,*,*,*,とき,トキ,トキ
に 助詞,格助詞,一般,*,*,*,に,ニ,ニ
ばいばい 名詞,サ変接続,*,*,*,*,売買,バイバイ,バイバイ
する 動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
EOS
```

5 つ作成した字辞書の内 4 番目の辞書以外は上記のように形態素解析を行った。しかし，4 番目の辞書は「かえる」の意味を「蛙」の意味にとってしまった。これは，学習コーパスの内容が違うために発生したと考えられる。

以上より，平仮名のみの形態素解析は，分かち書きは高い精度で行うことはできるが，分かち書きした単語が何を意味するかを推測することが難しいことが分かる。

7 おわりに

本研究では，全て平仮名で書かれた文を形態素解析する辞書の作成を行った。また，Wikipedia の形態素結果を平仮名に置き換えたコーパスで学習して実験したところ，作成した辞書の解析結果が ipadic 辞書の結果を上回った。

一方で，形態素解析結果における，品詞や，その単語が表す漢字の正答率は高くなく，さらなる学習が必要である。

これは，既存の辞書内の平仮名に変換する漢字において，平仮名に変換する際に正しく変換できていない文字が複数あることが原因であると考えられる。平仮名に変換する際の精度を高めることで結果が変わった可能性があるのかを再度検証していきたい。

また，分かち書きした単語が何を意味するかは MeCab ではラティス構造を用い，生起コストと連接コストを用い解析を行っている。生起コストと連接コストだけでなく，その単語の前後の文脈から生成される単語の種類を分類することができると考えられる。例えば,「ばいばいする」を売り買いをする売買なのか，さよならをするバイバイなのかを判断する際に，単語の前後の文に「市場」や「お金」等の単語があれば売買と判断でき,「帰る」や「下校」等の単語があれば，バイバイと判断できるのではないであろうか。これらは，仮名漢字変換や語義曖昧性解消の研究と関連して解くことが必要になると考えられる。

今回の実験で使用したコーパス内の文字数は 96,402 個である。コーパスの量を増やすことや漢字を平仮名に変換する精度を高めることでより高い精度で平仮名文を形態素解析する辞書の作成を行いたい。

参考文献

[1] 藤田早苗, 平博順, 小林哲生, 田中貴秋. 絵本のテキストを対象とした形態素解析. 自然言語処理, Vol. 21, No. 3, pp. 515–539, 2014.

[2] 工藤拓, 市川宙, David Talbot, 賀沢秀人. Web 上のひらがな交じり文に頑健な形態素解析. 言語処理学会 第 18 回年次大会 発表論文集, pp. 1272–1275, 2012.

[3] 大崎彩葉, 唐口翔平, 大迫拓矢, 佐々木俊哉, 北川善彬, 堺澤勇也, 小町守. Twitter 日本語形態素解析のためのコーパス構築. 言語処理学会 第 22 回年次大会 発表論文集, pp. 16–19, 2016.

[4] 林聖人, 山村毅. ひらがな語の追加と形態素解析の精度についての考察析. 愛知県立大学情報科学部平成 28 年度卒業論文要旨, pp. 1–1, 2017.