日本語形態素解析における辞書登録の自動化手法とアルゴリズムの提案

迫田　翼
放送大学大学院
1818000940@campus.ouj.ac.jp

1　はじめに

近年、対話システムを搭載したスマートフォンやコミュニケーションロボット、スマートスピーカーなどの対話システムの普及や SNS を利用したビッグデータ解析の台頭により、自然言語処理を行う機会が多くなってきている。

自然言語の解析手法の一つに形態素解析がある。形態素解析では文章を形態素と呼ばれる、意味の分かる最小の単位に分割し、品詞情報や読みの情報を持たせる。一般的に形態素の抽出はあらかじめ言葉の辞書を作成しておき、その辞書を参照することで形態素に分けている。

しかし、近年に活発化している対話システムや自然言語のビッグデータ解析において、形態素解析の対象となる文章は話し言葉である場合や、SNS で多く見られる新しい言葉、すなわち未知語である場合が多くなると考えられる。

そのため、形態素解析器が本来抱えている、未知語に対処できないという問題が浮き彫りになっている。また、未知語に対する処理として多くの研究がなされているが、時間と経過とともに未知語処理に対応した辞書が古くなり、同様の問題が発生する場合や、別途辞書を用意したとしても専門用語など分野によっては解析できないといって問題を抱えている。

本研究では、日本語の文章構造が持つ仕組みを利用した未知語の抽出と機械学習を用いた判別機能を利用し、解析対象の文章ごとに適した辞書を適時作成することで、先に述べたような辞書が古くなる問題、辞書が様々な分野を網羅できないといった問題の解決を試みる。

上記の方法を用いて対話システムで用いられるような文章や様々な分野の文章を形態素解析する場合に、より求められるような形態素を抽出することができるようになると考えられる。また、本研究で提案する手法はオンラインに頼らない仕組みのため、対話システム等の活用範囲の拡大が期待できる。

本手法は時間経過や分野による影響を受けにくく、活用範囲の制限を極力抑えた複合語処理におけるアルゴリズム提案を目的とする。

2　関連研究

日本語形態素解析における辞書の自動化手法については、既存研究が行われており、主に形態素解析器の特性を利用した未知語抽出の手法が提案されている。

柴田ら[1]の研究では、Wikipedia を用いて解析し、テキスト内にある未知語を解析する。JUMAN を用いて解析結果が未定義語一語になるもの、一文字形態素のみからなるものを一形態素と見なし処理をする。この Wikipedia 辞書に加えて膨大な Web テキストを脇村ら[2]の手法を用いて自動獲得辞書を作成する。そのため最終的なシステムとしては入力されたテキストに対して形態素解析を行うとき、JUMAN が持つ基本語彙辞書、Web 自動獲得辞書、Wikipedia 辞書の３つの辞書を用いて解析を行い、その後構文解析器では複合名詞に対応した Wikipedia 辞書を用いて分析を行っている。

このように既存研究では複数の辞書を活用することで未知語や複合語に対応できることが報告されている。

3　提案手法

本研究で提案する複合語処理の流れを図１に示す。

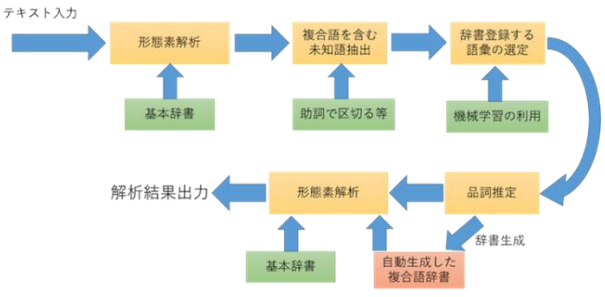

図 1　提案手法の処理の流れ

本研究で提案する手法では、初めに MeCab を用いて形態素解析を行う。次に形態素解析の結果得られた品詞情報を用いて、文章を「助詞」、「助動詞」、「記号（ただし品詞細分類 1 が一般であるもの)」、「括弧開」、「括弧閉」で分割する。分割して得られた文字列を複合語として適切か否かについて、機械学習を用いて判別し、適切と判断されたものを複合語辞書として登録する。

登録された複合語の品詞については各形態素の品詞情報から、「固有名詞」、「名詞（品詞細分類 1 が一般であるもの)」、数値処理、その他のように優先順位を設けて各形態素の品詞情報を適用している。数値処理に関しては「名詞, 数」の品詞を検出しかつ、先に挙げた優先順位の条件を満たす場合、品詞を一律「名詞, 一般」として扱う。

登録された複合語の読みについては、各形態素の読み情報を利用し、順番に結合することで読み情報とする。なお、登録した複合語の中に動詞が存在し動詞の原形に関する情報が複数ある場合は、2 つ目以降の情報を「()」で囲み登録する。

最後に MeCab の標準の辞書と本手法により作成された複合語辞書を用いて解析対象の文書に対して形態素解析を実施する。

4　機械学習

3 の提案手法で示した解析の流れのうち、機械学習について述べる。

機械学習は単純パーセプトロンをベースとしたものを用いるが、OK 用パーセプトロンと NG 用パーセプトロンを用意し、それぞれパーセプトロンの出力結果を SoftMax 関数にて入力された複合語の候補が辞書へ登録すべきか否かを判断する。

機械学習に用いる入力データは、3 の手順で品詞に分割した際の各形態素の品詞に対して固有の番号を割り振り、割り振られた番号の最大値でそれぞれの固有の番号を除算し、正規化したものを入力データとして、学習時と解析時にそれぞれ用いている。

学習データを用いて学習を行った場合は、各パーセプトロンの重みを記録したデータと正規化した値とそれに対応する品詞を対にして記録したルールデータをそれぞれ出力する。

解析時は、ルールデータを用いて、解析文書を数値化したうえで、パーセプトロン重みを適用して解析を行う。

5　実験

5.5 実験条件

今回提案したアルゴリズムの妥当性を検証するため、人によるアンケート調査を行う。

初めに、機械学習に必要な学習を行うため宮沢賢治の作品から「銀河鉄道の夜[3]」、「注文の多い料理店[4]」、「セロ弾きのゴーシュ[5]」、「グスコーブドリの伝記[6]」を学習用コーパスとして用意し、学習率 P は 0.001 とした。

解析するコーパスは、Yahoo ニュースのエンタメ[7]、IT・科学[8]、Wikipedia の機械学習に関する記事の見出し[9]、放送大学 HP の心理学のページ[10]を対象として評価を行った。

アンケート調査については本手法で抽出された複合語をコーパスごとに一覧で表示し、複合語と思われるものには「○」、そうでないものには「×」を記入させ、その結果を集計した。アンケート回答数は 41 人で、男女比、年齢比は表 1 のとおりである。

表 1　アンケート調査における男女、年齢の比

年代＼性別	男性	女性	不明	合計
20代	4.9%	4.9%	0.0%	9.8%
30代	12.2%	0.0%	0.0%	12.2%
40代	34.1%	7.3%	0.0%	41.5%
50代	14.6%	4.9%	0.0%	19.5%
不明	0.0%	0.0%	17.1%	17.1%
合計	65.9%	17.1%	17.1%	100.0%

5.6 実験結果

実験結果のうち、Yahoo ニュースのエンタメ分野より「吉本興業　岡本社長が２２日に会見へ」における本手法による語彙の抽出結果とアンケート結果を表 2 に示す。

表 3 には表 2 で示した抽出された語彙を基に品詞・読み推定を行った結果を示す。

表 2　エンタメ分野の語彙抽出結果とアンケート結果

語彙	○の割合
会見	90.2%
松本	87.5%
このまま	95.1%
ダウンタウン	95.1%
レギュラーコメンテーター	85.4%
ロンドンブーツ1号2号	87.5%
上層部	95.1%
全員クビ	85.4%
全面謝罪する	58.5%
危機感	95.1%
反社会的勢力	97.6%
契約解除	97.6%
宮迫博之	92.5%
岡本昭彦社長	95.0%
岡本社長	100.0%
断片的	95.1%
松本人志	95.0%
緊急生放送	92.7%
謝罪会見	97.6%
謹慎処分	100.0%
金銭受領	95.1%
関係者	100.0%
闇営業	90.2%
雨上がり決死隊	90.0%
10時	92.5%
2択	90.2%
2日前	95.1%
22日	92.5%
47	70.0%
49	70.0%
VTR出演	82.9%

表 3　エンタメ分野の品詞・読み推定の結果

語彙	品詞・読み
会見	記号,空白,*,*,*,*, 会見, カイケン, カイケン
松本	名詞,固有名詞,人名,姓,*,*, 松本, マツモト, マツモト
このまま	連体詞,*,*,*,*,*,このまま,コノママ,コノママ
ダウンタウン	名詞,一般,*,*,*,*,ダウンタウン,ダウンタウン,ダウンタウン
レギュラーコメンテーター	名詞,一般,*,*,*,*,レギュラーコメンテーター,レギュラーコメンテーター,レギュラーコメンテーター
ロンドンブーツ1号2号	名詞,固有名詞,地域,一般,*,*,ロンドンブーツ1号2号,ロンドンブーツイチゴウニゴウ,ロンドンブーツイチゴーニゴー
上層部	名詞,一般,*,*,*,*,上層部,ジョウソウブ,ジョーソーブ
全員クビ	名詞,一般,*,*,*,*,全員クビ,ゼンインクビ,ゼンインクビ
全面謝罪する	名詞,一般,*,*,*,*,全面謝罪する,ゼンメンシャザイスル,ゼンメンシャザイスル
危機感	名詞,一般,*,*,*,*,危機感,キキカン,キキカン
反社会的勢力	名詞,一般,*,*,*,*,反社会的勢力,ハンシャカイテキセイリョク,ハンシャカイテキセイリョク
契約解除	名詞,サ変接続,*,*,*,*,契約解除,ケイヤクカイジョ,ケイヤクカイジョ
宮迫博之	名詞,固有名詞,人名,姓,*,*,宮迫博之,ミヤサコヒロユキ,ミヤサコヒロユキ
岡本昭彦社長	名詞,固有名詞,人名,姓,*,*,岡本昭彦社長,オカモトアキヒコシャチョウ,オカモトアキヒコシャチョー
岡本社長	名詞,固有名詞,人名,姓,*,*,岡本社長,オカモトシャチョウ,オカモトシャチョー
断片的	名詞,一般,*,*,*,*,断片的,ダンペンテキ,ダンペンテキ
松本人志	名詞,固有名詞,人名,姓,*,*,松本人志,マツモトヒトシ,マツモトヒトシ
緊急生放送	名詞,一般,*,*,*,*,緊急生放送,キンキュウナマホウソウ,キンキューナマホーソー
謝罪会見	名詞,サ変接続,*,*,*,*,謝罪会見,シャザイカイケン,シャザイカイケン
謹慎処分	名詞,サ変接続,*,*,*,*,謹慎処分,キンシンショブン,キンシンショブン
金銭受領	名詞,一般,*,*,*,*,金銭受領,キンセンジュリョウ,キンセンジュリョー
関係者	名詞,サ変接続,*,*,*,*,関係者,カンケイシャ,カンケイシャ
闇営業	名詞,一般,*,*,*,*,闇営業,ヤミエイギョウ,ヤミエイギョー
雨上がり決死隊	名詞,一般,*,*,*,*,雨上がり決死隊,アメアガリケッシタイ,アメアガリケッシタイ
10時	名詞,数,*,*,*,*,10時,イチゼロジ,イチゼロジ
2択	名詞,一般,*,*,*,*,2,ニニ
2日前	名詞,数,*,*,*,*,2日前,ニニチマエ,ニニチマエ
22日	名詞,数,*,*,*,*,22日,ニニニチ,ニニニチ
47	名詞,数,*,*,*,*,47,ヨンナナ,ヨンナナ
49	名詞,数,*,*,*,*,49,ヨンキュウヨンキュー
VTR出演	名詞,一般,*,*,*,*,VTR出演,ブイティーアールシュツエン,ブイティーアールシュツエン

アンケートの収集結果として、初めに「吉本興業　岡本社長が２２日に会見へ[7]」の解析結果に対するアンケート調査の結果、80%以上「○」が付いた、つまり受け入れられた語彙の割合は 90.3%であった。

次に「参院選のキーワード１位は「難しい」...SNS の反応を分析してみた[8]」の解析結果に対するアンケート調査の結果、80%以上「○」が付いた語彙は全体の 72.5%であった。

「バックプロパゲーション[9]」の解析結果を表 4 に示す。アンケート調査の結果、80%以上「○」が付いた語彙は全体の 37.9%であった。

表 4　機械学習分野の解析結果とアンケート結果

語彙	○の割合
10][2]ら	10.0%
1960年	95.0%
1967年	95.0%
1969年	95.0%
1974年	95.0%
1986年	95.0%
2層	92.5%
3層以上	95.1%
ArthurE.Bryson	55.0%
B.Widrow	50.0%
M.E.Hoff,Jr.ら	43.9%
[1]	42.5%
[9]	42.5%
ござぎゃくでんぱほう	22.5%
デビッド・ラメルハートら	48.8%
デルタルール	57.5%
何度	87.8%
出力誤差	82.9%
可微分	61.0%
多段動的システム最適化手法	63.4%
機械学習	70.7%
活性化関数	63.4%
甘利俊一	85.0%
発表以降ニューラルネットワーク研究	45.0%
確率的勾配降下法	63.4%
英:Backpropagation	40.0%
英語版	97.6%
誤差伝播	56.1%
隠れ層	52.5%

最後に「認定心理士の資格取得を目指す方へ[10]」の解析結果に対するアンケート調査の結果、すべての語彙について 80%以上「○」が付いた。

今回の実験について、全体を通して 80%以上「○」が付いた語彙の割合は 75.2%であった。

6　考察

5 の実験結果より、全体を通して集計すると 80%以上「○」が付いたことで、複合語として受け入れられたと言える割合は 75.2%と比較的高い精度であることが分かった。唯一、「バックプロパゲーション[9]」の解析結果のみ 37.9%とその他の結果に比べて著しく低いが、アンケート結果から、「B.Widrow」といった人名や「活性化関数」や「隠れ層」といった機械学習の専門用語が低い数値になっている傾向があるため、その語彙が正しいか判断がつかなかったものと思われる。

品詞推定については、表 3 にある「松本人志」のように、品詞が「名詞・固有名詞・人名」となっており正確に推定できていると考えている。読み推定についても「マツモトヒトシ」とあり、正確に推定できている。

一方で、品詞推定については表 3 の「ロンドンブーツ 1 号 2 号」の品詞のように「名詞・固有名詞・地域」となってしまっている誤りがある。これは各形態素のうち「ロンドン」の品詞を優先的に適用したため、「地域」という品詞が当たったと考えられる。

また、読み推定についても表 3 の「10 時」では、読みが「10 時, イチゼロジ」となっており、一形態素の場合と複合語になった場合で読み方が変わる語彙については現状のままでは対応できないことが分かった。

7　展望と今後

今回の実験においては、機械学習に用いた学習データが宮沢賢治の作品であり、今回の解析で用いたコーパスとは時代の違う文書でありながら、比較的高い精度で複合語を抽出できている。このことから時間経過による辞書が古くなる問題の解決できる可能性を見出したと考えている。また、今回解析用として用意したコーパスはそれぞれ分野の異なるものでありながら、それぞれ辞書作成が行えている。そのため分野によって辞書が対応できなくなる問題についても本手法による有用性を見いだせたと考えている。

今回の手法ではスタンドアロンで動作するため、オフライン環境での対話システムでの利用が期待される。

また、時代の違うコーパスで学習させた場合でも、実験での語彙の抽出精度は比較的高かったことから新語や略語、特定の分野の言葉などが含まれていても対応可能であると考えられるため、データマイニング等でも活用することができる。

今後は特に品詞・読み推定について精度を上げるための手法について研究を進めていきたい。

8　参考文献

[1] 柴田 知秀, 村脇 有吾, 黒橋 禎夫, 河原 大輔. (2012). 実テキスト解析を支える語彙知識の自動獲得. 言語処理学会 第 18 回年次大会発表論文集, pp.81-84, 京都大学大学院情報学研究科

[2] 村脇 有吾, 黒橋 禎夫. (2010). 形態論的制約を用いたオンライン未知語獲得. 自然言語処理 Vol. 17 No.1, pp.56-75

[3] 宮沢賢治. 銀河鉄道の夜(青空文庫). https://www.aozora.gr.jp/cards/000081/card43737.html

[4] 宮沢賢治. 注文の多い料理店(青空文庫). https://www.aozora.gr.jp/cards/000081/card43754.html

[5] 宮沢賢治. セロ弾きのゴーシュ(青空文庫). https://www.aozora.gr.jp/cards/000081/card470.html

[6] 宮沢賢治. グスコーブドリの伝記(青空文庫). https://www.aozora.gr.jp/cards/000081/card1924.html

[7] 吉本興業　岡本社長が２２日に会見へ. https://news.yahoo.co.jp/pickup/6330710

[8] 参院選のキーワード 1 位は「難しい」...SNS の反応を分析してみた. https://news.yahoo.co.jp/pickup/6330580

[9] Wikipedia バックプロパゲーション. https://ja.wikipedia.org/wiki/バックプロパゲーション

[10] 認定心理士の資格取得を目指す方へ. https://www.ouj.ac.jp/hp/purpose/sikaku/psychology/psychologist/