形態素解析における連濁および反復形オノマトペの自動認識

笹野 遼平　　黒橋 禎夫
東京大学大学院情報理工学系研究科　京都大学大学院情報学研究科
ryohei@nlp.kuee.kyoto-u.ac.jp　　kuro@i.kyoto-u.ac.jp

1　はじめに

自然言語処理の基礎技術として形態素解析はすでにある程度高い精度が実現されており、近年は形態素解析結果を用いた構文解析や格解析、照応解析などの、より高度な解析技術が研究されている。しかしながら、形態素解析の誤りは形態素解析の後に続くこれらの処理にも影響を与えるため、構文解析や格解析、照応解析の研究が進むにつれて、頑健な形態素解析の必要性が高まってきている。

現状の形態素解析の問題点の 1 つは未知語への対応の弱さである。基本的に形態素解析では、入力された文を辞書に登録されている語に分解するため、未知語を含む文が入力された場合、正しい解析が得られない場合が多い。

しかし、未知語の中には “掘りごたつ” における“ごたつ”のように連濁と呼ばれる現象により辞書に登録されている語が濁音化したものや、“ほいほい”、“グニョグニョ” のような反復を含むオノマトペなど、既存の知識から容易に推測できるものがある。

そこで本研究では、既存の辞書および簡単なルールを用いて、形態素解析において形態素列の候補を作成する際に、辞書に登録されている語の初頭の清音が濁音化したものや、反復表現も形態素の候補とすることにより、連濁および反復形オノマトペの認識を行い、頑健な形態素解析器の構築を目指す。

2　形態素解析

連濁および反復形オノマトペの自動認識の説明に必要となる形態素解析の手順を簡単に紹介する。形態素解析は通常以下のような手順で行われる。

手順 1 入力された文に対し、文中の各位置から始まる可能性のある形態素をすべてを検索する。

手順 2 形態素の候補を列挙したグラフ構造 (ラティス構造) を作成する。

手順 3 形態素同士の組み合わせの中から、文として最も確からしい形態素の並びを決定する。

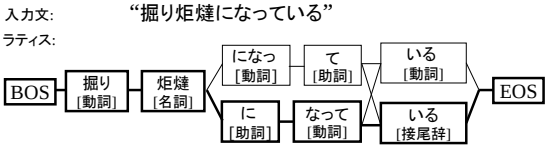

図 1: ラティス構造

例えば以下のような文が入力された場合、図 1 に示すようなラティスが作られ、最終的に太線で記されている組み合わせに決定される。

(1)　掘り炬燵になっている。

本研究では形態素解析器として JUMAN[3] を用いる。JUMAN では、**手順 3** において形態素の並びを決定する際、人手で設定した連接コストや単語生起コストから、それぞれの形態素列のコストを計算し、もっともコストの小さい形態素の並びに決定する。Chasen[4] や MeCab[2] など、機械学習を用いた形態素解析器においても基本的な解析の流れは同様である。

通常、**手順 1** では辞書に登録されている語が検索される。本研究では、辞書に登録されている語に加えて、辞書に登録されている語の初頭の清音が濁音化したものや反復表現も形態素の候補とし、これらの候補に対する**手順 3** におけるコストを適切に設定することにより、連濁および反復形オノマトペの認識を行う。

3　連濁の自動認識

3.1　連濁とは

連濁とは、日本語において複合語の後部要素の初頭にある清音が、濁音に変化する現象のことである。

(2)　掘りごたつになっている。

例えば、この文では “こたつ” が連濁現象により、初頭にある清音が濁音化し、“ごたつ” となっている。他にも “習慣づける” における “づける” など多くの例が存在する。連濁は複合語の後部要素初頭に清音があれば必ず起きるわけではなく、連濁を阻止する条件として以下のような性質が知られている。

語種による制約 連濁を起こすのは原則として和語であり、漢語や外来語は連濁を起こさない。

ライマンの法則 あらかじめ複合語の後部要素に濁音が含まれている場合、連濁は起こらない。

ただし、それぞれ例外が存在する。前者については、漢語であっても一般化・日常化した “会社” などの語は連濁を起こす。また、外来語についても少数ではあるが、“いろはがるた” の “かるた” のように連濁を起こすものも存在する。後者については、稀な例外として「なわばしご」などがある。

3.2 関連研究

連濁の認識に関する研究は、連濁規則に関する研究や、音声合成において、例文 (1) のように “炬燵” が漢字で表記されている場合に、濁音化して読むかどうかに関する研究は行われている [5]。しかし、濁音化した語が平仮名や片仮名で表記されている可能性のある文を入力として、連濁により濁音化した語を認識する研究は行われてこなかった。

従来、例文 (2) のような文の解析するための方法としては、“掘りごたつ” という語を辞書に登録するか、“ごたつ” という読みを辞書に登録する方法が一般的であった。しかし、連濁現象は特定の語のみ発生する現象ではないため、濁音化する可能性がある語すべてに濁音化した読みを登録するよりも、形態素解析において自動的に認識する方が望ましいと考えられる。

3.3 連濁の自動認識

本研究では、形態素解析において、平仮名または片仮名で表記される濁音化した表現を認識することを目的とする。濁音化した表現の認識は基本的に、形態素解析の**手順 1** において辞書を検索する際、濁音から始まる形態素の検索する場合に、その音を清音化した語から始まる語も検索することにより実現する。

例えば、“掘りごたつ” という入力があった場合は、“ご” から始まる形態素を検索する際に、“こ” から始まる語も検索し、結果として図 2 中で点線で示される “こたつ (名詞)” というパスが追加される。

ただし、“くろごま” 中の “ごま” のように、もともと濁音を含む形態素があればそちらを優先するために、濁音化した場合はもとの語 (ここでは “こま”) よりも単語生起コストを大きくする。単語生起コストをどのくらい大きくするかは、連濁の用例および連濁と誤って認識する可能性のある用例を用いて、品詞ごとに人手で決定した。

また、単純に清音化して検索できる形態素すべてを候補とすると、連濁により濁音化したと考えられないものまで候補としてしまうため、濁音化に次ような制約を加える。

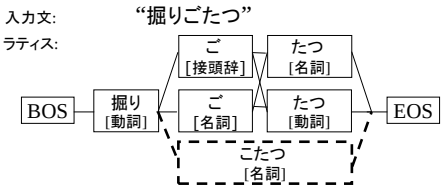

図 2: 連濁の認識

- 濁音化するのは普通名詞、動詞、形容詞のみ
- 濁音化するのは直前の形態素が名詞、動詞の連用形、名詞性接尾辞の場合のみ
- 1 文字からなる形態素は濁音化しない
- 代表表記が片仮名である形態素は濁音化しない
- もともと濁音を含む形態素は濁音化しない

まず、接続詞や助詞、人名、地名など、普通名詞、動詞、形容詞以外の形態素は連濁による影響を受けないものが多いと考えられるので除く。また、直前の語と複合語を形成している必要があるので、直前の形態素にも制限を加える。1 文字からなる形態素は “乳飲み子” における “子” のように、連濁により濁音化する場合も考えられるが、平仮名または片仮名で表記されることは稀であることから除く。

次に、**語種による制約**をどのように反映するかについて説明する。漢字を用いて表記されることが多い漢語が問題となることは少ない。x 語種による制約が必要となるのは、次のような 2 文を正しく解析したい場合である。

(3) a. 源氏ボタル (= 源氏 + ホタル)
b. 女性バイヤー (≠ 女性 + ハイヤー)

この例では、“ボタル” は連濁現象により “ホタル” が濁音化したものであるのに対し、“バイヤー” は “ハイヤー” が濁音化したものではない。この違いを認識するためには “ハイヤー” が外来語であることを認識し、さらに外来語は濁音化しないという制約が必要となる[1]。本研究では代表表記を用いることによりこの問題に対応する。JUMAN の辞書には、表記ゆれに対応するため、各語に対して代表表記というものが記されており、「ホタル」の場合「蛍」、「ハイヤー」の場合「ハイヤー」という代表表記が記述されている。代表表記が片仮名から始まる語は外来語であることが多いことから、代表表記が片仮名となる語は濁音化しないという制約を加える。

最後に、**ライマンの法則**を反映するため、もともと濁音を含む形態素は濁音化しないとう制約を加える。ただし、“はしご” のような例外については、ライマンの法則の例外であることを辞書中に記述することにより対応する。

[1] JUMAN の辞書には “ホタル”、“ハイヤー” は登録されているが、“ボタル”、“バイヤー” は登録されていない。

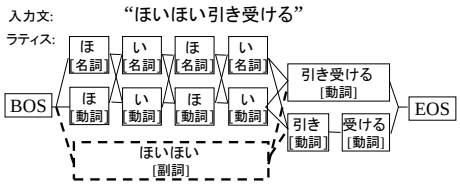

図 3: 反復形オノマトペの認識

4 反復形オノマトペの自動認識

4.1 反復形オノマトペとは

オノマトペとは、“ほいほい”、“ゆったり”、“グニョグニョ” などのような擬音語・擬声語・擬態語のことであり、日本語において頻繁に使われる。しかし、比較的自由に生成でき、新しい語が出現しやすいことから、辞書に載っていない語が多く、形態素解析の誤り原因の 1 つとなっている。オノマトペの多くは以下に挙げるようないくつかパターンで記述できる。

- ABAB (ほいほい、ガンガン)
- AっB り (うっかり、ひょっこり)
- A ん B り (しょんぼり、すんなり)
- AB ん AB ん (がたんがたん、くるんくるん)

このうち、“ABAB” や “AB ん AB ん” ような反復を含むオノマトペは数が多く、また、辞書に載っていない語も多い。そこで本研究では、オノマトペの自動認識の第 1 ステップとして、このような反復を含むオノマトペを反復形オノマトペと呼び、反復形オノマトペの自動認識を行う。

4.2 関連研究

オノマトペに関する研究としては、Web 上のコーパスからオノマトペの辞書を自動構築する奥村らの研究 [1] がある。奥村らは、典型的なオノマトペのパターンからオノマトペの候補語を生成した上で、その候補語をクエリとして Web 検索を行うことにより候補語を含む用例を獲得し、さらにその用例を分析することにより候補語がオノマトペであるかどうか判定し、オノマトペの辞書を構築している。

しかし、奥村らはパターンにマッチした文字列について、検索結果を用いたフィルタリングは行っているものの、オノマトペであるとする場合と、オノマトペでないとする場合、どちらが確からしいかの比較を行ってはいない。このため、新たにオノマトペと判断された語のうち、実際にオノマトペである語は約 55% しかない。また、獲得に失敗したオノマトペも多く、事前にオノマトペを獲得するよりも、形態素解析の時点で新たに認識できることが望ましいと考えられる。

表 1: 連濁の自動認識結果の分類

	A	**B**	**C**	**D**	合計
毎日新聞	85	47	14	0	146
Web	40	64	2	1	107
合計	125	111	16	1	253

4.3 反復形オノマトペの自動認識

形態素解析におけるオノマトペの自動認識は、形態素解析の **手順 1** において形態素の候補を検索する際に、2〜4 文字の繰り返し表現があれば形態素の候補に加えることにより行う。例えば、“ほいほい引き受ける” という入力があった場合は、最初の “ほ” から始まる形態素を検索する際に、“ほ (名詞)”、“ほ (動詞)” などに加えて、“ほいほい (副詞)” も候補し、図 3 中で点線で示すようなパスを追加する。

本研究では、“ほいほい” が 4 つの形態素に分割されるなど、 反復形オノマトペが必要以上に細かく分割されるのを防ぎ、全体で 1 形態素であることを認識するのを主な目的とする。このため、オノマトペの品詞としては、副詞、サ変名詞、形容詞などが考えられるが、本研究ではすべて副詞として処理する。

新たに副詞として生成した候補の単語生起コストは、反復形オノマトペの用例およびオノマトペ以外の反復表現の用例を用いて、これらを正しく解析できるように人手で決定した。基本的には、繰り返し表現の音数に比例し、濁音を含む場合、拗音を含む場合、カタカナで表記されている場合、それぞれコストが若干小さくなるように設定した。

5 実験

5.1 連濁の認識

毎日新聞 10 万文、Web 上のテキスト 3 万文を用いて連濁の自動認識を行った。それぞれ、146 個、107 個の表現が連濁によって濁音化していると認識された。解析結果を評価するため、連濁の認識を行わない場合と比較し、解析結果を以下の 4 つに分類した。

A 正しく解析できなかった表現を正しく解析
B もともと正しく分割・解析できていなかった表現を連濁により濁音化した表現と誤認識
C 未知語となっていた表現を連濁であると誤認識
D 正しく解析できていた表現を誤って解析

例えば、(4) のような文があった場合、濁音の自動認識を行わなかった場合は、“バシラン” は未知語として 1 つの形態素と認識されるが、濁音の自動認識を行うと “ハシ (箸)” と “ラン (蘭)” に分割されてしまうので、**C** に分類する。分類結果を表 1 に示す。

(4) フィリピン南部バシラン島。

表 2: 連濁の認識率

認識成功	認識失敗	再現率
83	17	83%

連濁の認識としては **A** 以外は誤りであるが、**B** に分類された表現はもともと正しく解析できていなかった表現なので、解析結果が悪化したとは言えない。また、**C** に分類される語はすべて固有表現中の表現であったので、固有表現認識を行うことにより間違いを修正できる可能性が考えられる。正しく解析できていた表現が誤って解析された **D** に分類されるものは 1 例のみであり、他の解析に悪影響を与えることなく連濁の認識を行うことができたと言える。

次に、連濁表現がどの程度の割合で認識できるかを調べた。与えられたテキストから連濁により濁音化した表現をすべて抜き出すことは困難であるため、以下の手順で実験に用いるコーパスを作成した。

1. 新聞記事中で出現回数が多い平仮名表記の形態素から、連濁により濁音化しうる形態素を選ぶ。
2. 選んだ形態素の初頭を濁音化した表現およびその片仮名表記を新聞記事から検索。
3. 連濁により濁音化した表現を含む文があればそれを収集する。

初頭の文字が単独で形態素となるかどうかにより認識の困難さが異なると考えられるので、濁音化が可能なカ行、サ行、タ行、ハ行の 20 音それぞれに対して、初頭がその音である 2 ないし 3 個の形態素を選び、それぞれの音ごとに 5 文ずつ連濁を含む文を収集した。

続いて、収集した合計 100 文の連濁を含む文を解析し、正しく認識できた連濁の数を調べた。結果を表 2 に示す。検出できなかった連濁 17 個のうち、直前の形態素が、名詞、動詞の連用形、名詞性接尾辞だと認識されなかったため検出できなかったものが 9 個、もともと濁音化した形態素が辞書中に存在しそちらが選ばれたものが 6 個、細かく分解する他の候補が選ばれたものが 2 個であった。

5.2 反復形オノマトペの認識

反復形オノマトペについては、繰り返し表現を自動抽出し、その中から反復形オノマトペとなっている表現を選び出すことにより、与えられたテキスト中の反復形オノマトペを抜き出すことが可能である。そこで、毎日新聞 10 万文、Web 上のテキスト 3 万文から、平仮名または片仮名の 2〜4 文字の繰り返し表現を抽出し、そのうち JUMAN の辞書に載っていないオノマトペとなっている表現に人手でオノマトペであるという情報を付与した上で、反復形オノマトペの自動認識を行い、解析結果の評価を行った。結果を表 3 に示す。

表 3: 反復形オノマトペの自動認識

	適合率	再現率
新聞記事	81.4(105/129)	92.9(105/113)
Web	62.5 (75/120)	87.2 (75/86)
合計	72.2(180/249)	90.5(180/199)

システムがオノマトペと認識した表現 249 個のうち 69 個は、実際にはオノマトペではなかった。しかし、このうち 20 個はもともと解析が誤っている表現であり、これらについては解析が悪化したとは言えない。また、43 個が次の例の “バーバー” のような固有表現であり、固有表現解析を行うことにより間違いを修正できる可能性がある。

(5) バーバー「弦楽のためのアダージョ」他。

もともと正しく解析できた表現を誤ってオノマトペと認識したのは、“ひまひま” や “タダメシタダメシ” など 6 例のみであり、他の解析に悪影響をほとんど与えることなく反復形オノマトペの認識を行うことができたと言える。

また、再現率も 90%を越えており、大多数のオノマトペの認識に成功していると言える。認識できなかったオノマトペ 19 個のうち、10 個が “鶴” の繰り返しと解析される “ツルツル” のように名詞の繰り返し、7 個が “割く” の繰り返しと解析される “さくさく” のように用言の繰り返しと解析され、2 個出現した “ワイワイ” が未知語と解析された。

6 おわりに

本研究では、形態素解析において形態素列の候補を作成する際に、辞書に登録されている語の初頭の清音が濁音化したものや、反復表現も形態素の候補とすることにより、連濁および反復形オノマトペの自動認識を行った。実験の結果、高い再現率、適合率で連濁および反復形オノマトペを認識できていることが確認できた。今後、固有表現抽出を用いることにより精度のさらなる向上を目指すこと、非反復形オノマトペの認識を行うことなどが考えられる。

参考文献

[1] 奥村敦史, 齋藤豪, 奥村学. Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築. 情報処理学会 自然言語処理研究会 2003-NL-154-10, pp. 63–70, 2003.

[2] 工藤拓. *Mecab*. http://mecab.sourceforge.jp/.

[3] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 5.1 使用説明書. 京都大学大学院 情報学研究科, 9 2005.

[4] 奈良先端科学技術大学院大学自然言語処理学講座松本研究室. 日本語形態素解析システム『茶筌』, 2003.

[5] 尾形エリカ, 林良子, 今泉敏, 平田直樹, 森浩一. 複合語の連濁・アクセント規則の認知機構. 日本音響学会聴覚研究会資料, 2000.