「読み」の概念を用いた漢字かな交じりのひらがな変換

林　聖人　　指導教員：山村　毅

1　はじめに

現代，我々は LINE や Twitter などの SNS が代表されるように文字を使ってコミュニケーションをとる．その中で誤りはつきものである，実際世の中には間違った文がたくさん存在するが，多大な情報を使う際，誤った文を含め少しでも多くのデータを利用したいと考えるのは自然なことである．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. また正しい文であっても小学生の教科書で見られるような，本来漢字で表記される単語がひらがなになっている文などにも弱いと言われている．

そこで本研究では, 堅牢な形態素解析システムの実現を目標に「読み」の概念を導入し漢字かな交じり文に対処する方法を提案する．

2　先行研究

2.1　概要と提案手法

著者ら [1] は現在の形態素解析器ではひらがな語に対する精度が悪いというところに着目した．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，漢字を用いる言葉が多い言葉は，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のみの文を正しく形態素解析できるか調査した．具体的には，形態素解析器として一般に用いられている MeCab の辞書 (IPADIC) にひらがな語を追加して形態素解析実験を行なった．ここで作成した辞書を用いた MeCab を「ひらがな MeCab」と呼ぶことにする．

2.2　評価実験と結果

辞書中の漢字の含まれている単語を，新たに単語として登録し辞書を拡張する．追加前の単語数は 392,126 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. これらの文を「そのまま（オリジナル）」または「すべてをひらがなに変換したもの（ひらがな）」に対して「ひらがな MeCab」で形態素解析した．また比較のために, 辞書を拡張しなかった元の MeCab を用いて, 同様に形態素解析を行なった．これらの結果得られた総計 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適合率・再現率を求めた．

表 1　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 1 に結果を示す．ひらがな語を追加することによりひらがな文に対するの解析精度を向上させることがわかったが，小学校中学年で見られるような漢字かな交じりの文には対応できない．

3　「読み」の概念

人間は誤りのある文や未知語などが含まれていても理解できる文へと柔軟に解釈し訂正を行なっている．その理由として意味の解釈を人間は常識のようなものを用いて行なっているからだと考えることができる

例えば，「わたしはは勉強をする」のような誤りの含む文を一般の形態素解析器が解析すると「はは」の部分が「母」というような結果が得られることがあるが人間の解釈では間違いを柔軟に判断し「私は勉強する」と理解する．このような人間の柔軟な解釈を「読み」と呼ぶことにする．

「読み」には音で判断する聴覚的「読み」,「言売」を「読」と判断する視覚的「読み」の二つの種類があると考える．

本研究では漢字かな交じり文を理解できるのは，読んで理解できるように「ひらがな」に変換しているからではないかと考えた．

4　提案手法

文をその読みであるひらがなに変換し，それを「ひらがな MeCab」で処理すれば漢字かな交じり文を処理できると考えられる．ひらがなへの変換は，漢字の読みの取得，読みの候補の絞り込みという手順で行う．

4.1　漢字の読みの取得

常用漢字表を利用して，全ての読み候補を生成する．例えば「安ぜんである」という文があり「安」という漢字は「アン，ヤス」という読みがあるので,「あんぜんである」と「やすぜんである」の二つの候補を生成する．

4.2　読み候補の絞り込み

「意味」が通じるように絞り込むのが本来だが，現在の自然言語処理技術では困難である．2 章で述べた「ひらがな MeCab」の時のように漢字かな交じりの単語を追加するのでは効率がとても悪いと考えられる．正しい「読み」を選択する上で必要なことは，文の前後からの判断からであると考える．先ほどの「安ぜんである」という文を見てみると,「あんぜん」か「やすぜん」となるが，ニュース記事や本などの文章をひらがなにした時,「やすぜん」という要素より「あんぜん」の要素の方が多く検出されると考えられる．文字列を区切る際にその要素の頻度が多いものの方がその文の正しい「読み」ではないかと考え，n-gram 頻度を利用して「妥当なもの」を選択することを提案する．

具体的な手順を以下に示す．

1. 全ての読みの候補を n-gram で分割する．
   例えば「あんぜんである」を 2-gram に分割すると，
   [’ あん’,’ んぜ’,’ ぜん’,’ んで’,’ であ’,’ ある’] のようになる．
2. 1 で作成した n-gram に対し，n-gram 辞書を用いて，その頻度を取得する．

3. 2 で取得した頻度を用いてその読みの妥当性を評価する．ここでは頻度の対数の和を用いる．

なお，手順 3 において頻度の対数の和を取るのは，頻度そのままの場合，一つの要素の値が極めて大きい数値の時，他の数値がどんな値でも極めて大きな数値が出た候補が選ばれることがあるため，これを軽減するために対数を導入する．

ここまでの様子を図示すると以下のようになる．

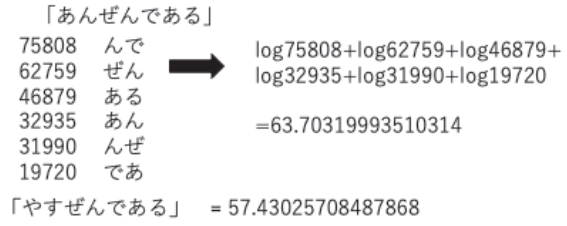

図 1　n-gram による妥当性の評価

5　評価実験

5.1　実験方法

4 章で述べた手法を実装し，その性能を評価した．

評価には毎日新聞 (2015 年) の文を漢字かな交じりに変換したものを 100 文（漢字 185 個）使用した．実験にあたり，2015 年の毎日新聞記事 1 年分を用いて，4 章で述べた n-gram 辞書を作成した．n-gram の n としてどんな値を用いるべきかを調べるため n を 2，3，4 と変えて，n-gram 辞書を作成し，それを用いてひらがな変換の正誤を調べた．図 2 に n-gram 辞書の一部を示す．

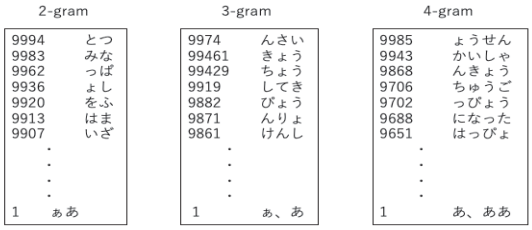

図 2　n-gram の辞書

5.2　実験結果

結果を表 2 に示す．この表で文正答率は，文全体として正しくひらがなに変換できている割合，漢字正答率は，個々の漢字を正しく変換できている割合である．

文正答率を見てみると 3-gram では 2-gram の時より 15 ポイント精度が向上し，4-gram では 33 ポイント向上していることがわかる．一方漢字正答率を見てみると 3-gram では 2-gram の時より約 9 ポイント精度が向上し，4-gram では約 16 ポイント向上していることがわかる．これらのことから漢字かな交じりの文章を n-gram と頻度を利用してひらがな変換する際，n-gram の n の値が大きくなるほど精度が向上することがわかる．特に文での正答率は 2-gram の時より 4-gram の時の方が極めて精度が向上していることがわかる．

表 2　ひらがな変換の評価結果

	2-gram	3-gram	4-gram
文正答率	38	53	71
漢字正答率	59.4	68.1	75.6

5.3　考察

図 3 に失敗例を示す．ここでは「こう」が正しい読みであるが，4-gram 辞書では「きょう」という要素が「こう」という要素より多く出現していたため，誤った「読み」が絞り込まれている．狭い範囲だけを見た場合，このように，偶発的に頻度の高い文字列が現れることがあるため，より正しく判定するには，n をもっと大きくする必要があるだろう．

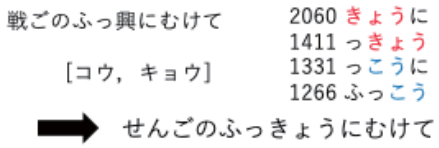

図 3　失敗例

なお，本手法で「夜露死苦おねがします」のようにあて文字の含まれている文を処理した場合（4-gram を使用），「よろしくおねがいします」と正しく変換された結果が得られた．

6　まとめ

文をその「読み」であるひらがなに変換し，それを「ひらがな MeCab」で処理すれば漢字かな混じり文を処理できると考え，本研究では n-gram を利用して読み候補の絞り込みを行い正しくひらがな変換されているか評価実験を行なった．

n の値を上げるにつれてひらがな変換の正答率は上がった．特に文単位で見た時に精度が極めて向上していることがわかった．しかし，頻度の高い読み候補の場合その「読み」が正しくなくても絞り込まれてしまうことがあった．一方これらとは別に，「夜露死苦おねがいします」などのあて字が含めれている文にも正しい読みを絞り込むことに成功した．

今後は，本研究の n-gram と頻度を利用した絞り込みの手法と形態素解析器を組み合わせることで正しい文でも漢字かな交じりの文章にも対応できるシステム開発．および n-gram の n の値および辞書データとの関係性についての調査．その他の自然言語処理の誤った表現や雑音についての対処を考えるなどの課題が挙げられる．

参考文献

[1] 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”, 電気・電子・情報関係学会東海支部連合大会講演論文集, C1-2, 名古屋大学, 2017