漢字の読み誤りの自動生成における候補生成能力の評価

ボラ　サワシュ †　　林 良彦 ‡

†‡大阪大学言語文化研究科 〒560-0043 大阪府豊中市待兼山町 1-8

E-mail: †bsavas@gs.lang.osaka-u.ac.jp, ‡hayashi@lang.osaka-u.ac.jp

1. はじめに

近年のインターネットの普及は様々な分野に影響を与えてきたが，語学教育の分野にも大きな変化をもたらした．現在，インターネット上において様々な言語学習のツールが提供されている．日本語教育においても，外国人の日本語習得に大きな問題となっている漢字学習のためのツールや漢字テストが公開されている[9]．しかし，学習者にとって自分のレベルに合った漢字の問題集を探すのは非常に困難である．また，日本語教師にとっては，漢字の読み問題の作成方法についての明示的な基準がないため，問題作成の作業には時間と労力を要する．

そこで我々は,日本語教師などの専門家の手を煩わせることなく，インターネット上の生のテキストからそのまま学習者のレベルに合ったテスト問題を自動生成し，漢字の読み方を学習できるようなシステムを開発することにした．本稿では，テスト問題においてユーザに選択肢として提示するための読み誤りの自動生成手法を提案し，特に，評価実験の結果からその候補生成能力と今後の課題を検討する．

2. 関連研究

漢字の読みに関する問題を自動生成しようとする研究は比較的少ないが，「おさる」と呼ばれる日本語練習問題自動生成ツール＆問題データベースが報告されている[4]．「おさる」で提案されている読み問題の自動生成は，「似た読み」，「別々読み」，「類義語」の 3 つのレベルにより行われる．ここで，「似た読み」の生成は，おもに音韻に関する 7 つのルールにより行われる[3]．また，「別々読み」の生成においては，漢字の他の読みに置き換える（強引→つよひき）手法や，類字語のルールにより同じ漢字を含む別の言葉に置き換える（会議→集会「シュウカイ」）手法が提案されている．しかし，「おさる」の読み誤り生成能力の評価は今後の課題として残されている．また，自動生成の精度を高めるためにコーパスを使用しているが，著作権の問題でシステムが公開されていない現状である．

一方，日本語学習者を対象とする辞書検索支援ツール[1]として「FOKS」が提案されている[5]．FOKS では，「浴衣」や「山車」などの一見正しい読み方が分かりづらい熟語の意味を調べるときに，ユーザが「よくい」，「やましゃ」というように，音訓取り混ぜた読み方で検索することができる．入力された読み仮名と熟語を構成する漢字をもとに生成される読み候補の類似性をローマ字表記ベースで求めることにより検索を実現している．その際さらに，日本語学習者によって間違えて入力されやすい例や，タイプミスなどのデータを基に候補を絞り込む手法が採用されている．

これらの関連研究に対し，我々は，既存の言語資源を基に独自の辞書・ルールを生成することにより，無償で誰でも使えるようなソフトウェアを目指して研究を進めている．特にこれまでは，テスト問題においてユーザに選択肢として提示するための十分な数の読み誤りを生成する方式について検討してきた．

3. 読み誤りの自動生成

3.1. 自動生成の処理手順

自動生成の処理手順と使用する辞書を図 1 に示す. まず，ユーザにより入力，あるいは，選択されたテキスト中に含まれている文を単語単位に分割する必要がある．このために，日本語の形態素解析を適用する．本システムは Java 言語で開発しているため，MeCab[14]の Java への移植版であるオープンソースの形態素解析器 SEN[13]を用いている．

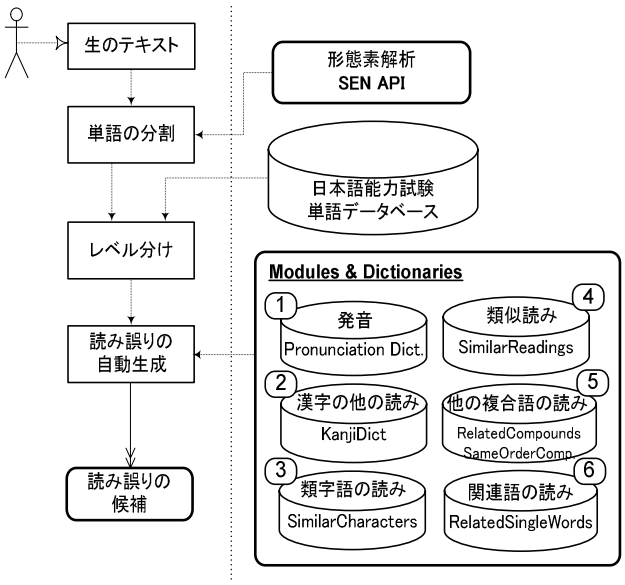

図 1　自動生成の処理手順と使用する辞書

形態素解析器で抽出された形態素を下記の節で詳述する単語の読みのパターンによって分類する．次に，読みの分類処理によって得られた単語リストをユーザの日本語のレベル[2]に合ったものに限定する．各単語のレベル判定はネット上で無償で公開されている日本語能力試験出題基準の語彙リスト（1 級～4 級）[8]を使用する．最後に，ユーザのレベル別に適合した単語を対象に読み誤りの候補を自動生

[1] http://www.foks.info/

[2] 日本語能力試験のレベル（1 級～4 級）

成する．ここでは，読み，形状，意味に基づく３つのモジュールとそれぞれに関連した辞書を利用する．

3.2. 単語の分類

形態素解析によりテキストから抽出された形態素の中には漢字を含まないものも存在するが，フィルタリング[3]の処理で漢字を含む単語のみを残す．次に，各単語を，送り仮名を含むものと含まないもの，熟語と単漢字に分類する．読み誤りの生成においては，この分類ごとにあらかじめ用意した，漢字の読み，形状，意味の類似性をパターン化したルールを適用する．

図 2 は，上記の単語の分類を示す．本稿においては「送り仮名のない単語」のみを扱う．

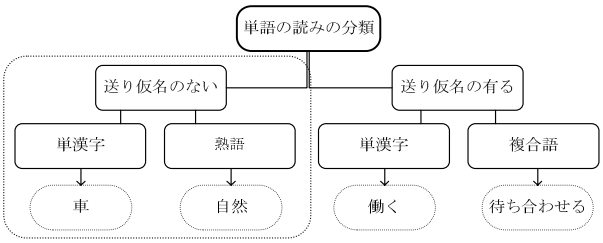

図２　単語の分類

3.3. 読み誤りのパターン化

自動生成の対象となる単語から誤りを含む読みを単語の読み，形状，意味の類似性をパターン化した下記のようなルールに基づいて生成する．これらのパターンを適応することによって多数の読みを生成することができる．本研究の現段階においては，有効な範囲でなるべく多くの読み誤りの候補を生成することを目指すが，実際の学習場面においてテスト問題の作成に適用する際には，適切な絞込みにより適切な数(3~4 個)の選択肢に絞り込む必要がある．

以下に誤り生成の 3 つのパターンについて説明する．

- 読みベース:
  - 発音の置換　(図 1 辞書#1)
    - 自然　「シ・ゼン」
      シ（清音）→ジ（濁音）＝ジゼン
  - 50 音順縦移動置換 (辞書不要)
    - 自然　「シ・ゼン」
      シ（い行）→キ　＝キゼン
  - 他の熟語の読みへの置換　(図 1 辞書#5)
    - 自然　「シ・ゼン」
      「自」を含む他の熟語 →自由　＝ジユウ
  - 類似発音の熟語の読みへの置換
    (図 1 辞書#4)
    - 自然「シゼン」
      = シゲン
  - 漢字の他の読みへの置換 (図 1 辞書#2)
    - 自然　「シ・ゼン」
      「然」の読み（ゼン，ネン）＝シネン
- 形状ベース
  - 類字語の読みへの置換 (図 1 辞書#3)
    - 自然　「シ・ゼン」
      自→目→目然　＝モクゼン
- 意味ベース
  - 関連語の読みへの置換 (図 1 辞書#6)
    - 北「キタ」
      北の類字語　→南，東，西　＝ミナミ

読みベースのルールを適用するためには，単語中のどの文字に対してルールを適応するかを定める必要がある．すなわち，単語を構成する各単語と読みの部分とを対応付ける必要がある(3.4 節)．また，他の熟語の読みへの変換や，他の読みへの変換を行うためには，これらの情報を参照するための辞書が必要となる

また，形状ベースのルールを適用するためには，ある漢字に対して形状が類似した漢字(類字語)を選択する必要があるが，実行時に形状のマッチング処理を行うことは困難であるため，あらかじめ類字語を収録した辞書を準備しておくことが望まれる．さらに，意味ベースのルールを適用するためにも，ある観点において意味が類似した漢字を抽出できる必要がある．ここで，意味の類似としては，外国語学習の観点に根ざした観点が考えうる．すなわち，いわゆる類義・反意といった関係だけでなく，より制約の弱い広い意味での連想関係に近い関係を扱う必要があると考えられる．

本研究では，図 1 に示すように各パターンに対応する専用の辞書を構築している．各辞書は必要に応じて拡張，編集可能な形式で設計されている．

3.4. 単語の読みの分割処理

単語正解の読みから読み誤りを生成する前に，まず単語の読みのどの部分を置き換えるか決めなければいけない[4]．例えば，「シゼン（自然）」の場合，単語の置き換える部分を「シ」に決定し，該当するパターンを適応することによって読み誤りの候補を生成することができる．

しかし，形態素解析を行うだけでは「シゼン」に対してどれが「自」の読みで，どれが「然」の読みであるか分からない．そこで，単語をそれぞれの読みに分割する処理を行う[5]．このために，KanjiDic[6] [10]から抽出した 6,356 個の漢字を使って約 500 個の読みからのみなるリストを作成し，これを利用する．

4. 辞書とその作成方法

本システムでは第 3 節で説明したようなルールに対応し

[3] 熟語・漢字仮名混じりの単語など．

[4] 後述の評価実験においては，単語の構成文字の全ての漢字に対して置換した候補を生成している．

[5] この処理の詳細に関しては[1]を参照されたい．

て図 1 に示したような複数の辞書を用意している．以下に各辞書の内容と作成方法を説明する．

読みベースのパターンに基づく熟語の読み誤りの生成には，RelatedCompounds，および， SameOrderCompounds (図 1 辞書#5) とよぶ 2 種類の辞書を使用する．

RelatedCompounds 辞書のエントリー例を図 4 に示す．この辞書には，読み誤り候補を生成する熟語と先頭の漢字が同一である熟語を Edict[7][11]から抽出してリストする．

郵	郵券/ユウケン, 郵政/ユウセイ, 郵政省/ユウセイショウ, 郵船/ユウセン, …

図 4　RelatedCompounds のエントリー例

これに対して，SameOrderCompounds 辞書は，単語の二文字目以降の各漢字に注目した読み誤り候補を保持する．より具体的には，以下の条件を満たす熟語を抽出する．

- 熟語の文字数が等しい
- 着目している同じ文字位置の漢字が等しい
- 熟語全体をローマ字表記したときの編集距離[12]が閾値(現在 3)以下である

図 5 にこの辞書のエントリー例を示す. 例えば,「郵便局」という熟語に対して，二文字以降である「**便**」と「**局**」が熟語の同じ文字位置(二文字目，三文字目)に現れる「郵**便**船」や「事務**局**」のような同文字数(文字数=3)の熟語が Edict から抽出されている．

便	郵便局/ユウビンキョク, 不便利/フベンリ, 男便所/オトコベンジョ…

図 5　SameOrderCompounds のエントリー例

読みに基づく誤りの自動生成のために，さらに SimilarReadings（類似読み）(図 1 辞書#4)とよぶもう一つの辞書を用意する．この辞書の作成にも Edict を用いた．Edict 中に収録されている各単語の読みをローマ字表記に変換し，辞書中の他の単語の読みのローマ字表記との間で編集距離を求め，編集距離が閾値以下である単語を抽出することにより作成する．今回は，編集距離の閾値を 1 に設定し，45058 語について平均 7 個の読みを抽出し，辞書化した．

読みベースの漢字の他の読みへの置換による読み誤りの生成には，KanjiDic (図 1　辞書#2)を使用し，発音の置換による読み誤りの生成には PronunciationDictionary(図 1 辞書#1)とよぶ辞書を用意している．この辞書には，3.4 節で述べた仮名リストをもとに各エントリーの読みを可能な限り「濁音」,「半濁音」,「長音」,「促音」,「拗音」のどれかに変換したものを収録する. 図 6 に ProdunciationDictionary のエントリーの例を示す.

[6] いわゆる JIS 第 1 水準，第 2 水準の単漢字を収録.

[7] 約 11 万のエントリーを持つ和英対訳辞書.

[9] [1]の報告の後に SameOrderCompounds 辞書を導入し，これにより再現率は平均約 10%程度上昇した.

キ	ギ, キャ, キュ, キョ, キイ
キク	キグ, ギク, ヒク
オツ	オス, オッ
キョ	ギョ, キュ, キャ, キヨ, キョウ

図 6　PronunciationDictionary のエントリー例

意味ベースのパターンに関しては，RelatedSingleWords (図 1 辞書#6) とよぶ辞書を使用する．この辞書は日本語能力試験出題基準の語彙リスト 3 級と 4 級に出題した単語に対して関連語を人手で収録したものである．図 7 に関連語辞書のエントリー例を示す.

また，形状ベースのパターンに関しては，日本語能力試験出題基準の語彙リスト（1 級～4 級）に含まれている単漢字を対象に手動で作成した SimilarCharacters(図 1 辞書#3)とよぶ類字語辞書を使用する．図 8 に類似語辞書エントリー例を示す.

父	チチ	ハハ, カレ, オトウト
頭	アタマ	カオ, ミミ, ハナ

図 7　関連語の辞書エントリー例

婬	娃,垠	人	入,八
西	四,両	海	悔,侮

図 8　類字語の辞書エントリー例

5. 評価実験

5.1. 評価実験

テスト問題の作成においては，十分な選択肢のバリエーションを生成できることがまずは重要であるため，提案手法の読み誤りの生成能力についての評価実験を行った．今回は, 特に読み誤りの候補を絞り込むことはせず(読みに基づく誤りの編集距離による絞込みを除く), 生成した読み誤り候補のどの程度が正解データにおける読み誤りの選択肢にも出現していたかという再現率を調べた.

今回，正解データとしたのは，日本語能力試験の過去問題集[6][7]から抽出した読み誤りの選択肢である. 本研究においては，送り仮名を含まないケースを対象にしたため，この基準に合致する試験問題を， 1 級から 200 問，2 級から 200 問，3 級から 110 問，4 級から 98 問抽出し，これらの問題における誤りの選択肢(各問につき 3 つ)を正解集合とした.

5.2. 評価結果

1 級から 4 級までの各問題についての実験結果を表 1[9]に示す. 表 1 において,「読み誤り数」とは全ての選択肢の数から正解(各問について一つ)を除いたものである.「一致数」とは，これらの選択肢のなかに含まれる本プログラムが生成した候補の数であり, 一致数と読み誤り数の比が「再現率」である.

5.3. 考察

表 1 に示すように，再現率は比較的低いレベルにとどまっている．しかしながら，漢字学習者の能力をテストする上で選択肢としてどのようなものが適切かということには確固たる基準は存在しない．したがって，過去問題における再現率のみによって提案手法の評価を行うことは不適切であると考えられる．すなわち，過去問題には出現しなかったかもしれないが，同等の適切さを持つ選択肢も多く存在するものと考えられ，本手法によりこれらの選択肢が生成できている可能性もあるのである．ただし，級ごとの再現率の傾向をみると，級が上がるごとに再現率が下がる(4 級を除く)傾向がうかがわれる．これは，上級になるほど，本方式で想定していないようなより多様なバリエーションを持つ選択肢が使われている可能性を示しており，さらなる誤り生成パターンの追求が必要と考えられる．

表 1　読み誤りの再現率

級	読み誤り数	一致数	再現率
1 級	600	239	39.7%
2 級	600	243	40.5%
3 級	327	132	41.1%
4 級	282	110	39.0%

6. まとめと今後の課題

本研究では指定された日本語テキストからユーザのレベルに合った単語を抽出し，これらの単語に対して選択式の読みテストの問題を生成するために読み誤りを生成する方式を提案した．提案手法は，読み，形状，意味という 3 つのレベルのルールに基づいている．読みベースのルールは，KanjiDic と Edict といった汎用の辞書をベースに構築した．形状ベースのルールのための類字語辞書，意味ベースのルールのための関連語辞書については，人手により小規模な辞書を構築し，その有効性を検証したが，今後はその大規模化や機械的な生成手法の検討が必要である．

日本語能力試験における過去問題を正解データとした予備的な評価実験の結果からは，提案手法の基本的な有効性を確認することができたが，以下のような項目を検討することにより，さらに読み誤りの生成能力の向上を検討していきたい．

- 能力試験のレベルが低いほど，「関連語の読みへの置換」が重要になってくるため，特に一文字の単語の場合に使用する関連語の辞書を更に充実させる必要がある．
- 形状ベースの類字語の辞書も網羅性が低いため，この辞書を充実させることが必要である．入手可能な文字認識エンジン（巴[2]など）の識別辞書から自動処理により類字語を計算し，KanjiDic に含まれている全ての単漢字に対して，類字語の辞書を構築することを検討したい．
- 現在の候補生成は，熟語単位ではなく単語単位で行っているため，例えば，「一週間」に対して「イッチュウカン」のような誤り候補をすることができていない．数量表現を含む場合など熟語の単位を超えた複合語に対する候補生成が必要である．

ところで，実際の学習ツールへの適用においては，生成能力だけでなく，学習者へ実際に提示する読み誤り候補をいかに適切に絞り込むかという問題も重要になってくる．各ルールを用いて候補を生成する際に何らかの評価値を与えることができれば，これを基に選択を行うことができる．あるいは，各種の読み誤り候補辞書を構築する際に，そのエントリーを絞り込んでおくことも考えられる．これらのためには，コーパスなどを用いた使用頻度に関する統計的なデータが有用であろう．また，テスト問題として提示される選択肢集合の集合としてみた際の適切性についても何らかの評価尺度を設定し，これに基づき，学習効果を考慮した選択肢集合の提示ができることが望まれる．このためには，各学習者の特性を適切にプロファイル化し，さらには，習熟度や学習履歴をも考慮することが必要となろう．

文　献

［1］Bora Savas, 林 良彦, 漢字の自律学習のための読み誤りの自動生成, 2006, 電子情報通信学会技術研究報告, TL―31～40

[2] 手書き文字認識エンジン Tomoe, http://tomoe.sourceforge.jp/

［3］ネワパル ブッシュハン ラザ, 森川聡, 松本祐治, 日本語学習支援システム「おさる」における漢字読み問題生成方法, 2002, 言語処理学会第８回年次大会

[4] Neupane Bhooshan Raj, Data Classification and Question Generation in Osaru “A web based Japanese language acquisition support system”, NAIST Master Thesis, 2003

[5] Slaven Bilac, Timothy Baldwin, Hozumi Tanaka, Construction of a Japanese learner-friendly dictionary interface, 言語処理学会第８回年次大会発表論文集, 460-463, 2002.

[6] かたくり日本語教師会著, 石井怜子（著者代表）,完全マスター漢字日本語能力試験 1 級レベル, スリーエーネットワーク, 東京, 2003 年

[7] かたくり日本語教師会著, 石井怜子（著者代表）,完全マスター漢字日本語能力試験 2 級レベル, スリーエーネットワーク, 東京, 2003 年

[8] Japanese language proficiency test vocabulary list, http://www.thbz.org/kanjimots/jlpt.php3

[9] Jim Breen’s Japanese page, http://www.csse.monash.edu.au/~jwb/japanese.html#links_software

[10] Jim Breen, KANJIDIC project, Faculty of Information Technology, Monash University, http://www.csse.monash.edu.au/~jwb/kanjidic.html

[11] Jim Breen, JMdict/EDICT project, Faculty of Information Technology, Monash University, http://www.csse.monash.edu.au/~jwb/edict.html

[12] The Levenshtein Algorithm, http://www.levenshtein.net/

[13] SEN (形態素解析器), https://sen.dev.java.net/

[14] 工藤拓, MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.jp/