

視覚障がい者用仮名漢字変換プログラムのための 説明語辞書の自動作成

後藤 和真 西田 昌史 綱川 隆司 西村 雅史

静岡大学情報学部

1. はじめに

視覚障がい者はコンピュータを利用する際、スクリーンリーダーという音声読み上げソフトを用いて画面を認識し、操作する。仮名漢字変換を行う場合、候補漢字の説明を聞くことで正しい漢字を選択する。現在広く使われている説明方法である詳細読みは、漢字をその漢字を含む単語と、音読み、訓読みで説明する。しかし渡辺ら[1]により、詳細読みは説明に使用者の語彙にない言葉や同音異義語が使われた場合想起しづらいことが指摘されている。

荒田ら[2]は漢字をその部品と位置関係で説明する構成読みを提案し、漢字の形を正確に想起できることを示したが、構成読みは複雑な字形の漢字には使いづらいという欠点がある。

また宮村[3]は視覚障がい者に漢字との関わりについてのアンケートを行い、多くの視覚障がい者が漢字の意味、読み、例文、字形の知識を必要としていることを明らかにした。

これらを受け西田ら[4]は、漢字の詳細読み、意味、例文、構成読み、読みといった情報をひとまとめに提示するシステムを提案し、システムにより正しく漢字を連想できることを示した。しかし検索対象となる漢字の説明は全て手動で登録されており、全て人手で辞書を作成するのは非常に負担が大きい。そこで Web 上から単語の情報を自動で収集し、自動的に辞書を作成する手法を提案する。

2. 提案システム

Web スクレイピングにより漢字の意味、例文、読みといった情報を自動で検索、抽出、加工し、辞書を作成する。スクレイピングには自動でブラウザを操作するためのライブラリである Selenium を使用する。Web ページのソースコードから操作したい箇所を指定し、値の代入やクリック、文字列の抽出などを自動で行う。なお入力 Excel ファイルに検索対象とする漢字二字からなる単語を書き込んでおく。

意味、読みはまず三省堂 Web Dictionary[5]から検索、抽出する。三省堂 Web Dictionary は単語の意味が簡潔にわかりやすく記されているため、音声の説明で単語を識別する本システムに適していると考えられる。検索窓に入力漢字を順番に入れ検索し、意味、読みを抽出していく。複数の意味がある場合、一般的に用いられる順に記載されているため、一つ目のものを採用する。また意味だけを登録できるように抽出した文字列を加工する。三省堂 Web Dictionary で検索できなかった単語は goo 辞書[6]で再検索する。また三省堂 Web Dictionary で取得した意味の説明が二字熟語のみであった場合、同音異義語がある場合やあまりなじみのない単語が使われる場合が多くわかりづらいことから、同様に再検索する。goo 辞書でも検索窓に漢字を順番に入れ検索していく。三省堂 Web Dictionary とは異なり goo 辞書では複数の読みが検索結果に出てくるため、全て辞書に登録する。

例文は goo 辞書から検索、抽出する。goo 辞書では単語の意味の説明内に簡単な例文を含むものが多いため、そこから抽出する。意味のとき同様入力漢字を順に検索し、例文を抽出していく。しかし goo 辞書では歴史書から抽出された例文も含まれておりその場合説明がわかりにくいものが多い。よって例文が歴史書からの抽出である場合と検索結果がなかった場合、自動で検出し用例.jp[7]で再検索する。用例.jp は単語を検索すると膨大な量の書籍からその単語が使われている箇所が抜き出され表示される。検索結果のうち一番上のものから一文を抽出する。

日本漢字能力検定での出題級数が 10 級から 2 級までの漢字 2 字からなる単語 330 単語に対し単語情報の自動収集を行った。抽出できた割合と単語数を表 1 に示す。

表 1 抽出できた割合と単語数

	意味	例文	読み
割合(%)	99.7	95.2	99.7
単語数	329	314	329

ほとんどの単語に対して単語の情報を抽出できていることを確認した。

3. 評価実験

3.1. 実験条件

晴眼者の大学生 12 名に対し、従来手法である手動で作成した辞書と提案手法である自動で作成した辞書を用いた長野のシステムを使用し、2通りの説明方法で評価実験を実施した。実験では問題文を提示し、該当箇所の単語に対して候補の中から正しい漢字を選択してもらった方法で行った。こちらが用意した問題文で使用されている漢字の同音異義語は 5 種用意した。この 5 種の候補の中から正しい漢字を選択してもらった。問題文の例を図 1 に示す。今回は自動収集できなかった単語を除き、問題文を 50 問用意した。

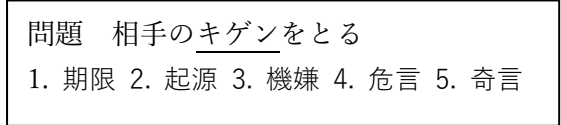


図 1 問題文と候補漢字の例

最初に練習問題によりシステムについての理解を得てもらった。パソコンの画面を隠した状態でシステムを使用し、音声を聞き問題に解答してもらい正答率と解答時間を記録し、提案手法と従来手法との正答率と解答時間の違いを比較した。実験終了後に漢字の連想しやすさなどについてのアンケート調査を行った。

3.2. 実験結果

表 2 に被験者ごとの正答率と 1 問あたりの解答時間を示す。なお被験者 A から F までは従来システム、提案システムの順で、被験者 G から L までは提案システム、従来システムの順で評価を行った。

表 2：正答率と 1 問あたりの解答時間

被験者	従来手法	提案手法	被験者	従来手法	提案手法
A (従来→提案)	68% (31秒)	72% (28秒)	G (提案→従来)	76% (46秒)	80% (41秒)
B (従来→提案)	86% (43秒)	94% (48秒)	H (提案→従来)	84% (29秒)	82% (31秒)
C (従来→提案)	72% (34秒)	78% (30秒)	I (提案→従来)	66% (25秒)	74% (27秒)
D (従来→提案)	72% (32秒)	82% (30秒)	J (提案→従来)	68% (22秒)	60% (33秒)
E (従来→提案)	76% (33秒)	82% (27秒)	K (提案→従来)	74% (28秒)	70% (35秒)
F (従来→提案)	80% (46秒)	84% (37秒)	L (提案→従来)	88% (32秒)	80% (41秒)

正答率と 1 問あたりの解答時間の平均値は、従来手法が 76% (33 秒)、提案手法が 78% (34 秒)であった。提案手法は人手で辞書を作成した従来手法と比べて、正答率と解答時間にほとんど差がなかったことから、辞書を自動で作成しても差し支えないことが明らかになった。

またアンケートの結果、12 名中 10 名が従来手法と提案

手法で分かりやすさは変わらないと回答した。よって提案手法と従来手法で分かりやすさにあまり差はないと考えられる。

また 12 名中 1 名は提案手法の方がわかりやすいと回答した。その理由は提案手法の方が意味の説明が短くてわかりやすいからだと答えている。簡潔にまとめられた三省堂 Web Dictionary から意味を抽出したことによると考えられる。

さらに 1 名は従来手法の方がややわかりやすいと回答した。その被験者は想起に例文をよく利用しており、例文が従来手法の方がわかりやすいと答えている。提案手法の例文には用例.jp から抽出した文が含まれており、それらの文は長く、実際に書籍で使われている形であるため辞書のための例文としてはわかりやすさが劣る。しかしこの被験者は提案手法の方が正答率が高く、回答時間が長いことから、じっくり聞いて考えれば正しい漢字を想起することはできると考えられる。

以上の結果から、提案手法による辞書の自動作成が有効であることがわかった。

4. おわりに

本研究では、視覚障がい者のための仮名漢字変換支援システムで使用する意味、例文といった情報を自動で収集する手法を提案した。評価実験の結果、従来の手動で作成した辞書とわかりやすさにあまり差がない辞書を、自動で作成することができた。しかし例文に長くわかりにくいものが含まれるという欠点も見つかった。例文のわかりやすさの向上と、抽出できた単語数の割合の向上が今後の課題である。

参考文献

- [1] 渡辺哲也, 渡辺文治, 藤沼輝好, 大杉成喜, 澤田真弓, 鎌田一雄:スクリーンリーダーの詳細読みの理解に影響する要因の検討-構成の分類と児童を対象とした漢字想起実験, 電子情報通信学会論文誌, D-I Vol. J88-D-I, No.4, pp.881-899, 2005.
- [2] 荒田龍朗, 岸和樹, 山口俊光, 渡辺哲也: 視覚障害者向け漢字の構成読みの開発とその評価, 電子情報通信学会論文誌 D Vol. J96-D No. 11 pp. 2746-2754 2013.
- [3] 視覚障害者と漢字 <http://kanji.eng.niigata-u.ac.jp/>
- [4] 西田昌史, 長野亮, 綱川隆司, 西村雅史:視覚障がい者を対象とした仮名漢字変換支援システムの構築, 第 18 回情報科学技術フォーラム(FIT), K-033, pp.377-378, Sep. 2019.
- [5] 三省堂 Web Dictionary <https://www.sanseido.biz/>
- [6] goo 辞書 <https://dictionary.goo.ne.jp/>
- [7] 用例.jp <http://yourei.jp/>