7ZJ-01

Word2vec を用いた詳細読みへの意味情報の付与

西山 千尋　　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、スクリーンリーダという音声読み上げソフトを用いて操作を行う。コンピュータに文章を入力し、仮名漢字変換を行う場合も、読み上げられた候補漢字の説明を聞くことで漢字を選択する。現在広く使用される詳細読みという説明方法では、漢字をその漢字を含む別の単語と、音読み、訓読みで説明する。しかし、渡辺ら[1]により、詳細読みは使用者の語彙にない言葉や同音異義語が説明に使われた場合、漢字を想起しづらくなるという欠点があることが指摘されている。

詳細読みの欠点を改善するため、西田ら[2]は変換候補の漢字の類義語といった意味情報で説明する（例えば、変換候補の漢字「乗法」を類義語「掛け算」で説明する）手法を提案し、使用者の負担が少なくより正確な仮名漢字変換が可能であることを示した。しかしこの方法にも、機械と器械といった似た意味の同音異義語の区別が難しい点や、名前や地名などの意味を持たない固有名詞には使用できない点などの欠点が存在する。また、意味情報は人手で抽出する必要があり手法の問題点となっていた。

そこで本研究では、詳細読みの説明語に対して Word2vec を用いた自動的な類義語の抽出を行い、意味情報として付与する手法を提案する。提案手法の有効性を確認するため、従来の詳細読みの説明と、提案手法による詳細読みの説明を聞き、漢字を連想して書き取る実験を行う。

2.従来手法

詳細読みでは、対象となる単語の漢字 1 文字ずつに対して、その漢字を含む別の単語、音読み、訓読みを用いて説明を行う。例として、「意思」という単語に対しては「意見のイ　思想のシ、思う」という説明を行う。このとき、「いけん」という単語には、違憲や異見といった同音異義語が存在し、音声を用いた情報だけでは、変換したい漢字の特定が困難である。このように、視覚障がい者が仮名漢字変換を行う場合、詳細読みの説明語に含まれる同音異義語を区別することが 1 つの課題である。

3.提案手法

以上の従来手法の問題点を解決するために、詳細読みの説明語に対して意味情報を付与する手法を提案する。提案手法でも、詳細読みと同様に、対象となる単語の漢字 1 文字ずつに対して、その漢字を含む別の単語、音読み、訓読みを用いて説明を行うが、その漢字を含む別の単語で説明する場合に、その単語の前に意味情報を加えて説明を行う。例として、「意思」という単語に対しては「見解の意見のイ考え方の思想のシ、思う」という説明を行う。ここでは、「意見」に対する意味情報として「見解」、「思想」に対する意味情報として「考え方」を付け加えている。

また、意味情報については Word2vec を用いて類義語を抽出する。Word2vec とは Mikolov ら[3]によって提案された単語の意味をベクトルで表現する手法である。ベクトル表現を行うことで、単語の意味について演算処理が可能となり、ある単語とある単語の間の意味の近さを計算することができる。詳細読みの説明語に対して、これを用いることによって、意味情報を付与したい単語と意味情報となる候補単語との意味的な近さを数値として算出し、類似度上位単語から意味情報となる単語を抽出する。例として、「意思」の詳細読み説明語となる「意見」の類似度上位単語を以下の表 1 として示す。

表 1 Word2vec を用いて抽出した「意見」の類似度上位 10 単語

単語	類似度
見解	0.820
見方	0.675
異論	0.636
言い分	0.626
考え方	0.621
賛否	0.616
論調	0.599
意見書	0.592
議論	0.592
結論	0.589

なお、Word2vec で抽出した類義語から最適な意味情報を決定する作業は人手で行った。そのとき、説明語と同じ読みを持つ単語、説明語そのものを含む単語を意味情報の候補から除くという基準を設けた。これは、「異論の意見のイ」という説明から 2 種類のイという漢字を想起できること、「意見書の意見のイ」という説明から漢字の特定が困難であることが懸念されるためである。

Semantic Information Indexing to Shosaiyomi Using Word2vec
Chihiro Nishiyama, Masafumi Nishida,
Takashi Tsunakawa, Masafumi Nishimura, Shizuoka University

4.評価実験

晴眼者の成人 10 名に漢字の書き取り実験を行った。実験は、詳細読みで説明を行う従来手法と、詳細読みの説明語に対して Word2vec を用いた自動的な類義語の抽出を行い意味情報として付与する提案手法について各々実施した。実験問題として 50 問 100 文字の漢字の問題を作成した。なお、提案手法の有効性を確認するため、詳細読みの説明語に同音異義語を多く含む漢字を問題に用いた。被験者はコンピュータの画面を隠してシステムを使用し、ボタンの押下で問題遷移、PC-Talker による説明音声の再生を行う。そして、ランダムな順に再生される漢字の説明を聞き、想起される漢字を解答用紙に記述する。実験では事前に数問の練習問題を行った。システムによる漢字の説明例を以下の図 1 に示す。

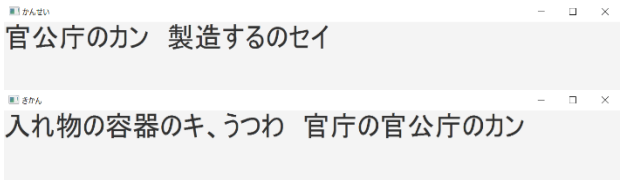

図 1 従来手法(上)と提案手法(下)による漢字の説明例

実験後はアンケートを行い、説明の分かりやすさ、漢字の想起のしやすさ、説明語の同音異義語の区別のしやすさについて 5 段階で評価をしてもらった。

以上の条件で実験を行い、被験者の一文字ごとの漢字の正答率と一問当たりの解答時間は以下の表 2、アンケート結果は以下の図 2 のようになった。

表 2 一文字ごとの漢字の正答率（一問当たりの解答時間）

被験者	従来手法	提案手法
A	80%（20 秒）	87%（23 秒）
B	74%（16 秒）	82%（20 秒）
C	78%（14 秒）	87%（14 秒）
D	67%（24 秒）	79%（26 秒）
E	86%（24 秒）	93%（18 秒）
F	84%（29 秒）	87%（19 秒）
G	60%（24 秒）	69%（17 秒）
H	83%（18 秒）	89%（20 秒）
I	73%（15 秒）	87%（14 秒）
J	88%（13 秒）	92%（14 秒）
平均正答率（平均解答時間）	77%（20 秒）	85%（18 秒）

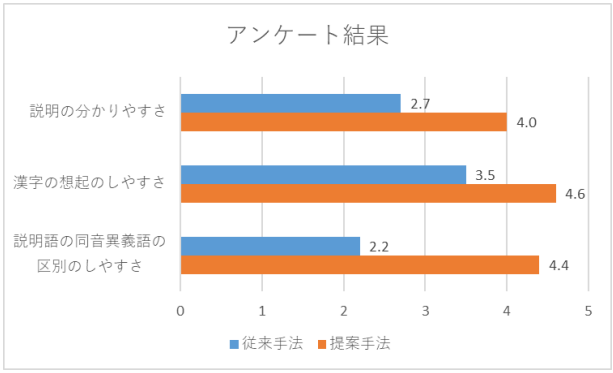

図 2 アンケート結果

表 2 より、一文字ごとの漢字の平均正答率について、従来手法では 77%、提案手法では 85%、二文字ごとの漢字の平均正答率について、従来手法では 59%、提案手法では 74%という結果になった。また、被験者 10 名全員、従来手法よりも提案手法の正答率が高かった。特に提案手法は、詳細読みの説明語に同音異義語が多く、訓読みによる説明がない漢字に有効であった。一問当たりの平均解答時間について、従来手法では 20 秒、提案手法では 18 秒という結果になり、大きな差はなかった。図 2 より、アンケート結果の平均点について、3 つの項目全てにおいて、従来手法よりも提案手法の方が高くなることが分かった。以上の結果から、従来手法よりも提案手法の方が漢字を正確に想起しやすいことが分かった。

5.おわりに

本研究では、詳細読みの説明語に対して Word2vec を用いた自動的な類義語の抽出を行い、意味情報として付与する手法を提案した。実験の結果、従来手法よりも提案手法の方が漢字を正確に想起しやすいことが分かった。今後は、視覚障がい者によって評価を行う予定である。

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] 西田昌史, 堀内靖雄, 黒岩眞吾, 市川熹, ”視覚障害支援のための意味情報に基づく仮名漢字変換, ”電子情報通信学会論文誌, VOL. J95-D, NO. 4, pp. 960-968, 2012/4.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, “Distributed Representations of Words and Phrases and their Compositionality,” NIPS, 2013.