テキストコーパスを用いた漢字詳細読みの自動生成

川崎 博章 † 笹野 遼平 ‡ 高村 大也 ‡ 奥村 学 ‡
† 東京工業大学 総合理工学研究科, ‡ 東京工業大学 精密工学研究所
kawa@lr.pi.titech.ac.jp, {sasano, takamura, oku}@pi.titech.ac.jp

はじめに

漢字詳細読みとは対象の漢字一文字を説明する読み方であり，スクリーンリーダに搭載されている重要な機能の 1 つである．多くの漢字には同音異字が存在しており，漢字詳細読みには音声による説明のみでユーザに漢字を正しく想起させることが求められる．しかし既存のスクリーンリーダの出力の中には，対象の漢字の想起が難しい漢字詳細読みが存在している．また，漢字詳細読みで用いる単語はユーザに慣れ親しんだものであることが望ましいが，時間の経過やユーザの背景の変化がある度に，人手で対応するコストは小さくないと考えられる．そこで本研究では，単語の親密度と同音異字語を考慮に入れた，コーパスを用いた漢字詳細読みの自動生成法を提案する．

関連研究

コンピュータの普及に伴い，漢字詳細読みの音声出力を機能として持つスクリーンリーダは視覚障害者の間で広まった．それに従い，スクリーンリーダにより出力される漢字詳細読みの問題も議論されるようになった．

大山ら [7] は，人名で使用されている漢字に焦点をあて，日本人の名前の漢字表記を自動的に説明し合成音声で出力する説明文生成の対話システム EXPLANET を提案している．EXPLANET は多くの同音の他の漢字から目的の漢字を明確に区別することが可能で，説明の際には目的の漢字の構成や他の単語を用いている．

渡辺ら [5] は漢字詳細読みのタイプについて分析を行なっており，その結果より漢字詳細読みは以下の 3 タイプに分類できる．

タイプ 1 対象の漢字を含む単語とその読み
“コウニュウ (購入) のコウ” (“購”)
タイプ 2 対象の漢字の独特な読み
“サクラ” (“桜”)
タイプ 3 対象漢字の特徴とその読み
“サンズイのカワ” (“河”)

渡辺らによると，タイプ 1 の漢字詳細読みが最も多く使用されている．タイプ 1 は統計的処理に適しており，また，タイプ 1 は基本的に全ての漢字に適用可能であることから，本論文ではタイプ 1 の漢字詳細読みの自動生成を試みる．

さらに渡辺らは対象の漢字を想起することができない要因について考察しており [4]，以下のような要因を挙げている．

要因 1 “チヨガミのヨ” という漢字詳細読みで用いられている “千代紙” のような低い親密度の単語の存在
要因 2 “購買” と “勾配” のような同音異字語の存在
要因 3 “爾” のような難解な漢字の存在

このうち要因 1 と要因 2 は漢字詳細読みに適切な単語を用いることにより改善できると考えられるが，要因 3 は想起させる漢字そのものが難しいことが要因であるため改善は難しい．本論文では漢字詳細読みによる対象漢字の想起率の向上を目的としており，要因 1 と要因 2 に焦点を当てる．

漢字詳細読みの自動生成

本論文で提案する漢字詳細読みの自動生成法はインタラクティブなシステムである．概要を図 1 に示す．第一段階として，各漢字に対し，できる限り曖昧性がなく親密度が高い単語を用いたタイプ 1 の漢字詳細読みを 1 つ出力する．

しかし，タイプ 1 の漢字詳細読み 1 つでは曖昧性なく 1 つの漢字を想起させるのが困難な漢字も存在する．例えば “科” という漢字の場合，それを含む最も一般的な単語は “科学” や “教科”，“単科” などであるが，“科学” には “化学”，“教科” には “強化”，“単科” には “炭化” や “単価” などの同音異字語が存在するため第一段階により生成される漢字詳細読みだけでは不十分である．そこで提案するシステムでは，第一段階の漢字詳細読みでは曖昧であるとユーザが返答した場合，第二段階として第一段階でユーザに提示した漢字詳細読みと組み合わせることによりユーザに漢字を曖

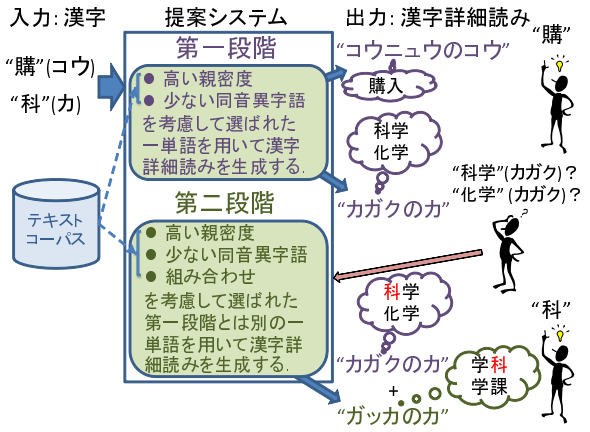

図 1: 提案システムの概要図

昧性なく想起させるような漢字詳細読みをさらに 1 つ出力する．以下では第一段階と第二段階の漢字詳細読みの生成法について詳述する．

第一段階の漢字詳細読み生成法

i. コーパスから，二文字以上かつ対象の漢字を含む単語を抽出する．

ii. 各単語に対しスコアを以下の式により計算する．

$$score_1(w) \triangleq p(w)^{\alpha} \cdot u_1(w)^{\beta} \quad (1)$$

ただし，w ：候補単語

$p(w)$ ：コーパス中の単語 w の出現確率

$u_1(w)$ ：単語 w と同じ読みを持つコーパス中の全単語の合計出現頻度に対する，単語 w の出現数の割合

$\alpha, \beta \in [0, 1]$ ：パラメータ

iii. 最も高いスコアとなる単語 w を用いて，漢字詳細読みを生成する．

式 (1) では，$p(w)$ が親密度度合いを，$u_1(w)$ が同音異字語の少なさを表す．従って，パラメータ α, β はそれぞれ親密度度合いと同音異字語の少なさに関係するパラメータとなる．例えば “購” という漢字を含む単語には “購読” や “購入”，“購買” などがある．このうち，“購読” は他の単語よりも出現確率は高いが，“鉱毒” などの同音異字語が存在する．項 $u_1(w)$ はこのような同音異字語を持つ単語のスコアを下げる働きをする．その結果，“購読” ではなく “購入” のような曖昧性のない単語を用いた漢字詳細読みが優先して出力される．

第二段階の漢字詳細読み生成法

i. コーパスから，二文字以上かつ説明したい漢字を含む単語を抽出する．

ii. 抽出した単語の組全てにスコアを付ける．スコアは以下の式により計算する．

$$score_2(w_1, w_2) \triangleq score_1(w_1) \cdot score_1(w_2) \cdot u_2(w_1, w_2)^{\gamma} \quad (3)$$

ただし，w_1 ：第一段階により選択された単語

w_2 ：第一段階とは別の漢字詳細読みを生成する候補単語

$u_2(w_1, w_2)$ ：単語の組み w_1, w_2から想起可能な漢字に対する，説明したい漢字の割合

$\gamma \in [0, 1]$ ：パラメータ

iii. 最も高いスコアの単語 w_2 を選択した後，w_2 を用いて漢字詳細読みを生成する．

$u_2(w_1, w_2)^{\gamma}$ は二単語用いたときの曖昧性の少なさを表している．例えば，“カガクのカ” と “タンカのカ” は，“科学” と “単科” から “科” という漢字と，“化学” と “炭化” から “化” という漢字の少なくとも 2 つの漢字を想起させる可能性があるが，項 $u_2(x, y)^{\gamma}$ はこのような曖昧な漢字詳細読みのスコアを下げる．この場合システムは “カガクのカ” と “ガッカのカ” のように曖昧のない漢字詳細読みの組を優先して出力する．

実験

実験設定

実験では以下の 3 つのコーパスを用いる．

- Google 日本語 N グラムコーパス [3]
- 読売新聞コーパス [6]
- 現代日本語書き言葉均衡コーパス (BCCWJ) [2]

想起対象の漢字が難しいことに起因するエラー (2 節の要因 3) をなるべく無視できるように，実験では Google コーパス中に現れる出現頻度上位 2,000 個の漢字を用いた．上記の 2,000 個の漢字の合計出現頻度は全出現漢字の 99% 以上を占めており．実用上の観点から十分であると考えられる．また，パラメータ (α, β, γ) は予備実験の結果に基づき $(0.1, 1.0, 1.0)$ に設定した．

表 1: 3 つのコーパスの比較の結果

	Google コーパス	読売新聞コーパス	BCCWJ	PC-Talker XP
a	179	170	185	185
b	15	15	9	9
c	6	15	6	6
IR[%]	89.5	85.0	92.5	92.5

つのコーパスの比較

3 つのコーパスの内，どのコーパスが提案手法に最も適しているかを調査するために，3 つのコーパスを用いて第一段階の漢字詳細読みを生成し評価した．比較対象としてスクリーンリーダ PC-Talker XP [1] に搭載されている漢字詳細読みを利用した．このため，各漢字に対し 4 つの漢字詳細読みが存在することになる．

出現頻度上位 2,000 個の漢字のうち PC-Talker XP による漢字詳細読みがタイプ 1 である漢字の中から，無作為に 100 個の漢字を選び評価に使用した．評価者による偏りが比較結果に影響を与えないように，各評価者には手法ごとに 25 個づつ，合計 100 個のカタカナで記した漢字詳細読みを提示した．漢字詳細読みの評価では書き取りを行うことも考えられるが，実際に漢字詳細読みが使用される場面では漢字を書き取る事が出来る必要はほとんどなく，漢字の想起が出来れば十分である場合が多いことから，漢字の想起の可否により漢字詳細読みの評価を行った．評価は 8 名の評価者により行い，各漢字詳細読みを 2 名が評価するように調整した．評価者には提示された漢字詳細読みから最もふさわしい漢字を選択した後に，次の選択肢から 1 つを選択するように伝えた．

a. 漢字を想起し，正解だった
b. 漢字を想起したが，不正解だった
c. 漢字を想起しなかった

各コーパスに対し，正しく漢字が想起された割合である想起率 (Identification Rate, IR) を次式により計算した．$n(x)$ は選択肢 x が選ばれた回数である．

$$\mathrm{IR} = \frac{n(\mathbf{a})}{n(\mathbf{a}) + n(\mathbf{b}) + n(\mathbf{c})} \times 100 \ [\%]$$

表 1 に結果を示す．読売新聞コーパスの想起率は他のコーパスに比べ低いが，その要因の 1 つとして親密度は高いが新聞では利用頻度の低い単語が存在することが考えられる．例えば Google コーパスや BCCWJ を使用した場合，“貌” という漢字に対して “美貌” という単語を用いた漢字詳細読みが生成されるが，読売新聞コーパスでは “美貌” という単語があまり出現しないため，“外貌” という親密度の低い単語が選ばれる．二番目に想起率の高い Google コーパスとの間には 0.05 水準での McNemar 検定で有意差は確認できなかった．次節でのスクリーンリーダと提案手法の比較の際には想起率の最も高い BCCWJ を利用する．

提案手法の評価

続いて提案手法全体の評価を行った．比較対象として PC-Talker XP の出力を利用した．使用する漢字として出現頻度上位 2,000 個から無作為に 100 個の漢字を抽出した．2 つの手法により合計 200 個の漢字詳細読みを用意し，4.2 節の評価と同様にバイアスがかからないように混ぜ，評価者 60 人には各々50 個の漢字詳細読みを示した．従って各漢字詳細読みは 15 人により評価されることになる．

提案手法は二段階構成であり，第一段階の漢字詳細読みの出力で評価者が 1 つの漢字に絞れないと判断した場合は第二段階の漢字詳細読みを出力する．従って評価の際には，まず提案システムは第一段階の漢字詳細読みを評価者に提示し，評価者の要求に応じて第二段階の漢字詳細読みを追加で提示する．評価者には以下の 5 つの選択肢 ($\mathbf{a}_1$,$\mathbf{b}_1$,$\mathbf{a}_2$,$\mathbf{b}_2$,$\mathbf{c}$) から適切なものを 1 つ選ぶように伝えた．

- 第一段階の漢字詳細読みのみを見て，1 つの漢字を想起した [$\mathbf{a}_1$. 正解だった，$\mathbf{b}_1$. 不正解だった]
- 第二段階の漢字詳細読みまで見て，1 つの漢字を想起した [$\mathbf{a}_2$. 正解だった，$\mathbf{b}_2$. 不正解だった]
- [$\mathbf{c}$. 漢字を想起することはなかった]

提案システムの第一段階または第二段階のいずれかで正しい漢字を想起することができれば良いので，選択肢 $\mathbf{a}_1$ と $\mathbf{a}_2$ を正解として想起率 IR_2 を次式により計算した．

$$\mathrm{IR}_2 = \frac{n(\mathbf{a}_1) + n(\mathbf{a}_2)}{n(\mathbf{a}_1)+n(\mathbf{b}_1)+n(\mathbf{a}_2)+n(\mathbf{b}_2)+n(\mathbf{c})} \times 100 \ [\%]$$

また，参考のために，提案手法の第一段階の出力による想起率 IR_1 を次式により計算した．

$$\mathrm{IR}_1 = \frac{n(\mathbf{a}_1)}{n(\mathbf{a}_1)+n(\mathbf{b}_1)+n(\mathbf{a}_2)+n(\mathbf{b}_2)+n(\mathbf{c})} \times 100 \ [\%]$$

表 2 に結果を示す．4.2 節の想起率 (92.5%) よりも本節の第一段階の想起率 (78.7%) の方が低い理由は 2 つある．1 つ目の理由は，評価方法の違いによるものであり，評価者が複数の漢字を想起した際に，4.2 節の評価の際には複数の漢字を 1 つに絞ってから答え合わせをするため正解に含まれる可能性があったが，今回

表 3: BCCWJ を用いて提案システムが生成した漢字詳細読みと PC-Talker XP による出力の例とその評価

漢字	提案システムの出力		PC-Talker XP
	第一段階	第二段階	
嘩	“フウフゲンカ (夫婦喧嘩) のカ”(4/15)	“オオゲンカ (大喧嘩) のカ”(5/15)	“ケンカスル (喧嘩する) のカ”(1/15)
課	“カダイ (課題) のカ”(13/15)	“カゼイ (課税) のカ”(15/15)	“カゼイスル (課税する) のカ”(6/15)
圭	“ケイジロウ (圭二郎) のケイ”(2/15)	“ケイイチ (圭一) のケイ”(4/15)	“ツチヲフタツカサネタ ケイ”(11/15)
藍	“ガラン (伽藍) のラン”(0/15)	“アイハラ (藍原) のアイ”(4/15)	“アイイロ (藍色) のアイ”(12/15)

“(n/15)” は，15 人の内 n 人が正解の選択肢を選んだことを意味する

表 2: 提案システムとスクリーンリーダの比較結果

	提案システム (BCCWJ)	PC-Talker XP
$\mathbf{a}_1$	1,181	1,301
$\mathbf{b}_1$	28	58
$\mathbf{a}_2$	163	-
$\mathbf{b}_2$	22	-
$\mathbf{c}$	106	141
IR[%]	IR_1:78.7 IR_2:89.6	86.7

の評価の際には必ず不正解となるためである．2 つ目の理由は，使用する漢字が異なることによるものである．4.2 節の評価の際には使用する漢字は PC-Talker XP による出力がタイプ 1 のものに限っていたが，今回の評価の際にはそのような制限はしておらず，今回の評価ではタイプ 1 に適していない漢字も評価対象に含まれているためである．

システム全体で見た場合，提案手法は PC-Talker XP よりも良い性能を示していることが確認できた．McNemar 検定の結果，BCCWJ を用いた第二段階までの提案手法の出力と，PC-Talker XP の間には，0.05 水準の有意差が有ることが確認できた．

表 3 に漢字詳細読みの例とその評価を示す．提案手法がスクリーンリーダよりも良い性能を示した漢字として，“喧” と “課” がある．“喧” という漢字の場合，スクリーンリーダの出力 (喧嘩する) では “献花する” という同音異字語を想起する可能性があるが，提案手法では同音異字語の無い単語を利用しており，想起率の向上に成功している．“課” という漢字の場合，提案手法の第一段階から漢字を想起出来なかった残りの 2 人は，第二段階の出力まで見ることにより正しい漢字を想起しているので，提案手法は 15 人全員に “課” という漢字を想起させることに成功している．スクリーンリーダのこの漢字に対する低い正解率の原因は，評価者が “加勢する” を想起したためであると考えられる[1]．

一方，提案手法がスクリーンリーダよりも悪い性能を示した漢字の例として，“圭” と “藍” がある．“圭” という漢字の場合，スクリーンリーダは “土の上に土が乗っている” という説明により “圭” という漢字の形を説明し，提案手法の場合と比べ多くの評価者に正しい漢字を想起させることに成功している．“藍” という漢字の場合，スクリーンリーダの出力は出現頻度は高くないものの “藍” という色を直接想起させる単語 (藍色) を用いることにより多くの評価者に正しい漢字を想起させている．

[1] “加勢” は (カセイ) と読むが，“多勢” は (タゼイ) と読むことから，“勢” は (ゼイ) と読むこともあるので，評価者は “加勢する” を (カゼイスル) と読み間違えたと考えられる．

おわりに

本論文では，漢字の親密度と同音異字語を考慮に入れた，テキストコーパスを用いた漢字詳細読みの自動生成法を提案した．評価者による評価の結果，システムにより生成された漢字詳細読みが，スクリーンリーダに搭載されているものよりも性能が良いことを確認した．今後の課題としては，ユーザへの適応が挙げられる．ユーザの属性は様々であるが，各ユーザが書いた文章などをコーパスとして選び，ユーザごとに提案手法を実行することで，ユーザごとに適切な漢字詳細読みを生成することが可能であると考えている．

参考文献

[1] 株式会社高知システム開発. PC-Talker XP. http://www.pctalker.net/.

[2] 国立国語研究所. 現代日本語書き言葉均衡コーパス. http://www.tokuteicorpus.jp/.

[3] 工藤拓, 賀沢秀人. Web 日本語 N グラム第 1 版. 2007.

[4] 渡辺哲也, 藤沼輝好, 渡辺文治, 澤田真弓, 鎌田一雄. 視覚障害者用スクリーンリーダの「詳細読み」に関する検討. 電子情報通信学会技術報告，HCS2002-41, 2003.

[5] 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, Vol.J88-D-I，No.4，pp.891–899, 2005.

[6] 読売新聞社. 読売新聞記事データ集. http://www.nichigai.co.jp/dcs/index2.html.

[7] 大山芳史, 浅野久子, 高木伸一郎. 姓名漢字表記を説明する対話システムの試作と評価. *IPSJ SIG-SLP*, Vol. 96, No. 123, pp. 53–58, 1996.