2008－MPS－68 (17)
2008／3／5

原言語音韻を考慮した多様なカタカナ異表記生成モデル

服部弘幸[†] 関 和広[††] 上原邦昭[†††]

言語表記における問題の一つに，文字表記の揺れ（異表記）がある．例えば，「ロサンゼルス」は「ロスアンゼルス」や「ロサンジェルス」のようにも表記が可能である．特に，日本語においてはこのようなカタカナ異表記が多く存在しており，自然言語を機械的に処理する際の障害となっている．これに対処するため，異表記生成などの研究が行われている．しかし，従来の研究ではカタカナ表記の字面を利用した書き換え規則による手法が主流であり，カタカナ語の原言語等，より深層的な特徴は考慮されていない．また，これらの研究では書き換え規則の獲得にコーパスを用いているため，学習データの影響を受けやすい．そこで本論文では，原言語の音韻がカタカナ表記に関係している点に着目し，英語由来のカタカナ語を確率的に原言語音素列に変換，さらにその音素列をカタカナ語に逆変換することで多様なカタカナ異表記を自動生成するモデルを提案する．また，提案モデルを情報検索システムの検索質問拡張に利用し，評価実験を行う．

A Generative Model for Diverse Katakana Variants based on English Phonetic Orthography

HIROYUKI HATTORI,[†] KAZUHIRO SEKI[††] and KUNIAKI UEHARA[†††]

In Japanese orthography, there is often more than one way to spell a phoneme sequence. This is especially true for katakana words which are typically transliterations from foreign languages. For example, "Los Angeles" can be written as "*rosuanjerusu*," "*rosanzerusu*," or "*rosuanzerusu*" in Japanese; they all are considered legitimate. This ambiguity becomes a critical problem for automatic processing when those variants need to be associated with the same concept. To deal with the problem, this paper proposes a novel approach to produce katakana variants for a given katakana word based on a generative model that considers phonetic orthography of the original language for the given word. The proposed model is empirically evaluated based on the variants it generated. It is also shown that the model is beneficial for information retrieval systems when applied to query expansion.

1. はじめに

人間が言語を表記する際，異なる表記を持ちながらも同じ意味を担う異表記同義語（以下，異表記と記す）が利用されることがある．[1] 特に日本語では，ひらがな・カタカナ・漢字といった複数の文字種を用いているため，このような異表記が生じやすい．例えば，「ロサンゼルス」が「ロスアンゼルス」や「ロサンジェルス」のように表記されることがある．情報検索，機械翻訳などの分野においては，これらの表記が与えられた場合，システムは他の表記も考慮することが望ましい．なぜなら，計算機を用いて自然言語を扱う際に，表記間の適合関係が取れないためにシステムの精度が低下してしまうからである．

このような問題に対して，表記の統一や異表記の自動生成といった様々な研究が行われている．中でも，カタカナ異表記生成に関する研究では，書き換え規則を用いた手法が主流となっている．[2] これらの手法では主にカタカナ表記の字面に着目し，カタカナ文字の書き換え規則を限られたデータから学習している．よって，作成される書き換え規則は，規則作成に利用した異表記（学習データ）の影響を受けやすく，データの量が少ない場合・事例に偏りがある場合に，不完全な書き換え規則が作成されてしまう．また，大量かつ偏りがないデータを収集できたとしても，差異が大きい異表記に関しては有効な規則が作成できないおそれがある．

本研究では，機械翻訳などの分野で示された原言語の音韻とカタカナ表記の持つ関係性[3] に着目し，カタカナ語と英語間の音素対応を用いることでカタカナ異表記を生成するモデルを提案する．本モデルは，学習データとして異表記を必要としないため，既存の異表

† 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
†† 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University
††† 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

記に起因する上述のような問題を持たない．また，従来手法と全く異なる方法で異表記を生成するため，書き換え規則では生成が難しい多様な異表記を生成できる可能性がある．

2. 提案手法

カタカナ語と英語間の音素の対応関係に関する研究が Knight と Graehl[3] によって行われた．この研究では，CMU 音素辞書☆と日本語・英語間のバイリンガル辞書を用いて，カタカナ音と英音素に関する対応関係を学習する．表 1 にこの対応関係の一部を示す（以下，この対応関係を「Knight らの対応表」と記す）．表中のカタカナ音とは，Knight らの研究で用いられる一つ以上のローマ字の組み合わせである．

表 1 カタカナ音と英音素の対応の一部
Table 1 A fragment of mappings between katakana and English phonetic representations.

英音素 e	カタカナ音 k	表記確率 $P(k\|e)$
AA	o	0.566
	a	0.382
G	g	0.598
UH	u	0.794

提案モデルでは，この Knight らの対応表を使用し，以下の流れで入力されたカタカナ語に対する異表記を生成する．なお，原言語として英語を使用するにあたって，本来，カタカナ語に対する原言語の推定が必要である．しかし，本研究ではこの問題に関しては取り扱っていない．

(1) カタカナ音への変換
(2) 英音素への変換
(3) カタカナ音への逆変換
(4) 候補語の生成および異表記の選定

以降の節で，それぞれの処理について説明する．

2.1 カタカナ音への変換

カタカナ語とローマ字の間には，ほぼ一対一の関係がある．そこで，カタカナ文字・ローマ字対応表[3] を用いて与えられたカタカナ語をローマ字に変換する．

ローマ字変換を行った後，Knight らの対応表から変換可能なカタカナ音を導出する．図 1 に，例としてカタカナ「ディテール」に対応する可能なカタカナ音列を示す．このとき，1 つのカタカナ音は小文字のアルファベット 1〜5 文字と「/」によって表現される．「/」は，1 つのカタカナ音が 2 文字以上で表現される際に，全体の音数を調整するために挿入される記号である．

以下の節では，「ディテール」に関する例を取り扱い，説明を行っていく．

2.2 英音素への変換

前節で得られたカタカナ音を Knight らの対応表を用いて英音素に変換する．具体的には，図 1 で示されるカタカナ音の並びから任意のカタカナ音列を撰択し，それぞれのカタカナ音に対応可能な英音素を列挙する．図 2 に，全てのカタカナ音列に対して変換可能な英音素列を導出した例を示す．

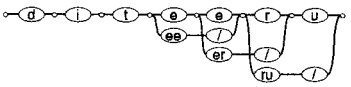

図 1 「d-i-t-e-e-r-u」に対応する可能なカタカナ音列
Fig. 1 Possible partitions for "d-i-t-e-e-r-u".

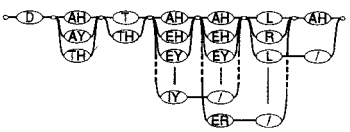

図 2 「d-i-t-e-e-r-u」に対応する可能な英音素列
Fig. 2 Possible English phonemes for "d-i-t-e-e-r-u".

ここで導出された英音素列はあくまで可能な英音素列に過ぎないため，この中から与えられたカタカナ語に対応し，かつ原言語としてもっともらしい英音素列を次のように推定する．まず，

- $K = k_1 k_2 ... k_n$: 与えられたカタカナ音列
- $E = e_1 e_2 ... e_n$: 導出された任意の英音素列

として，問題を以下のように定義する．

$$\hat{E} = \underset{E}{\operatorname{argmax}} P(E|K) \tag{1}$$

$$= \underset{E}{\operatorname{argmax}} P(K|E)P(E) \tag{2}$$

ここで $\hat{E}$ が求める英音素列である．上式を，英音素を隠れ状態とした隠れマルコフモデル（HMM）としてとらえ，またカタカナ音間の独立性を仮定すると，

$$P(K|E)P(E) = \Pi_{i=1}^{n} P(k_i|e_i)P(e_i|e_{i-1}) \tag{3}$$

が成り立つ．ただし，$P(e_1|e_0) = P(e_1)$ とする．式 (4) の第一因子 $P(k_i|e_i)$ は記号出力確率を，第二因子 $P(e_i|e_{i-1})$ は状態遷移確率を表わしている．本研究では，記号出力確率を Knight らの対応表中の表記確率によって取得し，状態遷移確率を CMU 音素辞書から得られる 127,000 の英音素列に基づいて推定した．

2.3 カタカナ音への逆変換

次に，前節で同定された英音素列 $\hat{E}$ をカタカナ音へ逆変換する．具体的には，$\hat{E}$ を構成する各英音素を Knight らの対応表に応じて可能なカタカナ音に変換することで，カタカナ音の組み合わせを得る．なお，各カタカナ音には Knight らの対応表によって表記確率が付与されている．

2.4 候補語の生成および異表記の選定

前節で得られたカタカナ音の組み合わせの中から，カタカナとして可能な組み合わせだけを異表記候補として生成する．しかし，生成される候補の数は非常に多くなるため，カタカナ文字の n グラム言語モデルによって「カタカナ語らしさ」を評価することで，候補

☆ http://www.speech.cs.cmu.edu/cgi-bin/cmudict

を選定する．ここでは，英和辞書 EDICT☆☆に含まれるカタカナ語 13,124 語を用いて 3 グラム言語モデルを構築した．

また，Web を巨大なコーパスと見なし，そこに表われない語は異表記として不適切であるという仮定のもと，候補語のさらなる選別を行う．本研究では，候補語を Yahoo!検索 API☆☆☆によって検索し，検索記事数が 1 以上の場合，候補語として採用した．

3. 評価実験

3.1 生成された異表記の有効性

本節では，提案モデルによって生成された異表記の質をアンケートによって評価する．

評価には，異表記を生じやすいと思われる観点から Infoseek マルチ辞書☆4などから選定した 25 単語を使用する．評価語を提案モデルの入力として異表記を生成し，それらを「自分なら異表記として使用してもよい（強肯定）」と「異表記として使われているかもしれない（弱肯定）」,「異表記ではない（否定）」という観点から被検者に判定してもらう．なお，被検者は工学系大学（院）の学生 17 名である．

表 5 に評価語の一覧と，各単語に対するアンケート結果を「強肯定率」,「弱肯定率」として示す．なお，表中の「–」は提案モデルによって異表記が生成されなかったことを示している．強肯定率は，各評価語について算出した

$$\frac{\text{被検者が強肯定と判定した異表記数}}{\text{提案モデルにより生成された異表記数}} \times 100\ (\%)$$

の被検者間の平均を示す．同様に，弱肯定率は各評価語について算出した

$$\frac{\text{被検者が弱肯定と判定した異表記数}}{\text{提案モデルにより生成された異表記数}} \times 100\ (\%)$$

の被検者間の平均を示す．評価語 25 個中，異表記が生成された 24 個に対する強肯定率の平均は 18.56%であり，標準偏差は 14.49%であった．また，弱肯定率の平均は 13.98%であり，標準偏差は 7.07%であった．

実験結果より，提案モデルによって生成された異表記のうち，18.56%は正しい異表記であり，弱肯定まで含めると 32.54%が異表記として使われ得ることが示された．なお，少くとも一人の被検者によって強肯定または弱肯定と判定された異表記 195 語のうち，174 語（89.2%）は書き換え規則を用いた既存手法（3.2 節参照）では生成できなかった．これは，多様な異表記を生成するという観点において，原言語の音韻を用いた本提案モデルの有効性を示すものである．

3.2 情報検索への応用

本研究で提案したカタカナ異表記生成モデルは，言語表現の多様性に対応するものであり，情報検索・機械翻訳などへの応用が期待できる．本稿では，提案モデルを情報検索の検索質問拡張に利用することで，その効果を検証する．なお，検索質問拡張とは，検索質問にその関連語（ここでは異表記）を追加することで，情報検索の精度向上を図る処理である．

検索課題には，NTCIR-3 の Web 検索テストコレクション[4] を利用し，カタカナ語を含む 26 件の検索質問に対して，提案モデルを用いて検索質問拡張を行った．

実験では，検索質問拡張を行わないシステムを（*Base*）とする．また，比較手法として EDICT から収集した約 750 組のカタカナ異表記より，編集操作を抽出し，5 回以上観測された操作 118 個を書き換え規則として持つシステム（*Rule*）を実装した．そして，提案モデルによるシステムを（*Phone*）とする．なお，*Rule* に関しても Yahoo!検索 API による候補語の選定を行った．

図 3 に検索結果に対する R-P 曲線を示す．図は，縦軸に適合率 P，横軸に再現率 R を示している．

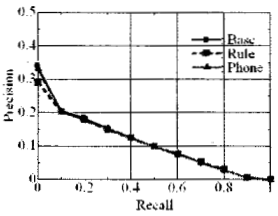

図 3 NTCIR データセットに対する R-P 曲線．
Fig. 3 Recall-precision curves for NTCIR dataset.

表 2 評価語一覧と結果
Table 2 Words for evaluations and its results.

評価語	異表記数	強肯定率	弱肯定率	合計
アイデンティティー	12	17.13	12.50	29.63
イノベーション	11	14.14	16.67	30.81
ウィニングボール	1	11.11	33.33	44.44
カバレージ	13	14.96	9.83	24.79
グラフィクス	8	18.06	12.50	30.56
シェーカー	23	8.70	8.45	17.15
スパゲッティ	6	29.63	12.04	41.67
ダイヤモンド	21	5.03	5.03	10.05
テイスト	13	6.84	6.41	13.25
ディテール	12	16.67	11.11	27.78
ネイル	2	41.67	8.33	50.00
パフューム	3	27.78	9.26	37.04
フリーエージェント	9	16.67	20.37	37.04
メーデー	32	8.51	6.25	14.76
モラトリアム	22	5.81	11.36	17.17
ユーザーネーム	7	23.81	11.90	35.71
ロサンゼルス	20	12.78	14.72	27.50
ソルトレイクシティ	0	–	–	–
エイジシュート	9	19.14	16.05	35.19
メーキャップアーチスト	13	26.92	16.24	43.16
アドバーティンメント	1	72.22	27.78	100.00
シャキラ	29	1.34	5.17	6.51
ケイティーホームズ	15	26.30	25.19	51.48
ウェアラブル	7	11.90	14.29	26.19
サプリメント	4	8.33	20.83	29.17
全体	293	18.56	13.98	32.54

☆☆ http://www.rdt.monash.edu.au/ jwb/japanese.html
☆☆☆ http://developer.yahoo.co.jp/search/
☆4 http://dictionary.www.infoseek.co.jp

図3からは，再現率が低い部分において *Rule* の適合率が低下していることが分かる．これは，他の手法と比べて *Rule* の検索結果が上位に非正解文書を多く含んでいるためであり，生成された異表記が悪影響をおよぼしている可能性を示唆している．これに対して，*Phone* は *Base* と比べてほぼ低下が見られない．しかし，向上もしていない結果となっている．

3.3 Web 検索への適用

前節の実験結果より，NTCIR-3 の Web 検索テストコレクションでは，異表記が検索に与える影響を適切に評価できないこと分かった，よって本節では，他の評価方法・データに基づいて提案モデルの有効性を検証する．

3.3.1 再現率に関する検証

再現率に関する検証を行うために，本稿では再現率に近似する他の評価指標を定義する．

まず，Yahoo!検索 API を用いた Web 検索システムを構築した．このシステムを（*Base*）とし，書き換え規則によるモデルを適用したシステムを（*Rule*），提案モデルを適用したシステムを（*Phone*）として使用する．また，検索語として 3.1 節で使用した評価語 25 単語を使用し，各システムに入力することで，検索された記事数がどの程度増加したかを確認する．そして，この結果を以下の式

$$\frac{\text{各システムの検索記事数} - \textit{Base}\text{の検索記事数}}{\textit{Base}\text{の検索記事数}}$$

に適用し，算出される値を「増加率」として定義する．ただし *Base* の検索記事数がゼロの場合はその値を ∞ とする．

増加率は，全ての異表記を検索質問として使用した結果である．そのため，厳密な意味での再現率と違い，あくまで近似的な意味合いを持つ評価値である．

全評価語に対する増加率の平均は *Phone* で 90.56，*Rule* で 3.36 であり，標準偏差は *Phone* で 310.86，*Rule* で 10.25 であった．

3.3.2 適合率に関する検証

本来，異表記の生成という処理は適合率に関与しない．しかし，前節で定義した増加率を評価指標として使用することで，正しくない異表記が検索システムへ与えられた場合の適合率への影響を考慮しなければならない．そこで，検索結果を人手で評価することで，提案モデルによって生成した異表記の情報検索への適用が，適合率にどのような影響をおよぼすかについて調べる．

ここでは，検索語として 4.1 節で用いた評価語を使用する．また，*Phone* と *Rule* を使用する．

まず，評価語 25 単語を入力し，質問置換を行う．そして，それらを検索語として与えた検索結果の上位各最大 20 記事を評価対象とする．評価者には，検索結果を「記事中で正しい異表記が使用されている（適合）」，「記事中に正しい異表記が使用されていない（非適合）」という観点で判断してもらう．なお，評価者は 4.1 節と同じ，工学系大学（院）の学生 17 人である．また，評価者は評価対象となる記事がどの手法による検索結果であるかは分からない．適合率は，

$$\frac{\text{適合記事数}}{\text{各システムの検索結果}}$$

で算出される値であり，全評価語に対する適合率の平均は *Rule* で 0.58，*Phone* で 0.62 であり，標準偏差は *Rule* で 0.34，*Phone* で 0.30 であった．

実験結果から *Phone* が *Rule* と比べて約 6.9%高い適合率を持つことが確認された．これより，*Phone* の作成するカタカナ異表記は従来手法と比べて，適合率への影響が少ないと考えられる．

また，この実験結果と 3.3.1 節の実験結果から，提案モデルによって生成された異表記を情報検索へ適用した場合，より希少な表記を入力し，正しい異表記が生成された場合に有効に働くと考えられる．そして，*Rule* との比較から，提案モデルによって生成される異表記は増加率，適合率の両観点から情報検索に有効であると考えられる．

4. 結　　論

本論文では，原言語音韻の表記確率を用いたカタカナ異表記の自動生成モデルを提案した．ローマ字と英音素の関係を利用することで，カタカナ語を英音素へ確率的に変換し，その後，カタカナ語への逆変換を行うことで候補語を作成した．そして，候補語を言語モデルと Web 検索を使用することで選定した．結果として，提案モデルによって生成されたカタカナ異表記のうち，被検者が許容できるものは，全体の約 18.56%～32.54%であることが確認できた．

また，生成された異表記を情報検索へと適用し，従来手法との比較を行うことで提案モデルの有効性を調べた．実験は，情報検索の検知から再現率に近似した評価指標である増加率および適合率によって判定することとした．実験の結果，提案モデルの全評価語に対する増加率の平均は 90.56，適合率の平均は 0.62 であった．この結果は，従来手法と比較しても高いことから，提案モデルによって生成される異表記は情報検索に有効に働くと考えられる．

参 考 文 献

1) 神門 典子: 情報検索システムの評価プロジェクト: NTCIR ワークショップ, 情報処理, 41(6), pp.689-697 (1999).
2) Masuyama, T, Sekine, S and Nakagawa, H: Automatic construction of Japanese katakana variant list from large corpus, Proceedings of the 20th International Conference on Computational Linguistics (COLING'04), Vol.2, pp.1214–1219 (2004).
3) Knight, K and Graehl, J: Machine transliteration, Computational Linguistics, Vol.24, No.4, pp.599–612 (1998).
4) Eguchi, K, Oyama, K, Ishida, E, Kando, N and Kuriyama, K: Overview of the Web retrieval task at the third NTCIR workshop, National Institute of Informatics Technical Report, NII-2003-002E (2003).