2007－NL－181 (10)
2007／9／25

英音素変換を用いたカタカナ異表記の自動生成

服部弘幸[†] 関 和広[††] 上原邦昭[†††]

情報検索における問題の一つに，文字表記の揺れ（異表記）があげられる．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ため，これらのうち任意の表記が検索語として与えられた場合，情報検索システムは他の表記も考慮することが望ましい．特に，日本語においては上記のようなカタカナ異表記が多く存在しており，これに対処するために様々な研究が行われている．しかし，これらの研究では異表記の生成に限られたコーパスを用いているため，データの過疎性に起因する問題が生じやすい．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複数のカタカナ異表記を自動生成する手法を提案する．また，NTCIR-3 の Web 検索テストコレクションを用いた評価実験について報告する．

Automatic Katakana Variants Generation via English Phonemes

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

In information retrieval and other text processing applications, there has been a problem concerned with variant notations. For example, "Los Angeles" can be written as "rosuanjerusu," "rosanzerusu," or "rosuanzerusu" in Japanese. Thus, it would be desirable that a search system considers all the notations given any of them as a query. Although, there has been much research conducted for dealing with the problem, the previous work typically relied on the katakana rewriting rules derived from Japanese corpora or search engine logs, which apt to be suffered from the data sparseness problem. This paper proposes---based on our observation that a number of katakana variants are influenced by the pronunciation in the source language—a method to automatically generate katakana variants by back-transliterating a katakana word. The proposed method is evaluated on the NTCIR-3 Web retrieval test collection.

1. はじめに

情報検索や機械翻訳において，異なる表記を持ちながらも同じ意味を担う語である異表記同義語（以下，異表記と記す）と呼ばれる文字表記の揺れが問題となっている.[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のでこのような問題が生じやすい．例えば，「取り扱い」が「取扱」，「とり扱い」と記されたり「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．よって，これらのうちの一表記が検索語として与えられた場合，情報検索システムは他の表記も考慮して検索を行うことが望ましい．

日本語の中でもカタカナ語は，外国語を音訳する際に頻繁に使用されるため異表記を生じやすく，新聞・書籍・放送・インターネットなどのメディアに登場する言葉の中で，新語として増加している点が特徴的である.[2] しかし，このようなカタカナ語は日常的な使用において，決まった定義がないままに用いられているので個人によってその表記に違い（揺れ）が生じやすい．典型的な表記の揺れの例として，コンピュータとコンピューター，ディテールとデテール，スパゲッティとスパゲティが挙げられる．これらの違いは，情報検索や機械翻訳といった分野で計算機を用いて処理を行う際に，表記間の適合関係が取れないのでシステムの精度低下を引き起こす原因となる．

現在，Web 上で広く利用されている検索エンジンの多くは，入力された検索語に加えて，異表記やスペルミスを考慮した候補語を提示することができる．しかし，日々新しい単語が増えている今日，情報検索システムが異表記に十分対処できていないという報告がある.[3]

† 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

この問題に対して本研究では，原言語の音韻がカタカナ表記に関係する点[4]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，さらにそれらの音素列をカタカナ語へ逆変換することで複数のカタカナ異表記を自動生成する手法を提案する．そして，提案手法により生成されたカタカナ異表記を使って検索質問を拡張することで，情報検索の精度向上を図る．

以下，2節でカタカナ異表記生成の関連研究について述べる．次に，3節で提案する英音素変換を使ったカタカナ異表記生成方法について述べ，4節で本手法の評価実験について述べる．5節で本論文の結論と今後の課題について述べる．

2. 関連研究

カタカナ異表記の生成に関する研究として，これまでコーパスを用いてカタカナ語の書き換え規則を学習させることで異表記を生成する研究[5]や，異表記を統一的な表記に置き換える研究[6]が行われてきた．しかし，これらの手法は限定的なコーパスを基に異表記の生成を行うため，データの過疎性に関する問題を起こしやすい．例えば，Web 情報検索に必要とされる語彙には，MSN の検索エンジンログを学習データとして使用しても不十分であるという報告がある．[7]

一方，機械翻訳や音声認識に関する研究として，カタカナ語と英語間の音素対応に関する研究がある．[8]このような研究は主にカタカナ語から英語，または英語からカタカナ語といった言語間の対応関係を導出している．そのため，これらの手法を使うことでコーパスを使用する手法とは異なる異表記生成が可能になると考えられる．

そこで本研究では，原言語の音韻がカタカナ表記と関係する点に着目し，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自動生成する手法を提案する．これにより，原言語の音韻に左右される特殊な表記や学習データに存在しないような新語に対しても柔軟に異表記を生成することが期待できる．

3. 提案手法

カタカナ語と英語間，英語と英音素間には通常，一対一の対応関係が存在しない．しかし，両言語を限定した数の音素表現に変換することで，それらの対応を実現できる．[9]

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[8]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ることで，カタカナ音と英音素のペア 8000 組に関する対応関係を学習させている．カタカナ音とは，カタカナ音素と英音素の対応をとるために Knight らが用いた一つ以上のカタカナ音素の組み合わせである．

本研究では，このカタカナ音と英音素間の関係を使用し，与えられたカタカナ語を可能なカタカナ音列，続いて英音素列へと変換する．そして，得られた英音素列をカタカナ語に逆変換することでカタカナ異表記候補語（以下，略して候補語と記す）を生成する．しかし，ここで生成される候補語は対応関係において可能なものを網羅的に列挙するため膨大な数になる．そこで，音素の連接確率と単語の共起を用いることでよりもっともらしいカタカナ異表記（以下，単に異表記と記す）を選定する．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を取り扱っていない．

提案手法では，以下の流れで異表記を生成する．

(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

以下，それぞれの処理について説明する．

3.1 カタカナ音への変換

カタカナ語とその音となるローマ字の間には，ほぼ一対一の関係がある．ここでは，Knight らのカタカナ文字・ローマ字対応表[8]を用い，与えられたカタカナ語をローマ字に変換する．対応表を表 1 に示す．ここで，吃音を表す「ッ」は，その後ろに来る文字の子音部分を重複させ，長音を表す「ー」は，その前に来る文字の母音部分を重複させて表現する．また，表 1 に加えて複合カタカナ文字に対応した表 2[9]も併せて使用する．表 2 は，Knight らの音素表記に合わせて改変を加えている (例えば，ジャ：ja → ジャ：jya)．

表 1 カタカナ文字・ローマ字対応表
Table 1 Katakana characters and their phonetic (roma-ji) representations.

ア: a	タ: ta	マ: ma	ギ: gi	ビ: bi
イ: i	チ: chi	ミ: mi	グ: gu	ブ: bu
ウ: u	ツ: tsu	ム: mu	ゲ: ge	ベ: be
エ: e	テ: te	メ: me	ゴ: go	ボ: bo
オ: o	ト: to	モ: mo	ザ: za	パ: pa
カ: ka	ナ: na	ヤ: ya	ジ: ji	ピ: pi
キ: ki	ニ: ni	ユ: yu	ズ: zu	プ: pu
ク: ku	ヌ: nu	ヨ: yo	ゼ: ze	ペ: pe
ケ: ke	ネ: ne	ラ: ra	ゾ: zo	ポ: po
コ: ko	ノ: no	リ: ri	ダ: da	ン: n
サ: sa	ハ: ha	ル: ru	ヂ: ji	ヴ: v
シ: shi	ヒ: hi	レ: re	ヅ: zu	ー: <YY>
ス: su	フ: hu	ロ: ro	デ: de	ッ: <XX>
セ: se	ヘ: he	ワ: wa	ド: do	—
ソ: so	ホ: ho	ガ: ga	バ: ba	—

ローマ字変換を行った後，Knight らの対応関係から変換可能なカタカナ音を導出する．図 1 に，例とし

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

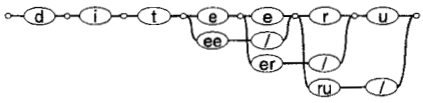

図 1 「d-i-t-e-e-r-u」に対応する可能なカタカナ音列
Fig. 1 Possible partitions for " d-i-t-e-e-r-u".

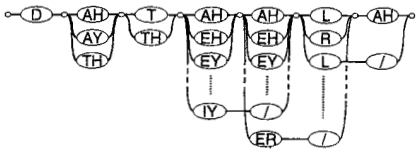

図 2 「d-i-t-e-e-r-u」に対応する可能な英音素列
Fig. 2 Possible English phonemes for " d-i-t-e-e-r-u".

表 2 複合カタカナ文字・ローマ字対応表
Table 2 Compound Katakana Characters and their phonetic representations.

ディ: di	ツィ: tsi	チョ: chyo	ピュ: pyu
デゥ: du	ツェ: tse	ニャ: nya	ピョ: pyo
ティ: ti	ツォ: tso	ニュ: nyu	ギャ: gya
テゥ: tu	シェ: she	ニョ: nyo	ギュ: gyu
シィ: si	ジェ: je	ヒャ: hya	ギョ: gyo
ウィ: wi	チェ: che	ヒュ: hyu	ジャ: jya
ウェ: we	キャ: kya	ヒョ: hyo	ジュ: jyu
ウォ: wo	キュ: kyu	ミャ: mya	ジョ: jyo
ヴァ: va	キョ: kyo	ミュ: myu	ヂャ: dya
ヴィ: vi	シャ: shya	ミョ: myo	ヂュ: dyu
ヴェ: ve	シュ: shyu	リャ: rya	ヂョ: dyo
ヴォ: vo	ショ: shyo	リュ: ryu	ビャ: bya
ヴュ: vyu	チャ: chya	リョ: ryo	ビュ: byu
ツァ: tsa	チュ: chyu	ピャ: pya	ビョ: byo

てカタカナ「ディテール」を与えたときに選択された変換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．

3.2 英音素への変換

前節で得られたカタカナ音を Knight らの対応関係を用いて英音素に変換する．図 1 で示されるカタカナ音の並びからカタカナ音列を 1 つ撰択し，対応関係から可能な英音素を網羅的に選択する．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．このとき，1 つの英音素は大文字のアルファベット 1〜2 文字と「/」によって表現される．

しかし，ここで選択された英音素はあくまで可能な英音素であり，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推定する必要がある．まず，

- $K = k_1 k_2 ... k_X$: 観測されたカタカナ音列
- $E = e_1 e_2 ... e_X$: 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．

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

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

$$= \underset{E}{\operatorname{argmax}} P(K|E)P(E) \tag{3}$$

これを，英音素が隠れ状態とした隠れマルコフモデル (HMM) ととらえ，かつカタカナ音間の独立性を仮定すると，

$$P(K|E)P(E) = \Pi_{i=1}^{n} P(k_i|e_i)P(e_i|e_{i-1}) \tag{4}$$

式 4 の第一因子 $P(k_i|e_i)$ は記号 (カタカナ音) 出力確率であり，Knight らの対応関係から得られる．一方，第二因子 $P(e_i|e_{i-1})$ は状態遷移確率を表わし，CMU 音素辞書から与えられる 127,000 の英単語の音素列から学習している．また，状態遷移先に「/」を含むような音については，全体のカタカナ音数と英音素数を調整するために次の時刻で「/」を挿入している．「ディテール」のカタカナ音に対して最も確率の高かった英音素列は「D-IH-T-EY-L」であった．

3.3 カタカナ音への逆変換

次に，前節で推定された最も確率の高い英音素列をカタカナ音へ逆変換する．推定された英音素列を Knight らの対応関係によっ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が生成される．表 3 に英音素「D-IH-T-EY-L」が与えられたときのカタカナ音の組み合わせを示す．なお，各カタカナ音には対応表によって出力確率が付与されている．

表 3 「D-IH-T-EY-L」の各音素に対応するカタカナ音
Table 3 Katakana phonemes corresponding to " D-IH-T-EY-L".

D	IH	T	EY	L
d	i	t	ee	r
do	e	to	a	ru
ddo	—	tto	e	—
—	—	—	ei	—

3.4 候補語の生成および異表記の選定

前節で与えられたカタカナ音の組み合わせの中から表 1，2 に基づき，可能なカタカナ語の組み合わせだけを候補語として生成する．また，英和辞書 EDICT☆☆ よりカタカナ語のみを抽出してカタカナ文字に対する tri-

☆☆ http://www.rdt.monash.edu.au/ jwb/japanese.html

gram モデルを作成する．これによって前節で導出した確率にカタカナ語の並びに対する確率を掛け合わせ，カタカナ語らしい候補語を選定する．また，候補語と与えられた検索語を共に Yahoo!検索 API☆☆☆を使って検索することで，両単語が共起するような記事の件数を取得した．これをフィルタとして候補語の選定に使用する．以下に，閾値として共起する記事数を 10 以上とした場合の「ディテール」に対して生成された候補語上位 5 つとその記事数を示す．

(1) ディテール 1970000
(2) ディテイル 329
(3) デテール 195
(4) ディテル 86
(5) ディタル 36

なお，この方法によって多くのノイズを除去できるものの，低頻度の正しい異表記まで除去してしまう可能性がある．この点については，今後さらなる検討が必要である．

4. 評価実験

本研究で提案した異表記生成法は，情報検索の精度向上を目的としている．そこで，提案手法によって生成されたカタカナ異表記を用いて検索質問拡張を行い，その効果を検証する．以下，実験に使用したデータ，実験結果等について詳述する．

4.1 使用データ・評価方法

検索課題には，NTCIR-3 の Web 検索テストコレクション[10] を利用した．テストレベルは NTCIR-3 に準拠する DM2&RL1 を対象としている．DM2&RL1 は，適合判定の際に文書本体のみを考慮し，検索質問に対して返ってきた記事が正解データによってランク A (適合) 以上と判定された場合に，その記事を正解とする．コレクションには 47 件の検索質問が与えられており，このうちカタカナ語を含む 26 件の質問に対して拡張を行った．

実験では，検索質問拡張を行わなず，ベクトル空間モデルを使って tfidf およびコサイン類似度によって検索を行う手法をベースライン (*Base*) とする．また，比較対象として英和辞書 EDICT から収集した約 750 組のカタカナ異表記より，各ペアに対する編集操作を抽出し，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規則に基づく手法 (*Rule*) を実装した．そして，本論文で提案した手法によって検索質問中のカタカナ語を拡張する手法 *Phone* とする．なお，*Rule* に対しても Yahoo!検索 API によるフィルタを使用した．

4.2 実験結果

表 4 に検索質問中に含まれるカタカナ語と，各手法の検索結果を示す．表中の値は，各検索質問に対する検索結果 (最大 1000 件) に含まれる正解文書数の割合を示している．今回，実験に使用した検索質問全 26 件に対する評価は，*Base* が 0.0283，*Rule* が 0.0288，提案手法である *Phone* が 0.0282 であった．表 4 において，アスタリスク (*) が記されている結果は異表記が存在しないと判定された結果を示している．

表 4 上位 1000 件の適合率
Table 4 Precision at top 1000 retrieved documents.

カタカナ語	Base	Rule	Phone
サルサ	0.0500	0.0510	0.0490
オーロラ	0.0050	0.0050	0.0050
オゾン層，オゾンホール	0.0650	0.0650	0.0650*
ゲノム	0.0220	0.0220	0.0220*
ベースボール	0.0020	0.0040	0.0020
ロープワーク	0.0180	0.0180*	0.0180*
インターネット	0.0020	0.0020	0.0020
テービング	0.0110	0.0110	0.0110
レビュー	0.0010	0.0010	0.0000
デジタルコンテンツ，ネットワーク	0.0340	0.0350	0.0330
スピーカー	0.0080	0.0080	0.0060
アカデミー賞	0.0150	0.0160	0.0160
キューブリック	0.0720	0.0720	0.0720
ゲーム	0.0280	0.0300	0.0260
パイプオルガン，コンサートホール	0.0330	0.0330	0.0330*
バイク，ツーリング，レポート	0.0380	0.0380	0.0380
アニメーション	0.0050	0.0050	0.0050
モネ	0.0690	0.0680	0.0730
イースター，キリスト	0.0630	0.0630	0.0630
シフォンケーキ	0.0540	0.0540	0.0540*
アロマセラピー，アロマオイル，アロマキャンドル	0.0000	0.0000	0.0000*
カプサイシン	0.0350	0.0350	0.0350*
アントシアニン，ブルーベリー	0.0180	0.0180	0.0180
ポリフェノール	0.0650	0.0710	0.0650*
N ゲージ，HO ゲージ	0.0010	0.0010	0.0010*
グレートバリアリーフ，オーストラリア	0.0220	0.0220	0.0220
全体	0.0283	0.0288	0.0282

4.3 考　　察

実験結果から，カタカナ異表記を使った検索質問拡張による情報検索への顕著な効果は観測できなかった．そこで，効果が表われない原因を特定するため，元の検索語を生成した候補語によって置換し，質問置換によって候補語のみを使用した検索結果が正解となる記事を取得しているかを観測する．この時，異表記が存在しないと判定されたカタカナ語を含む質問を省略し，異表記が存在する語と存在しない語が混在すると判定された質問は，異表記が存在しないカタカナ語に対してのみ元のカタカナ語を使用して検索を行った．また，検索結果が 0 件になった場合も省略している．実験結果を表 5 に示す．

実験結果から，いくつかの質問で正解判定となる記

☆☆☆ http://developer.yahoo.co.jp/search/

表 5 候補語による質問置換を行ったときの上位 1000 件の適合率
Table 5 Precision at top 1000 retrieved documents retrieved only by katakana variants.

カタカナ語	Rule	Phone
サルサ	—	0.0000
オーロラ	0.0050	0.0000
ベースボール	0.0000	0.0000
インターネット	0.0020	0.0000
テーピング	—	0.0000
レビュー	0.0000	0.0010
スピーカー	0.0080	0.0080
アカデミー賞	0.0010	0.0010
キューブリック	0.0010	0.0000
ゲーム	0.0000	0.0000
アニメーション	0.0000	0.0000
モネ	0.0000	0.0270
イースター，キリスト	0.0630	0.0010
シフォンケーキ	0.0000	—
カプサイシン	0.0000	—
アントシアニン，ブルーベリー	0.0180	—
ポリフェノール	0.0020	—
N ゲージ，HO ゲージ	0.0010	—
全体	0.0063	0.0150

事を検索できていないことが分かる．そこで各質問に対する Rule および Phone の検索結果の上位 10 記事を人手によって調べたところ，次のようなことが分かった．

- 「カプサイシン」を候補語「キャプサイシン (Rule によって作成)」などで置換した結果，人手ではランク A 以上と思われるが，正解データでは対象となっていない記事が 1 件あった．
- 「デジタルコンテンツ」を候補語「ディジタルコンテンツ (Rule)」などで置換した結果，人手ではランク A 以上と思われるが，正解データでは対象となっていない記事が 2 件あった．
- 「ポリフェノール」を候補語「ポリフェノル ポリフエノール (Rule)」などで置換した結果，正解データでは対象となっていない記事が 6 件あった．
- 「ベースボール」を候補語「ベイスボール (Phone)」などで置換した結果，正解データでは対象となっていない記事が 8 件あった．

例えば「カプサイシン とうがらし 効能」という検索質問に対するランク A の正解データおよび検索結果の一部を以下に抜粋する．

- 正解データ
  「…この辛みはカプサイシンと呼ばれる刺激性を有する成分によります．この刺激性を利用して，ペースト状にしたトウガラシを湿布に使い，挫骨神経痛や痛風の痛み止めに利用しました．また，凍傷による手足の指の麻痺の回復にも活用されています．…」
- 候補語を使った質問置換による検索結果
  「…唐辛子やカレーライスのような場合には，キャプサイシンと呼ばれる辛味成分が含まれており，それは温度の上昇で興奮する脳の 温度感受性ニューロンや皮膚にある 温度受容器 を刺激するように働きます．このような場合，動物はよけいに熱く感じて体温を下げるように反応を起こします．…」

次に「ポリフェノール」に対する質問置換を行った場合の適合判定結果と，文書中に含まれていた検索語あるいはその異表記を表 6 に示す．

表 6 候補語による質問置換を行った「ポリフェノール」に関する検索結果
Table 6 Top 10 documents retrieved by query substitution for "polifenol".

ランキング	判定	検索語もしくは異表記
1	A	ポリフェノール，ポリフエノール
2	C (不適合)	ポリフエノール
3	—	ポリフェノル
4	—	ポリフエノール
5	—	ポリフェノル
6	C	ポリフェノール，ポリフエノール
7	C	ポリフェノール，ポリフエノール
8	—	ポリフエノール
9	—	ポリフェノル
10	—	ポリフエノール

表 6 から，「ポリフェノル」「ポリフエノール」といった候補語のみを含む記事 6 件が適合判定なしと判定されていることが分かる．

以上の事実は，本実験で使用したテストコレクションが異表記に十分対応していないことを示唆しており，これが原因となって候補語を使った検索質問拡張の効果が観測できなかったと考えられる．

一方，検索結果を人手で調べて実際に精度が悪かった例も存在した．「モネ」や「ゲーム」のような短い単語は，音素変換を用いる Phone では「ミニ」や「マネ」，「ガム」や「ガン」などの別の意味を持つ候補語を多く生成して精度に影響を与えていた．また，書き換え規則を使用した Rule は少ない候補語の中で「キャプサイシン」や「ポリフエノール」などの有用な異表記を生成していた．しかし，前述したような特殊な変換を必要とする異表記や，新語への異表記対応は提案手法が有用であると考える．今後は，違うデータセットを使った提案手法の有用性の検証や，提案手法と他の手法を組み合わせることで異表記の生成をより知識的に行えるように改良していく．

5. 結　　論

本論文では，英音素変換を用いたカタカナ異表記の自動生成手法について説明した．提案手法より生成された候補語を使って検索質問拡張を行い，情報検索への効果を調べた．結果として，検索質問拡張を行った

ことによる情報検索への効果は顕著に表われなかった．検索結果を人手で検証したところ，実験に利用したテストコレクションで適合文書と判定されていなくとも，実際には適合文書である例が見つかった．ただし，短いカタカナ語に関しては異なる意味を持つ異表記を生成してしまうことがあり，今後フィルタの改良や他の手法との組み合わせによる，より知識的な異表記生成を行う必要がある．

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) 三省堂編修所: ネットでよくひくカタカナ新語辞典, 三省堂 (2005).
3) 久保村千明, 亀田弘之: 片仮名異表記処理能力を備え持つ情報検索システム, 信学技報, TL2000-25, pp.418-428 (2000).
4) 松崎 寛: 外来語音の表記のゆれに関する定量的研究, 東北大学文学部日本語学科論集, Vol.3, pp. 83–94 (1993).
5) 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).
6) Shishibori, M and Aoe, J: A method for generation and normalization of katakana variant notations, IEICE Transaction on Information and System D-II(2), J77-D-II, pp.380–387 (1994).
7) Qing, C, Mu, L and Ming, Z: Improving query spelling correction using Web search results, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.181–189 (2007).
8) Knight, K and Graehl, J: Machine transliteration, Computational Linguistics, Vol.24, No.4, pp.599–612 (1998).
9) Gregory, G, Yan, Q and David, A, E: Mining the Web to create a language model for mapping between English names and phrases and Japanese, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'04), pp.110–116 (2004).
10) 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).