情報処理学会論文誌　数理モデル化と応用　Vol. 2　No. 1　145–155 (Feb. 2009)

英語音韻を考慮した情報検索のための多様なカタカナ異表記生成

服部弘幸[†1]　関　和広[†2]　上原邦昭[†3]

日本語，特にカタカナ語では，異なる表記を持ちながら同じ対象を指す異表記同義語が多く存在する．たとえば，「ロサンジェルス」は「ロサンゼルス」，「ロスアンジェルス」，「ロスアンゼルス」のように表記することもできる．このような表記の多様性は，文字を単なる記号として扱う処理，たとえば情報検索などにおいて処理精度を低下させる要因の 1 つとなっている．具体的には，検索語として「ロサンジェルス」が与えられたとき，通常の情報検索では異表記のみを含む文書はけっして検索されることがない．この問題への対処法は，表記の統一，異表記の生成のいずれかに大別でき，後者の異表記生成には，これまでカタカナ書き換え規則に基づく手法が提案されている．本研究では，2 言語間の音素の不一致によって前述のようなカタカナ異表記が生じている場合がある点に注目し，表層的なカタカナ書き換え規則ではなく，より根源的な音素レベルでの異表記生成を試みる．提案手法では，従来研究の音素間対応を基に確率的音素変換モデルを構築し，カタカナ語から英語への逆翻字，英語からカタカナ語への翻字を連続的に行うことで，従来の表層的な特徴を利用した書き換え規則では得られない多様なカタカナ異表記を生成する．提案手法の妥当性を検証するため，生成された異表記に関して人手で評価を行う．さらに，生成された異表記を検索質問置換に利用し，情報検索における提案手法の有効性を示す．

Generating Diverse Katakana Variants via Backward-Forward Transliteration for Information Retrieval

Hiroyuki Hattori,[†1] Kazuhiro Seki[†2] and Kuniaki Uehara[†3]

In Japanese, it is quite common for the same word to be written in multiple ways. This is especially true for katakana words which are typically used for transliterating foreign languages. For example, “Los Angeles” can be written in katakana as “ロサンジェルス (*rosanjerusu*),” “ロサンゼルス (*rosanzerusu*),” “ロスアンジェルス (*rosuanjerusu*),” or “ロスアンゼルス (*rosuanzerusu*),” all considered legitimate. This ambiguity becomes a critical problem for automatic processing such as information retrieval. To tackle this problem, we propose a simple but effective approach for generating katakana variants for a given katakana word based on phonemic representation of the original language for a given word. The proposed approach is first evaluated through a manual assessment of the variants it generates. It is also shown that the approach is beneficial for information retrieval when applied for query replacement, retrieving a large number of potentially relevant documents.

1. はじめに

我々が日常用いる自然言語では，異なる表記を持ちながら同じ意味を担う「異表記同義語」（以下，異表記）が利用されることがある．特に日本語では，ひらがな・カタカナ・漢字など複数の文字種を用いていることもあり，このような異表記が生じやすい[1)]．たとえば，「取り扱い」が「取扱」と記される送り仮名による異表記，「ワイシャツ」が「Y シャツ」と記される文字種による異表記，「ロサンゼルス」が「ロスアンジェルス」のように記されるカタカナ表記の曖昧性による異表記などがある．人間の談話理解の過程では，このような表記の多様性は無意識に，あるいは容易に同一概念として解釈されうるのに対し，計算機で同様の解釈を得るには明示的な処理が必要となる．たとえば，日英機械翻訳を考えた場合，ロサンジェルスあるいはロスアンジェルスの正しい語訳を得るためには，これらの異表記を同一概念，すなわち Los Angeles に関連付ける必要がある．

上述の機械翻訳，あるいは情報検索など，自然言語処理の関連研究分野においては，このような表記の多様性に対処するためにいくつかの研究が行われてきている．これらの研究は，異表記問題を扱う枠組みの方向性から，表記を統一する方法と異表記を生成する方法に大別できる．特に後者では，カタカナ文字列書き換え規則（たとえば「ゼ → ジェ」など）を用いた手法が主流となっている．これに対し，本研究では，情報検索システムへの応用を念頭に，英語と日本語の音韻の不一致に着目することで多様なカタカナ異表記を生成する手法を提案する．さらに，提案手法によって生成された異表記を様々な観点から評価し，その有効性を検証する．

†1 グーグル株式会社
Google, Inc.
†2 神戸大学自然科学系先端融合研究環
Organization of Advanced Science and Technology, Kobe University
†3 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

なお，類似の問題として綴り訂正問題[2] がある．これは，システムのユーザが誤った語を入力した際，正しい綴りの語を提示する処理であり，現在のウェブ検索エンジンには標準的な機能である．たとえば，Yahoo! Japan[⋆1] にアーティスト名 Alisha Keys（誤）を与えた場合，Alicia Keys（正）が提示される．本研究で扱う異表記生成問題は，標準的な正しい表記を提示するのではなく，与えられた語の異表記として使われうる表記を網羅的に生成し，漏れのない検索を実現することを目的とする点，異表記は基本的に綴り誤りではないという点において綴り訂正問題とは異なる．

以下，2 章でカタカナ異表記処理の関連研究についてまとめる．続いて，3 章で提案手法について詳述し，4 章で提案モデルの評価実験について報告する．最後に，5 章で本論文のまとめと今後の課題について述べる．

2. 関 連 研 究

本章では，情報検索のためのカタカナ異表記処理に焦点を当て，関連研究をまとめる．

従来のカタカナ異表記問題に対する解決法は，表記の統一・標準化，または異表記生成の 2 つに大別できる．前者の統一・標準化[3),4)] による処理は比較的古くから存在し，あらかじめ与えられた分類規則を基に異表記の検出および統一を行い，索引付けと検索に利用する[5),6)]．冒頭の「ロサンジェルス」，「ロサンゼルス」，「ロスアンジェルス」，「ロスアンゼルス」の場合，表記の類似性などによってこれらが異表記であることが検出され，検索対象文書集合中におけるそれぞれの使用頻度などに基づいて，たとえば「ロサンゼルス」が標準的表記として索引付け・検索に利用される．一方，後者の異表記生成では，索引付けは文書中に現れた字面どおりに行い，検索の際にユーザから与えられた検索語の異表記を生成，検索に供することで対応する．たとえば，「ロサンゼルス」という検索語が与えられた場合，その異表記である「ロスアンゼルス」「ロサンジェルス」などを何らかの方法で獲得し，所与の検索語とともに（ブーリアン OR）検索に利用する．なお，このように検索質問を同義語などで拡張することを「検索質問拡張」という．所与の検索語に関して異表記を得る素朴な方法は，個々のカタカナ語に対してその異表記を列挙した異表記辞書を整備することである．しかし，この方法では，辞書の網羅性，既存エントリの保守・管理，新出カタカナ語といった問題への対応が必要となる．

異表記生成のもう 1 つの方法として，カタカナ文字列書き換え規則（たとえば「ゼ → ジェ」など）による方法がある．この方法は静的な辞書を持たず，与えられた任意の語に対して異表記が生成可能であるため，新語にも対応することができる．このような特長から，近年の研究では後者の書き換え規則による手法が主流となっている．一例として，久保村ら[7)] は，書籍や新聞記事，インターネットなどからカタカナ語を収集し，人手で異表記を同定・分析することにより書き換え規則の作成を試みた．分析の結果，久保村らはカタカナ異表記を小型文字異表記（例：フ**ィ**ルムとフ**イ**ルム），長音記号異表記（例：パ**ター**ンとパタン），その他（例：デ**ィ**ジタルとデジタル）に分類し，これらの類型に対応する計 258 種類の書き換え規則を作成した．これらの書き換え規則に基づいて作成した異表記（候補）を人手で評価したところ，平均 17.6%の候補が異表記として利用可能であり，さらに，生成された異表記を用いて検索質問拡張を施した場合，100 件中 47 件の検索語について既存の検索エンジンの検索ページ数を増加させることができたと報告している．

関連研究として，Masuyama ら[8)] は，文書中のカタカナ異表記を自動的に同定するため，大規模なコーパスから抽出したカタカナ語とその文脈を利用して，文字列間の類似度を推定する手法を提案した．類似研究では，文字列間の類似度を測る指標として編集距離を用いることが多い．しかし，挿入・削除などの編集操作に関するペナルティは，予備実験などから経験的に決められることが多かった．Masuyama らは，コーパスから収集したカタカナ語の前後の一定の窓内の語を文脈ととらえ，類似の文脈を持つ候補のみを異表記と扱い，編集距離の重みを自動的に算出した．評価実験として別のコーパスからすべてのカタカナ語を収集し，前述の重みを用いて算出した編集距離に基づき異表記候補ペアを獲得，さらにもう 1 度文脈類似度で各候補ペアを評価することで，カタカナ異表記の組を同定した．無作為に選んだ 682 組に関して精度を評価したところ，98.6%の再現率および 86.3%の適合率で異表記ペアを同定できたと報告している．

これらの研究では，すでに存在するカタカナ異表記間の表記の差異から，規則作成者の内省，あるいは統計的情報などに基づいて，書き換え規則や編集距離の重みを手動あるいは自動で作成する．この結果得られる規則・重みは，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などには，不完全な書き換え規則が作成されてしまう可能性がある．また，大量かつ偏りがないデータを収集できたとしても，編集距離などに基づいて異表記間の差異を検出するという方法では，表層的な差異が大きい異表記に関しては有効な規則を作成することができない．

本研究では，カタカナ語の表層的な違いではなく，より深層的な英語と日本語の音韻が持つ関係性[9)] に着目し，音素の不一致を利用してカタカナ異表記を生成する手法を提案する．

⋆1 http://search.yahoo.co.jp/

本手法は，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，従来手法とまった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．

3. 提案手法

3.1 あらまし

カタカナ異表記が生じる原因の 1 つとして，日本語と他の言語で使われる音素の違いがある[10]．ある言語で使われる音素は日本語では存在せず，逆に日本語で使われる音素は他の言語には存在しないことがある．たとえば，英語の母音/æ/は日本語のどの母音にも完全には一致しない．そのため，日本人がこの音を聴いた場合，類似する他の音，たとえば/a/や/e/として理解するため，それぞれの音に対応する表記が生まれてしまう．例として，人名の Chandler/tʃæ̲ndlə/はチャンドラーともチェンドラーとも表記可能である．この観測に基づき，本研究では，英語由来のカタカナ語をいったん英語の音素表記に変換（逆翻字）し，さらに日本語音素に変換（翻字）し直すことで，多様なカタカナ異表記を生成する．なお，本論文では原言語として英語だけを想定するものの，原理的に本提案手法は他の言語にも適応可能である．

提案手法では，次の 4 ステップによって所与のカタカナ語に対する異表記を動的に生成する．

(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および候補語の生成
(4)　異表記の選定

以降の節で，それぞれの処理について詳説する．

3.2 カタカナ音への変換

このステップでは，所与のカタカナ語をローマ字に変換したのち，ローマ字を英語音素と対応可能な日本語音素列に分割する．前者のローマ字変換については，Knight ら[9] のカタカナ文字・ローマ字対応表（**表 1**）を用いる．なお，吃音「ッ」は，その後ろに来る文字の子音部分を重複させ，長音「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** もあわせて使用する．表 2 では，Gregory ら[11] の表を基に，Knight らの音素表記に合わせて改変を加えている．例として，カタカナ語「ディテール」は，これらの表に基づき，ローマ字「diteeru」に変換される．

表 1　カタカナ文字・ローマ字対応表
Table 1　Katakana characters and their phonetic 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	
ス：su	フ：hu	ロ：ro	デ：de	
セ：se	ヘ：he	ワ：wa	ド：do	
ソ：so	ホ：ho	ガ：ga	バ：ba	

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

ローマ字変換ののち，Knight らが導出した音素対応確率表（一部を**表 3** に示す）に基づき，英語音素と対応可能な日本語音素列（これを「カタカナ音」と呼ぶ）への分割を行う．この際，一般的に日本語音素列と英語音素との対応には複数の可能性があるため，複数の分割の仕方が存在する．たとえば，「diteeru」末尾の「ru」は，全体で 1 つの英音素（L）に

表 3　Knight ら[9] の音素対応確率表の一部
Table 3　A fragment of English-Japanese phonemic mappings.

英音素	カタカナ音	音素対応確率
D	d	0.535
	do	0.329
ER	aa	0.719
	a	0.081
	ar	0.063
	er	0.042
EY	ee	0.641
	a	0.122
	e	0.114
IH	i	0.908
L	r	0.621
	ru	0.362
T	t	0.463
	to	0.305
	tto	0.103
UH	u	0.794
	uu	0.098

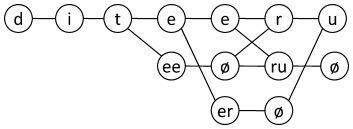

図 1　ディテール（diteeru）に対応する可能な分割（カタカナ音列）
Fig. 1　Possible partitions for “diteeru”.

対応する可能性もあるし，「r」と「u」のそれぞれが 1 つの英音素（L と UH）にも対応しうる．ここでは，Knight らの音素対応確率表に現れるすべてのカタカナ音を利用して分割を行い，その結果得られる可能なカタカナ音列を網羅的に取得する．なお，この処理は表 3 のカタカナ音のみを利用し，英音素および音素対応確率には（現時点では）依存しないことに注意を要する．

例として，**図 1** にカタカナ語「ディテール」に対応する可能な分割（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「ϕ」によって表現される．「ϕ」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無音記号である．

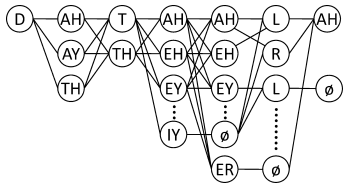

図 2　ディテール（diteeru）に対応する可能な英音素列
Fig. 2　Possible English phoneme sequences for “diteeru”.

3.3　英音素への変換

本研究で想定するカタカナ語は英語の外来語であり，そのようなカタカナ語は原言語の音訳であることから，あるカタカナ音列と元の英音素列の間には何らかの依存関係がある．これを統計的自然言語処理で一般的に用いられる雑音のある通信路モデル（noisy channel model）にあてはめれば，元の英音素列が情報源系列，カタカナ音列が観測系列に相当する．このモデルを基に，本ステップでは所与のカタカナ音列を生成した英音素列を確率的に求める．

まず，3.2 節の処理で得られた（複数の）カタカナ音列をそれに対応する英音素列に変換する．しかし，一般的に個々のカタカナ音（たとえば「a」）は複数の英音素（たとえば「ER」と「EY」）に対応するため，1 つのカタカナ音列（たとえば d-i-t-ee-ru）は個々の音素間対応の組合せ数分の英音素列に対応することになる．**図 2** に，図 1 の各カタカナ音を対応可能な英音素に展開した例の一部を示す．たとえば，図 1 の中ほどのカタカナ音「e」と「ee」は，図 2 の同じ位置にある英音素「AH」「EH」「EY」「IY」などに対応している．

ここで，それぞれのカタカナ音列 $J = j_1 \dots j_n$（j_i はカタカナ音）がある英音素列 $E = e_1 \dots e_n$（e_i は英音素）から生成された確率を $P(E|J)$ とし，これを最大化する英音素列 $\hat{E}$ を求める．

$$\begin{aligned} \hat{E} &= \arg\max_E P(E|J) \\ &= \arg\max_E P(J|E)P(E) \end{aligned} \tag{1}$$

同式右辺における変形・整理はベイズ則による．さらに，カタカナ音間の独立性と英音素の単純マルコフ過程（英音素は直前の英音素のみに依存して決定）を仮定すると，次式が成り立つ．

$$P(J|E)P(E) = \prod_i P(j_i|e_i)P(e_i|e_{i-1}) \quad (2)$$

ここで，数式を簡潔に表現するため，便宜上 $P(e_1|e_0) = P(e_1)$ とした．なお，上述の仮定は必ずしも現実の事象を反映していないものの，複雑な問題を計算機上で扱う際には効果的な場合があり，Knight ら[9] の逆翻字モデルも同様の仮定をおいている．

式 (2) を，英音素 e_i を隠れ状態とする隠れマルコフモデル[12] ととらえた場合，右辺の第 1 因子と第 2 因子は，それぞれ記号出力確率と状態遷移確率に対応する．記号出力確率 $P(j_i|e_i)$ の推定値としては，カタカナ語と英語のペア 8,000 組から Knight ら[9] が推定した値を用いた．一方，状態遷移確率 $P(e_i|e_{i-1})$ の推定には，CMU 発音辞書[⋆1]収録の 127,000 の英単語を用いた．なお，状態遷移の可能な組合せ数 1,571（= CMU 発音辞書で利用されている音素数 39 の 2 乗）に比して，確率パラメータ推定に用いた単語数は十分大きいと考えられるため，確率ゼロの状態遷移は英音素の遷移としてありえないと仮定し，あえて確率の平滑化などは行っていない．

前出の例「ディテール」の可能なカタカナ音列 J に関して式 (2) を最大にする英音素列 $\hat{E}$ を求めたところ，正しい音素列である「D-IH-T-EY-L」が得られた．

3.4　カタカナ音への逆変換および候補語の生成

このようにして得られた最も確からしい英音素列 $\hat{E}$ を可能なすべてのカタカナ音列 J' に逆変換し，それぞれの J' ごとにカタカナ異表記候補 K' を生成する．J' から K' への変換は表 1 と表 2 に基づいて一意に行われるため，この過程で曖昧性は生じない．

3.5　異表記の選定

以上の手続きにより生成されたすべての K' は潜在的に元のカタカナ語 K の異表記となりうるものの，現実には通常使われない異表記も多く生成される．そこで，これらの誤りを除外するため，本研究では 2 つの指標を利用する．1 つは $\hat{E}$ が J' に対応し，かつカタカナ語 K' が生成される確率 $P(K') \cdot \prod_i P(j'_i|\hat{e}_i)$ である．この確率の第 1 因子 $P(K')$ は，文字ベースの n グラム言語モデルによって，第 2 因子は，式 (2) の第 1 因子と同様に Knight らの推定値を利用して得られる．なお，本研究では n は 3 とし（すなわちトライグラム），日英辞書 EDICT[⋆2]に収録された 13,124 のカタカナ語に基づき，平滑化にラプラス法を用いた最尤推定で得た．この推定確率がある閾値以下の候補 K' は生成誤りとして除外する．

⋆1 http://www.speech.cs.cmu.edu/cgi-bin/cmudict
⋆2 http://www.csse.monash.edu.au/~jwb/j_edict.html

表 4　ディテールの異表記として生成されたカタカナ語の例
Table 4　Examples of katakana variants generated for ディテール.

異表記	文書頻度	$P(K')\prod_i P(j'_i\|\hat{e}_i)$
ディテイル	329	0.000002
デテール	195	0.000017
ディテル	86	0.000003
ディタル	36	0.000003

なお，閾値は経験的に 10^{-6} とした．

もう 1 つの指標は，現実世界における K' のカタカナ語としての利用例の存在である．本研究では，ウェブを巨大なコーパスと見なし，K' のカタカナ語としての可否を判定した．具体的には，Yahoo! API によって K' の出現文書頻度（K' をクエリとしたときの検索ページ数）を獲得し，K' が少なくとも 1 つのウェブページで利用されている場合は，正しい異表記であると見なした．

上述の 2 つの指標を段階的に利用している理由は，片方の指標だけでは異表記をうまく選定できないことによる．前者の確率（の第 1 因子）はカタカナ文字間の局所的な連接確率であるため，語全体としては不正なカタカナ語を排除しきれない．一方，後者の出現文書頻度は元々の入力カタカナ語をまったく考慮しないため，K' が異義語であった場合，これを排除することができない．

以上の手続きにより，前述のディテールに関して得られた異表記の例を**表 4** に示す．なお，上述の文書頻度による誤表記除去処理により，ディテールに関しては，異表記候補を 46 語から 12 語へ削減することができた．

4. 評 価 実 験

本章では，まず提案手法により生成された異表記の妥当性を人手で検証する．続いて，生成された異表記を情報検索に用いた場合の効果を実験的に検証する．

4.1　生成された異表記の妥当性

4.1.1　評 価 方 法

評価用のカタカナ語として，Infoseek マルチ辞書[⋆3]のカタカナ見出し語，Yahoo!辞書内

⋆3 http://dictionary.www.infoseek.co.jp

の新語探険サイト⋆1などから，異表記を生じやすいと思われる 25 単語を選定した．これらは主に，長音記号や促音，拗音を含む語，あるいは含むような異表記も可能であると思われる語，さらに日本語として一般に定着していないと思われる新語などであり，選定の基準は著者の主観的な判断による．そして，これらの評価語を入力として提案手法により異表記を生成，評価した．評価の観点は，久保村ら[7] の評価実験を参考に，「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）」の三択とした．評価者は工学系大学（院）の学生 17 名である．

4.1.2 実験結果

表 5 に評価語の一覧と各評価語ごとに生成された異表記数・および評価結果を「強肯定率」および「弱肯定率」で示す．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意味する．強肯定率は，各評価語について算出した

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

の評価者間の平均を示す．同様に，弱肯定率は各評価語について算出した

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

の評価者間の平均を示す．評価語 25 単語における強肯定率の平均は 18.56%であり，弱肯定率の平均は 13.98%であった．すなわち，提案モデルによって生成された異表記のうち，2 割弱は正しい異表記であり，弱肯定まで含めると 32.54%が正しい異表記であることが分かった．評価に用いたカタカナ語が異なるため直接の比較はできないものの，この結果は，カタカナ書き換え規則を用いた久保村ら[7] の報告（弱肯定まで含めて 17.6%）から有望であると考えられる．書き換え規則を用いた従来手法とのより厳密な比較については，情報検索における精度向上という観点から 4.2 節で議論する．

4.1.3 考　　察

表 5 に示した個々の評価語ごとの結果は，評価者間の平均によって算出している．しかしながら，たとえ少数の評価者のみに肯定的に判定された異表記であったとしても，必ずしも誤りであるとは限らない．また，異表記と誤表記の境界は必ずしも明確ではなく，特に本研究が目指す情報検索への応用を考えた際は，少数の個人によって使われる異表記（あるいは誤表記）をも生成できることが，検索漏れを防ぐという観点からは望ましい場合がある．たとえば，「ケイティーホームズ」（女優）に対して生成された異表記「カティーホームズ」は，一般的には誤りであると考えられる．しかし，この表記でウェブを検索すると，ケイティーホームズを意味して利用されている例（のみ）を数ページ見つけることができる⋆2．この観測から，少数の評価者が肯定的に評価した異表記も情報検索には有用である可能性がある．そこで，評価語 25 語に対して生成された計 293 の異表記のうち，少くとも 1 人の評価者によって強肯定または弱肯定と判定された異表記を数えたところ，その数は 195（66.6%）であった．さらに，このうち 174（89.2%）の異表記は，表層的なカタカナ書き換え規則を用

表 5　評価語ごとの異表記判定結果
Table 5　Individual results for quality judgment of generated katakana variants.

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
アドバーテインメント	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
平均		18.56	13.98	32.54

⋆1 http://dic.yahoo.co.jp/newword/

⋆2 Google (http://google.com) において，2008 年 4 月 14 日現在．

いた既存手法（4.2.1 項参照）では生成することができなかった．これら 174 の異表記が必ずしも情報検索に有効であるとはいえないものの，この結果は，多様な異表記を生成するという目的に鑑みて，日英の音素間対応を用いた本提案手法の特長を示すものである．

4.2 情報検索における効果

本節では，提案手法によって生成された異表記を検索質問拡張に用いた場合の効果について，検索精度の観点から検証を行う．

4.2.1 評価方法

検索実験には，NTCIR-3 の Web 検索テストコレクション[13] を利用し，テストレベルは NTCIR-3 に準拠する DM2&RL1 とした．DM2&RL1 では，所与の検索質問に関して検索された記事が適合度 H（高適合）および適合度 A（適合）である場合，適合記事と見なされる．コレクションには 47 件の検索質問が含まれ，このうちカタカナ語を含む 26 件の質問（**図 3**）に対して提案モデルを用いて検索質問拡張を行った．1 つの検索質問に複数のカタカナ語が含まれる場合は，各々のカタカナ語について異表記を生成し，すべての異表記をまとめて検索質問拡張に用いた．

ベースラインの検索システムとしては，ベクトル空間モデル[14]，tfidf 単語重み付け[15]，およびコサイン類似度によって検索を行う一般的なシステム（以降 *Base* と呼ぶ）を用いた．*Base* を基に，提案手法によって生成された異表記を用いて検索質問拡張を行う検索システムを *Phone* と呼ぶ．また，比較のため，カタカナ書き換え規則に基づく従来の異表記生成手法を実装した．具体的には，英和辞書 EDICT からカタカナ異表記の組（計 750 組）を抽出し，各組に関して編集距離を算出した．この過程で，計 5 回以上観測された編集操作（挿入・削除・置換）118 個を書き換え規則として採用した．獲得した書き換え規則の例として，「ブ → ヴ」，「イッ→ イ」などがある．この書き換え規則によって生成された異表記を用いて検索質問拡張を行う検索システムを *Rule* と呼ぶ．なお，*Rule* に関しても Yahoo! API を用いた文書頻度による候補語の選定（3.5 節参照）を行った．

「サルサ」，「オーロラ」，「オゾン層，オゾンホール」，「ゲノム」，「ベースボール」，「ロープワーク」，「インターネット」，「テーピング」，「レビュー」，「デジタルコンテンツ，ネットワーク」，「スピーカー」，「アカデミー賞」，「キューブリック」，「ゲーム」，「パイプオルガン，コンサートホール」，「バイク，ツーリング，レポート」，「アニメーション」，「モネ」，「イースター，キリスト」，「シフォンケーキ」，「アロマセラピー，アロマオイル，アロマキャンドル」，「カプサイシン」，「アントシアニン，ブルーベリー」，「ポリフェノール」，「N ゲージ，HO ゲージ」，「グレートバリアリーフ，オーストラリア」

図 3　評価実験に用いた 26 件の検索質問に含まれるカタカナ語
Fig. 3　Twenty-six katakana queries from NTCIR-3.

4.2.2 実験結果

図 3 にあげた検索質問を基に，3 つのシステム，*Base*，*Phone*，*Rule* のそれぞれについて検索を行い，検索結果の上位 1,000 件を評価した．**図 4** にそれぞれのシステムの *R*–*P* 曲線を示す．*R*–*P* 曲線は，縦軸に適合率 *Precision*，横軸に再現率 *Recall* を持ち，それぞれ以下のように定義される．

$$\begin{aligned} Precision &= \frac{\mathrm{TP}}{\mathrm{TP+FP}} \\ Recall &= \frac{\mathrm{TP}}{\mathrm{TP+FN}} \end{aligned} \tag{3}$$

ここで，TP（true positive）は検索された文書中の正解文書数，FP（false positive）は検索された文書中の非正解文書数，FN（false negative）は検索されなかった正解文書数である．

図 4 から，*Base* の性能が最も高く，*Phone*，*Rule* の順に若干の性能の低下が見られた．特に，再現率が 0〜0.1 の範囲で，既存手法 *Rule* の適合率の低下が顕著であった．この結果

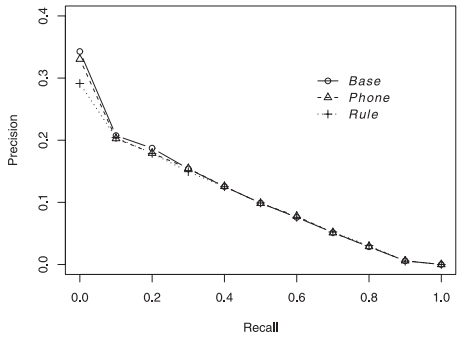

図 4　NTCIR テストコレクションにおける *R*–*P* 曲線
Fig. 4　*R*–*P* curves for NTCIR dataset.

は，他のシステムと比べて *Rule* の検索結果の上位に非正解文書が多く含まれていることを示しており，生成された異表記が悪影響を及ぼしていることを示唆している．一方，*Base* と比較して *Phone* の適合率低下は微小ではあったものの，期待した性能向上は見られなかった．

4.2.3 考　　察

前項の実験では，カタカナ異表記による検索質問拡張の顕著な効果を確認することはできず，逆に既存手法 *Rule* では適合率の低下が見られた．そこで，情報検索における異表記の影響をさらに詳細に分析するため，元の検索語の代わりに，生成された異表記だけを検索質問として用いて同様の検索実験を行った（これを「検索質問置換」と呼ぶ）．この実験によって，情報検索における異表記の有用性を生成された異表記のみに注目して評価することができる．**表 6** に，*Rule*，*Phone* それぞれの結果を上位 1,000 件の適合率で示す．表中の「—」は，異表記が生成されなかったことを表す．なお，検索結果が 0 件の検索質問は省略した．また，1 つの検索質問中で異表記が生成されるカタカナ語と生成されないカタカナ語が混在する場合は，異表記が生成されないカタカナ語についてのみ，元のカタカナ語を使用して検索を行った．

表 6　異表記による検索質問置換を行ったときの上位 1,000 件の適合率
Table 6　Precision at top 1,000 retrieved documents by query replacement.

元のカタカナ語	*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

表 6 から，全体的には提案モデル *Phone* が *Rule* より高い適合率を示すものの，多くの検索質問で適合文書を検索できていないことが分かった．そこで，各検索質問に関する *Rule* および *Phone* の検索結果の上位 10 記事を人手で分析したところ，次のような問題が明らかになった．なお，以下の「未判定」とは，所与の検索質問に関して適合性判定がなされていないことを意味する．

- 「カプサイシン」を異表記「キャプサイシン」などで置換した結果，NTCIR の正解データでは未判定であるものの，適合と思われる記事が 1 件あった．
- 「デジタルコンテンツ」を異表記「ディジタルコンテンツ」などで置換した結果，NTCIR の正解データでは未判定であるものの，適合と思われる記事が 2 件あった．
- 「ポリフェノール」を異表記「ポリフェノル」，「ポリフエノール」などで置換した結果，NTCIR の正解データでは未判定の記事が 6 件あった．
- 「ベースボール」を異表記「ベイスボール」などで置換した結果，NTCIR の正解データでは未判定の記事が 8 件あった．

例として，「カプサイシン とうがらし 効能」という検索質問に対する（NTCIR の正解データに基づく）適合文書，および未判定の検索結果の一部を以下に抜粋する．

- 適合文書
  「. . . 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．· · ·」
- 未判定の文書
  「. . . 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．· · ·」

このように，いずれの文書も同一概念の「capsaicin」に言及しており，かつ未判定の文書の内容も検索質問に適合していると考えられる．

もう 1 つの例として，「ポリフェノール」を検索質問置換したときの検索結果上位 10 件の（NTCIR 正解データによる）適合度，文書中に元のカタカナ語が含まれていたか否か，文書中に実際に含まれていた異表記を**表 7** に示す．表中の適合度「C」は不適合を意味する．

表 7 から，「ポリフェノル」，「ポリフエノール」といった異表記のみを含む文書 6 件が未

表 7「ポリフェノール」を検索質問置換したときの検索結果上位 10 件
Table 7　Top 10 documents retrieved by query replacement for “polyphenol”.

順位	適合度	元のカタカナ語を含む（Y/N）	出現異表記
1	A	Y	ポリフエノール
2	C	N	ポリフエノール
3	—	N	ポリフェノル
4	—	N	ポリフエノール
5	—	N	ポリフェノル
6	C	Y	ポリフエノール
7	C	Y	ポリフエノール
8	—	N	ポリフエノール
9	—	N	ポリフェノル
10	—	N	ポリフエノール

判定であることが分かる．これは，NTCIR の正解データがプーリング[16] によって作成されていることによると考えられる．プーリングでは，複数の異なる検索システムから得られた検索結果の上位 n 件（n は任意）の和集合のみを対象に適合性判定を行う．よって，プーリングに用いた検索システムが異表記を考慮していなければ，異表記のみを含む文書はそもそも評価の対象になっておらず，適合性判定は行われない．これが原因となって，前項の実験では異表記による検索質問拡張の効果が観測できなかった可能性がある．そこで次項では，異表記を考慮したシステムである *Rule* と *Phone* を用いて小規模なプーリングを行うことで，提案手法を再評価する．

なお，上述のカプサイシンとポリフェノールに関して，提案手法 *Phone* は有効な異表記を生成することができなかった．これは，カプサイシンについては，音素変換・逆変換の際にカ → キャと変換するような音素の対応確率が低かったことによる．ポリフェノールについては，候補「ポリフェノル」は生成されたものの，3.5 節の生成確率による選定によって除去されてしまった．提案手法は，書き換え規則に基づく手法とは異なる異表記を生成する傾向があるため，将来的には両者の結果を統合することを考えている．

4.2.4 追 加 実 験

情報検索システムを厳密に評価するためには，文書集合中のすべての文書について所与の検索質問との適合性を判定する必要がある．しかし，一般的にこれは不可能であるため，プーリングによってテストコレクションを構築するなどの代替法がとられる．前項で用いた NTCIR-3 の Web 検索テストコレクションはカタカナ異表記を考慮していないと考えられるため，本項では 4.1 節で用いた 25 のカタカナ語を用いて，情報検索における提案手法の有効性を検証する．NTCIR の検索質問と異なり，これらのカタカナ語は異表記を生じやすいという観点から選択されているため，異表記生成の手法の優劣がより顕著に現れると期待できる．

まず，カタカナ異表記を考慮することによってどれだけの多くの情報を見つけることができるのかを調査するため，4.1 節で少なくとも 1 人の評価者によって弱肯定または強肯定と判定された異表記のみを使って検索質問を拡張し，ウェブ検索を行った．検索には，Yahoo! 検索 API を用いた．その結果，元のカタカナ語だけを使って検索した場合と比較し，平均で 60.1 倍のウェブページを得ることができた．検索質問拡張を行えば検索されるページ数が増加することは当然の帰結であるものの，この実験では，潜在的に使われうる異表記のみを利用していることに留意する必要がある．よって，これらの異表記によって検索されたページは検索質問（元のカタカナ語）に関係する文書である可能性が高いと考えられる．

さらに実験結果を個別に分析したところ，特に「カバレージ」,「グラフィクス」,「メーキャップアーチスト」,「ケイティーホームズ」といったカタカナ語で，検索ページ数の増加が大きかった．これは，これらの語の異表記として生成された語も標準的に使われている表記であることを意味する．顕著な例として，元のカタカナ語「メーキャップアーチスト」単独での検索ページ数が 1,420 件であったのに対し，この語から生成された異表記「メイクアップアーティスト」の検索ページ数は，2,040,000 件であった（1,437 倍）．また，「グラフィクス」とその異表記「グラフィックス」の検索ページ数はそれぞれ 512,000 件と 12,400,000 件であった（238 倍）．これらの結果は，異表記を考慮しない一般の検索システムでは，相当数の関連情報が検索結果に含まれていないという事実を裏付けている．

続いて，検索結果に対して人手で適合性判定を行い，提案手法 *Phone* とカタカナ書き換え規則に基づく既存手法 *Rule* の定量的な比較実験を行った．具体的には，*Phone* と *Rule* を検索システムとして利用した小規模なプーリングによって，次のように正解データを作成した．まず，前述のカタカナ語 25 語について，*Rule* と *Phone* のそれぞれを用いて異表記を生成し，検索質問置換によりウェブ検索を行った．そして，それぞれの検索結果の上位 20 件について人手で適合性の判定を下した．なお，評価者は 4.1 節と同一の学生 17 名であり，評価対象のウェブページがどちらの手法によって得られた検索結果であるかは知らされていない．

この正解データを用いて，*Phone* と *Rule* の検索結果上位 100 件を評価した．この実験によって得られた R–P 曲線を**図 5** に示す．この結果から，すべての再現率の範囲で提案手法 *Phone* の適合率が従来手法 *Rule* を上回っていることが分かる．特に，再現率 0.2 以上

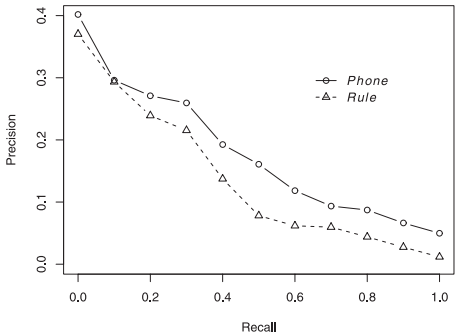

図 5　提案手法（*Phone*）と既存手法（*Rule*）の *R*–*P* 曲線の比較
Fig. 5　Comparison of *R*–*P* curves for our proposed and existing approaches.

でつねに 0.03〜0.08 ポイント程度高い適合率を示した．この結果は，検索結果の上位ではいずれの手法も適合ページをそれほど遜色なく検索できているものの，*Rule* よりも *Phone* の方が正しい（あるいはより頻繁に使われる）異表記を生成しているため，全体としてはより多くの適合文書を検索できていることを示している．

4.3　誤り例分析

4.1 節と 4.2 節における一連の評価実験から，カタカナ異表記生成における提案手法の有効性と情報検索におけるその効果が確認できた．しかしながら，表 5 に示されたように，生成される異表記の中には誤りも少なからず含まれる．また，必ずしも誤りではない場合でも，情報検索に利用した際にシステム性能に悪影響を与えることがあった．以下，これらの代表的な例をあげる．

- **まったくの異義語を生成してしまう例：**「テイスト」の異表記として生成された「テスト」や「シャキラ」（アーティスト名）の異表記として生成された「シャキア」（ゲームのキャラクタ名）など．このように異なる意味を持つ語を生成してしまう例は，特に語長が短い単語に多く見られた．
- **異義語の異表記を生成してしまう例：**「メーデー」の異表記として生成された「メイデー」など．この例は一見正しい異表記を生成しているように思える．しかし，メーデーは元来多義語であり，よく知られている労働祭の意味のほかに，無線電話の国際救難信号の意味もある．本研究の実験では，後者の語義を想定していたのに対し，生成された異表記「メイデー」は前者の語義の異表記として使われることがほとんどであった．その結果，メイデーによって検索質問置換を行ったことによって，前者の語義に関連した文書がほとんど検索されなくなってしまい，情報検索の精度が低下した．

前者の問題に関しては，元のカタカナ語の文字数によって異表記生成処理適用の有無を切り替える，あるいは語長の変化を許さないなどの方法で，ある程度の改善が可能であると考えられる．一方，後者の問題に対応するためには，単語の持つ意味まで考慮した異表記生成および選定を行う必要がある．しかし，元のカタカナ語に多義性がある場合，ユーザの検索意図を検索語のみから推定することは困難であり，当該ユーザの検索履歴や対話的なインタフェースによる多義性解消などの方策が必要である．

5. おわりに

本論文では，日本語と英語の音素対応の不一致に着目し，日本語から英語（音素）への逆翻字，英語（音素）から日本語への翻字を連続的に適用することで，多様なカタカナ異表記を生成する手法を提案した．翻字あるいは逆翻字に関する研究は古くからあるものの，両者を複合的に利用してカタカナ異表記を生成する試みが報告された例はない．異表記を持ちやすいと思われるカタカナ語 25 単語を対象に，提案手法によって生成されたカタカナ異表記を人手で調査したところ，平均 32.5%の異表記に関して肯定的な結果が得られた．また，生成された異表記のうち，66.6%は少なくとも 1 人の評価者によって肯定的に判定されており，このうち 89.2%はカタカナ書き換え規則による従来手法では生成することができなかった．これら肯定的な評価が得られた異表記のみを用いて検索質問拡張を行ったところ，平均 60.1 倍の検索ページが得られた．また，*R*–*P* 曲線による情報検索の性能評価においても，従来手法に比して提案手法の優位性が示された．

今後の課題として，語長などを考慮した異表記の生成，翻字・逆翻字のモデルの精緻化，他言語由来のカタカナ語への適用などがあげられる．

参　考　文　献

1) 武部良明：表記の「ゆれ」，日本語学，Vol.2, pp.43–49 (1983).
2) Brill, E. and Moore, R.C.: An improved error model for noisy channel spelling correction, *Proc. 38th Annual Meeting of the Association for Computational Linguistics*, pp.286–293 (2000).
3) 島津美和子，吉村祐美子，平川秀樹，天野真家：カタカナ異形表記・誤記修正機能の開発・評価，情報処理学会第 44 回全国大会，pp.3–249–250 (1992).
4) 奥村　薫，建石由圭，脇田早紀子，金子　宏：日本語校正支援システム「FleCS」，情

報処理学会研究報告，Vol.87, No.11, pp.83–90 (1992).
5) 獅々堀正幹，津田和彦，青江順一：片仮名異表記の生成および統一手法，電子情報通信学会論文誌，Vol.J77-D-II, No.2, pp.380–387 (1994).
6) 久保田淳市，庄田幸恵，河合眞宏，玉川博文，杉村領一：カタカナ表記の統一方式：予備分類とグラフ比較によるカタカナ表記のゆらぎ検出法，情報処理学会論文誌，Vol.35, No.12, pp.2745–2750 (1994).
7) 久保村千明，亀田弘之：片仮名異表記処理能力を備えもつ情報検索システム，電子情報通信学会論文誌，Vol.J86-D-II, No.3, pp.418–428 (2003).
8) Masuyama, T. and Nakagawa, H.: Web-based acquisition of Japanese katakana variants, *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.338–344 (2005).
9) Knight, K. and Graehl, J.: Machine Transliteration, *Computational Linguistics*, Vol.24, No.4, pp.599–612 (1998).
10) 小泉　保：日本語の正書法，大修館書店 (1978).
11) 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, *Proc. IEEE/WIC/ACM International Conference on Web Intelligence*, pp.110–116 (2004).
12) Frederick, J.: *Statistical Methods for Speech Recognition*, MIT Press (1998).
13) Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K.: Overview of the Web retrieval task at the third NTCIR workshop, Technical Report NII-2003-002E, National Institute of Informatics (2003).
14) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc. (1983).
15) Jones, K.S.: Statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, Vol.28, No.1, pp.11–20 (1972).
16) Voorhees, E.M. and Harman, D.K. (Eds.): *TREC: Experiment and Evaluation in Information Retrieval*, The MIT Press (2005).

(平成 20 年 4 月 17 日受付)
(平成 20 年 6 月 6 日再受付)
(平成 20 年 6 月 27 日採録)

服部　弘幸

平成 20 年神戸大学大学院自然科学研究科情報知能学専攻修士課程修了．同年グーグル株式会社入社．

関　　和広

平成 14 年図書館情報大学情報メディア研究科修士課程修了．平成 18 年インディアナ大学図書館情報学研究科博士課程修了．同年より神戸大学助手（現助教）．情報検索，自然言語処理，機械学習の研究に従事．Ph.D. 電子情報通信学会，自然言語処理学会，ACM SIGIR 各会員．

上原　邦昭（正会員）

昭和 53 年大阪大学基礎工学部情報工学科卒業．昭和 58 年同大学院博士後期課程単位取得退学．同産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同都市安全研究センター教授を経て，現在，同大学院工学研究科教授．工学博士．人工知能，特に機械学習，マルチメディア処理の研究に従事．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．