アルファベット表記とカタカナ表記の対応規則の生成

尾上 徹　　梅村 恭司　　岡部 正幸
豊橋技術科学大学 情報工学系　情報知能工学系　情報メディア基盤センター

1　はじめに

日本人は外国語を表記する際，カタカナ表記に置き換えることが多い．外国語で記述された情報から検索を行う際，調べたい事柄のカタカナ表記だけ知っていたとしても元の綴りを知らなければ検索することは難しい．特に人名の場合には，ある表記に対して複数のカタカナ表記があり，その複数のカタカナ表記全てを網羅的に列挙する辞書を作成することは合理的ではない．そこで，これを解決する手段としてアルファベット表記とカタカナ表記の対応規則（以下単に規則）の自動生成を考える．この規則は外国語のアルファベット表記とカタカナ表記が対応したデータの集合から自動的に取り出す．このとき，アルファベット表記の言語の母音を示す文字以上の知識は用いない．人名の発音記号が入手できればアルファベット表記を用いるよりも正確なシステムが作れると思われるが，入手はコストが高いため，よって，本研究ではこれを用いない．

カタカナによる英単語の検索はこれまでにも行なわれている．例えば，宮内 [1] は英単語の発音記号からカタカナ表記を作り，検索するカタカナの表記ゆれを変換表で解消し，検索を行なった．しかし，この方法は規則の抽出に人手を用いており，発音記号の情報を用いている．発音記号を直接用いずに読みを得る方法として，住吉ら [2] があるが，これも英文字列を変換する変換テーブルを人手により作成する必要がある．

カタカナ表記でアルファベット人名の検索を行う際に重要となることは，検索者が意図したものが含まれるならそれにヒットすることである．このため，入力に一つの人名綴りを対応させる必要はなく，対応する候補集合の中に意図したものが含まれていればよいと考えることができる．よって，本研究では入力に対するマッチングの候補に正解が含まれていることを規則の検索性能とした．そして，カタカナ表記とアルファベット表記を一意に結びつけることは目的とはしないこととした．本研究を実際の検索システムに用いる場合，検索結果に対して，結果を絞り込む何かしらの処理（例えばユーザによる絞込みなど）が加えられることを想定している．

カタカナによるアルファベットの人名検索において，置換は何度も行われるためできる限り単純であることが望まれる．よって本研究では単純な置換による検索でも効果を発揮する規則を生成することを目的とした．

また，本研究は現在の計算機で実現できるかどうかは考慮するが，計算不可を少なくすることは目的としていない．このため，本研究は現在の計算機で実現できる範囲で実験を行い検討した．

人名事典からのアルファベット表記に対応するカタカナ表記の規則の抽出は，増田ら [3] により提案されている．本研究では，増田らの手法 [3] をベースに抽出された規則の改良について扱う．なお規則の評価は，増田らの用いた規則の評価尺度（綴りの復元率，読みの復元率）と，新たに提案する規則の評価尺度（逆綴り復元率，逆読み復元率）によって行なうこととした．その結果，規則の性能が有意に向上したことを報告する．そして，なぜそのような結果となったかについても考察を行なう．

2　増田らの対応規則抽出法

我々がベースとして用いる増田らの手法 [3] は，一つの分割点を発見することにより対応規則を抽出する方法であるといえる．この手法は，人名辞書データを入力とし，アルファベット綴りとそれに対応する読みをそれぞれ二分割することで，アルファベット綴りに対応するカタカナの規則（以降，対応規則もしくは単に規則と表記）を生成する方法である．図1に分割の例を示す．分割に際して，次の二つの日本語知識と，文字列の出現頻度を用いる．

- 母音はカタカナ表記の区切り（変音記号のついた母音も母音とする）
- 促音（ッ）と長音（ー）は語頭に現れない

図 1: 人名辞書からの対応規則の抽出例

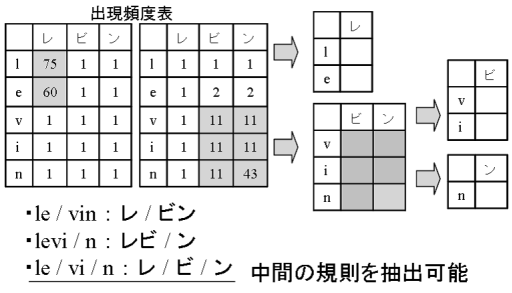

図 2: 人名辞書からの対応規則の抽出例

3 提案手法

3.1 概要

増田らの手法の分割点を発見することによる対応規則抽出方法は人名辞書データに対して，1 つの分割点を定めることで対応規則を取り出すものであった．1 つの分割点を定めることで対応規則を抽出する場合，中間部分の文字列を規則として取り出すことができないという問題がある．これに対して本章で提案する手法は，複数の分割点を定めることにより，より多くの規則を抽出するものである．複数の分割点を定めることで，中間部分の規則を抽出することができ，且つ分割の見込みがある部分で分割を適宜行なうことで，短い規則を複数取り出すことができる．ただし，分割点が 1 つの場合に抽出される規則も，この手法を適応した結果得られる規則に含まれるため，増田らの手法により得られる対応規則集合を部分集合として完全に含んでいる．図 reffig:rdiv に本手法による規則抽出の例を示す．図 reffig:masu と比較すると，中間部分の規則が抽出され，得られる規則が増えていることがわかる．

3.2 規則の抽出

本手法は，外国語に依存した知識を用いずに対応規則を人名辞書から抽出する．扱うデータには英語のみではなくドイツ語，フランス語等複数の言語が含まれ，読み仮名は全て日本語（カタカナ）である．そこで，本手法では増田らの手法と同様に，次の 2 つの日本語の知識を用いる．

1. 母音はカタカナ表記の区切り（変音記号のついた母音も母音とする）
2. 促音（ッ）と長音（ー）は語頭に現れない

本手法は複数の分割点をアルファベット文字列とその読みのカタカナ文字列に定めることでルールを得る．これは，まず一つ分割点を定め，その結果得られる語尾・語頭規則それぞれを繰り返し分割することで実現している．一つ分割点を定める操作は増田らの対応規則抽出アルゴリズムと同様にして実現できる．ただし，規則の再分割には，分割により生成されうる規則全ての出現頻度が必要となるため，これを事前に求めておくことが必要である．この手法は 1 以上の分割点を定めるため，アルファベット文字列が 1 文字もしくはカタカナ文字列が 1 音節の場合，分割できないため規則は生成されない．

4 実験

4.1 規則の抽出と評価に用いるデータ

カタカナとアルファベットの対応規則を抽出する対象として，外国人名のアルファベット表記とカタカナ表記が対応した言語混合の人名事典 [4] より成形したデータ）を用いる．この人名事典は 31847 個の人物名の対応データからなる．このうち 24000 個を無作為に取り出し，これをテストデータとする．そして，残り 7847 個を学習データ（対応規則抽出の対象）とする．この人名事典から無作為に抽出して生成する，テストデータ（対応データ 24000 個）と学習データ（対応データ 7847 個）の組をデータセットと呼ぶこととする．本実験では，このデータセットを 10 個構成し，それぞれについて比較する規則抽出法で規則の抽出を行い，規則の性能の比較を行なう．

4.2 評価尺度

本研究では，増田らの用いた綴りの復元率と読みの復元率という評価尺度と，新たに定める逆綴り復元率

と逆読み復元率という 4 つの評価尺度で，対応規則の性能を測り，対応規則抽出法の優劣を考える．

綴りの復元率 綴りの復元率とは，テストデータの各アルファベット表記を，対応規則集合を用いてカタカナに変換できる（読みの候補を作ることができる）割合である．これは，十分な量の規則が生成できているかを測る．ただし，置き換えられた読みが，元のデータと同じか異なるかは考えない．対応規則の当てはめは先頭から最長一致で当てはめる単純な方法で行った．

読みの復元率 読みの復元率とは，綴りを復元できたデータの集合において，読みが正しかった割合を表す．この比率は規則が生成できた場合に，元データにある正解のカタカナ表記で目的のデータが発見できる確率に相当する．これにより規則の性能，すなわち，実際のデータベース検索における検索可能な確率を測る．読みが正しいとは，アルファベット文字列の読み方を規則から生成し，その生成された読み方の集合に正解（辞書に記された読み方）が含まれている場合である．ただし，辞書に複数とおりの読み方がある（例えば adrian に対してアドリアン，エイドリアンという読み方がある）場合，それらは別々のデータとして扱われるため，adrian-アドリアンのペアに対してエイドリアンという読みしか生成できなければ，そのペアについて不正解として扱われる．

逆綴り復元率 逆綴り復元率とは，テストデータの各カタカナ表記を，対応規則集合を用いてアルファベットに変換できる割合である．綴りの復元率と同様に，置き換えて得られたアルファベット表記が元のデータと同じか異なるかは考えない．また，規則の適応も先頭からの最長一致で綴りの復元率と同様に行なう．

逆読み復元率 逆読み復元率とは，カタカナをアルファベット表記に置き換えられたものの内，その綴りが元のデータ（正解）と等しかったものの割合である．

増田らは読みの復元率を，元のデータにある正解のカタカナ表記で目的のデータが発見できる確率に相当すると述べたが，アルファベットを置換してカタカナを作り，それが正解とマッチングするかという判定方法であるため，これはアルファベット表記でカタカナ表記を検索するシステムに用いる対応規則集合の評価に相当するといえる．このため，カタカナ表記によるアルファベット表記検索システムに用いる対応規則集合の評価には不十分であると考えられる．ゆえに，本研究では逆綴り復元率と逆読み復元率として上記のものを定め，これを評価尺度に加え，上に述べる 4 つの尺度から評価を行なう．

表 1: R と RDiv の評価

	平均		標準偏差	
	R	RDiv	R	RDiv
綴りの復元率	0.949	0.983	0.003	0.004
読みの復元率	0.286	0.355	0.006	0.007
逆綴り復元率	0.902	0.990	0.006	0.001
逆読み復元率	0.261	0.376	0.005	0.004

4.3 実験手順

まず，4.1 節のようにしてデータセットを 10 個ランダムに作成する．

次に比較のために増田らの手法（分割点を発見することによる対応規則抽出方法）[3] を用いて学習データから抽出することで対応規則集合 R を，提案手法である複数の分割点を定めることによる対応規則抽出法により学習データから規則を抽出しすることで対応規則集合 RDiv を作成する．

以上のようにして得られた対応規則集合 R，RDiv それぞれを用いた場合のテストデータ（24000）に対する綴りの復元率，読みの復元率，逆綴り復元率，逆読み復元率を求め，これによりそれぞれの対応規則集合の評価を行なう．

4.4 実験結果

評価の結果，10 個のデータセット全てにおいて，提案手法は 4 つの評価尺度全てで増田らの手法を上回ることを確認した．これに対して，符号検定を行うと全ての評価尺度において増田らの手法よりも提案手法が優れているということが危険率 1%でいえる．10 個のデータセットにおける 4 つの評価尺度の平均と標準偏差を表 1 に示す．

5 考察

実験の結果，提案手法は 4 つの尺度で有意に結果の改善を達成した．特に，逆読み復元率を改善できたことで，提案手法はカタカナによるアルファベット人名検索において増田らの手法よりも優れているということができる．本節ではこのような結果となった要因について例を用いて考える．

増田らの手法で取り出すことができず，この提案手法では抽出できる規則は，元の文字列の語頭及び語

表 2: 文字列”アミアネシス”への R の適応

アミ	アネ	シス		
ami	ynet	sisso	s	sice

表 3: 文字列”アミアネシス”への RDiv の適応

アミ	アネ		シス		
ami	a	ane	sice	s	ssis
	ynet	sis	shis	sisso	

尾を含まない規則である．増田らの手法ではアルファベット・カタカナ対を 2 つに分割するため，語頭規則には文字列の語頭が，語尾規則には文字列の語尾が必ず含まれる．このため，文字列の中間を抜き出した規則を作ることはできない．この規則が R と RDiv の差集合であると考えられ，この差集合が結果改善の要因であったと考えられる．

では，実際にどのようにして改善されたか，改善につながった規則はどのようにして抽出されたか実例を用いて考える．例えば，データセット 1 のテストデータに”アミアネシス amianesis”という対がある．逆読み復元率による評価では，”アミアネシス”に規則を適応して正解の綴り”amianesis”を生成できれば逆読み復元率は大きくなる．この評価において，R，RDiv それぞれにより表 2 ，表 3 のように規則を適応された．この結果，R は正解を作れず，RDiv は正解を生成することができた．

表 2 ，表 3 を比較すると，両集合ともに文字列”アミアネシス”に対してアミ，アネ，シスの規則を適応していることが分かる．R の変換結果を見ると，RDiv では割り当てることができた規則である，”アネ”と”ane”，”シス”と”sis”の規則がないために正解を導けなかったことが分かる．

”アネ”と”ane”の規則について考える．学習データを調べたところ，”アネ”と”ane”が共起するアルファベット・カタカナ対は，”eanes エアネス”という対 1 つしかないことが分かった．この対から”アネ”と”ane”という規則を作ることは，増田らの手法では不可能であるが，提案手法ではこれは可能である．抽出の過程を見ると，”エアネ eane”と”ス s”に分割され，”エアネ eane”がさらに”エ e”と”アネ ane”分割されることで，”アネ ane”規則が抽出されることが分かった．このように，中間部分から取り出される規則が結果の改善に寄与していることが実例からも分かる．

このように複数の分割点を文字列複数回の分割により抽出できた規則集合が結果の改善につながる場合があることがこの実例より分かった．この，R にない二種類の規則集合によって結果が改善されたと考えられる．

6 おわりに

本研究では，複数の分割点を発見することによる対応規則抽出法の提案を行った．そして，アルファベット表記とカタカナ表記の対データ（人名辞書）と，アルファベット表記の母音を示す文字の知識のみから人手を介さずに自動的に得られる規則の性能の改善を行なった．性能の評価は，増田らの評価尺度に，新たに提案する逆綴り復元率，逆読み復元率の二つの評価尺度を加え，計 4 つの評価尺度により評価を行った．

人名辞書データ（対応データ 31847 個）からランダムに 24000 個のテストデータと 7847 個の学習データ（規則抽出対象）を抽出することで作成したデータの組を 10 個作り，実験を行なった．この結果，全ての評価尺度で，10 回のデータセットの内全てにおいて提案手法 RDiv は増田らの手法 R を上回った．このため，この結果には統計的有意差があるといえる．この結果から，カタカナによるアルファベット人名検索においては我々の提案手法が，増田らの手法に比べて優れているということができる．

本研究は，現在の計算機で実現できる範囲で実験を行い，検討した．今後の課題として，速度やメモリ効率といった性能の向上が挙げられる．

参考文献

[1] 宮内 忠信：カタカナ表記からの英単語検索システムの実現, 情報処理学会研究報告. 自然言語処理研究会報告 93(79), 119-126, 1993-09-16

[2] 住吉 英樹, 相沢 輝昭： 英語固有名詞の片カナ変換, 情報処理学会論文誌 35(1), 35-45, 1994-01-15

[3] 増田 恵子, 梅村 恭司：人名辞書から名前読み付与規則を抽出するアルゴリズム, 情報処理学会論文誌 40(7), 2927-2936, 1999-07-15

[4] 星野 裕，加藤 博子，永田 健二： 8 万人西洋人名よみ方綴り方辞典, 日外アソシエーツ（1994）