言語横断検索を目的としたカタカナ-アルファベット対応規則の抽出法

稲葉 祥 *　　足達 花絵 †　　岡部 正幸 ‡　　梅村 恭司 §

1　はじめに

一般的に，人名など外国固有の単語を日本語で表記する場合カタカナで記述することが多い．アルファベット表記の外国語で記載されている情報を検索する際に，カタカナの表記のみを知っていても元のアルファベット表記が分からない場合，検索するのは難しい．特に人名の場合，検索したいアルファベット表記に読まれ得るカタカナ表記が複数存在する場合があり，カタカナ表記をアルファベット表記とすべて対応付ける辞書を作成することは大変膨大な数のためコストが掛かる．

そこで，アルファベット表記とカタカナ表記の単語が対応した対データの集合（教師データ）からアルファベット表記に対応する日本語読みを表したカタカナ表記の規則（対応規則）を自動生成することを考える．対応規則を自動生成する方法として EM アルゴリズムを用いた Jiampojamarn らによる多対多アライメント [2] が提案されている．この手法では対応規則学習時に，より文字数の多い対応規則が有利になる欠点がある．この欠点を改良した久保らの多対多最小パターンアライメントアルゴリズム [3] が存在する．久保らの手法ではもっともらしい対応規則を生成し，生成される対応規則数は多いが，対応規則の字数は少ない．久保らの手法と対抗する一般化された方式はないだろうか．その手法に増田らの対応規則を抽出する手法 [1] がある．増田らの手法では文字列の出現頻度を用いて与えられた対データを分割し正しい対応規則を生成する．生成される対応規則の数は少ないが対応規則に現れる文字数は長い．

久保らの手法と増田らの手法には弱点があり相互に補うことで，この 2 つの手法よりも未知のアルファベット表記に対してカタカナの読みを付与できる対応規則を抽出することを考えた．

本研究では教師データを増田らの手法と，久保らの手法の 2 つの手法を適用して得られた対応規則を用いることで，未知のアルファベット表記集合に対するカタカナ表記の正解候補数がそれぞれの手法のみ適用した場合に比べ向上したことを報告する．

*豊橋技術科学大学 情報・知能工学課程，
inaba@ss.cs.tut.ac.jp
†豊橋技術科学大学 大学院 情報・工学専攻，
adachi@ss.cs.tut.ac.jp
‡豊橋技術科学大学 情報メディア基盤センター，
okabe@imc.tut.ac.jp
§豊橋技術科学大学 情報・知能工学系，
umemura@tut.jp

2　従来法

2.1　増田らの手法

増田らの手法 [1] は与えられた対データのアルファベット表記とカタカナ表記のそれぞれで 1 つの分割点を決定することで，分割点より前部分と後部分で対応付け対応規則として抽出する方法である．分割に際し以下の 2 つの日本語知識と文字列の出現頻度を用いる．以下の日本語知識はカタカナ表記に対する規則である．

- 母音（a,i,u,e,o）はカタカナ表記の区切りとする
- 促音（ッ）と長音（ー）は語頭に現れない

表 1 の例を用いて対応規則の抽出方法を説明する．表 1(a) は対データのアルファベット表記 (adam) とカタカナ表記 (アダン) を語頭からそれぞれ任意文字数取り出して部分文字列の組を生成し，与えられた教師データにおいて生成した部分文字列の組の出現頻度を表している．表 1(b) は同様にして語尾から任意文字数の部分文字列の組を生成し，与えられたデータベースにおいての出現頻度を表している．表 1(a) では「ada」と「アダ」の対応は 1 回出現することを示している．表 1(b) では「m」と「ン」の対応は 20 回出現することを示している．増田らの手法では部分文字列の文字数が変化した際にその前後の出現頻度の比が 1/3 以下の場合分割点とみなし，分割点前後を分割し対応規則

として抽出する．表 1 の例では語頭から部分文字列を取り出していった場合,「a」と「ア」,「dam」と「ダン」の対応規則が抽出され，語尾から部分文字列を取り出していった場合「ada」と「アダ」,「m」と「ン」の対応規則が抽出される．この例の概念図を図 1 に示す．図 1 の ① は語頭から， ② は語尾から走査した時の分割点である．

表 1: 対応規則の抽出例

(a) 左結合

	ア	アダ	アダン
a	18	1	1
ad	2	2	1
ada	1	1	1
adam	1	1	1

(b) 右結合

	アダン	ダン	ン
adam	1	1	1
dam	1	1	1
am	1	1	2
m	1	1	20

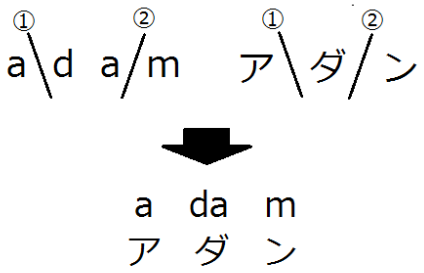

図 1: 増田らの手法の概念図

2.2 久保らの手法

久保らの手法 [3] は EM アルゴリズムを用いた Jiampojamarn の多対多アライメントアルゴリズムを改良したアルゴリズムである．多対多アライメントアルゴリズムでは学習時に 1 以下の乗算回数が少ない文字数の多いアライメントが有利となる問題点がある．久保らの手法はこの問題点を，カタカナ表記とアルファベット表記の文字数の和を乗算回数にすることで，未知語に対する頑健性が失われることを改善した．多対多最小パターンアライメントアルゴリズムと呼ばれている．

アルファベット表記とカタカナ表記の対データを d とし，対データの集合である教師データを D とする．図 2 にアルファベット表記とカタカナ表記の対応付けの概念図を示す．図 2 の場合状態 ①， ⑥， ⑮， ⑳ が選択され,「a」と「ア」,「da」と「ダ」,「m」と「ン」が対応していることを示す．この表記と読みのパターンを表し個々の状態遷移を表す変数を u と定義する．図 2 から対データ d はパターン u の系列により表されていると考えることができる．この系列をパターン系列と呼び，変数を $\boldsymbol{u}$ とする．対応規則を抽出する際はデータ d において考えられるすべてのパターン系列 $\boldsymbol{u}$ を考慮する．その系列の集合を $\boldsymbol{U}$ と定義する．

多対多アライメントでは与えられた対データ d の正しいパターン系列 $\boldsymbol{u}$ を推定するために，各パターン u の出現確率パラメータ p_u を EM アルゴリズムにより推定する．p_u を更新前のパラメータ，$\hat{p}_u$ を更新後のパラメータとすると，E ステップでは

$$\gamma_{\boldsymbol{u}} = \frac{\prod_{u \in \boldsymbol{u}} p_u^{n_u}}{\sum_{\boldsymbol{u} \in \boldsymbol{U}} \prod_{u \in \boldsymbol{u}} p_u^{n_u}} \tag{1}$$

を計算する．$\gamma_{\boldsymbol{u}}$ はパターン系列を $\boldsymbol{u}$ とした時の尤度である．n_u は総文字数，i_u はアルファベットの文字数，j_u はカタカナの文字数であり，

$$n_u = i_u + j_u$$

の関係である．また，M ステップでは

$$\hat{p}_u = \frac{\sum_{\boldsymbol{u} \in \boldsymbol{U}_u} \gamma_{\boldsymbol{u}}}{\sum_{u \in u_{all}} \sum_{\boldsymbol{u} \in \boldsymbol{U}_u} \gamma_{\boldsymbol{u}}} \tag{2}$$

を計算する．u_{all} はパターン u の全種類の集合，$\boldsymbol{U}_u$ は u が出現するパターン系列 $\boldsymbol{u}$ の集合である．この E ステップと M ステップを Forward-Backward アルゴリズムを用いて計算し，パラメータ値が収束するまで繰り返す．そして，推定したパラメータ $\hat{p}_u$ を用いて，与えられた対データ d の尤も正しい $\boldsymbol{u}$ と判断されたパターン系列 $\hat{\boldsymbol{u}}$ は Vitabi アルゴリズムにより推定される．推定されたパターン系列 $\hat{u}$ より対応規則を抽出する．図 2 の例では「a」と「ア」,「da」と「ダ」,「m」と「ン」の対応規則が抽出される．

3 提案手法

本研究では言語横断検索においてより正解候補が多く含むために，教師データから増田らの手法によって対応規則を抽出し，抽出された対応規則を入力として久保らの手法に与え対応規則を得る手法を提案する．

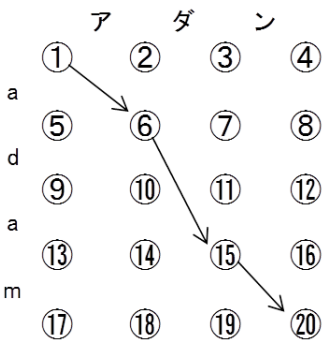

図 2: 多対多アライメントの概念図

久保らの手法は必ずしも正しい対応規則ではないことに着目し，増田らの手法によって正しい分割箇所で予め分割する．図 3 に提案手法の概念図を示す．

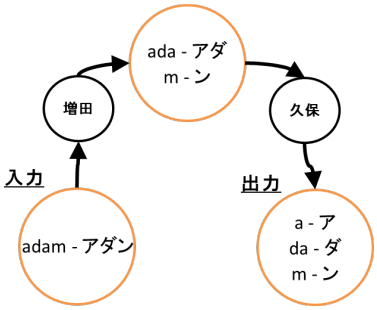

図 3: 提案手法

4 評価対象

対応規則の抽出に用いる教師データは本評価では人名のアルファベット表記とカタカナ表記が対応して記載された人名辞書を用いて行う．教師データは 29912 個の対データから成る．教師データから 90%を対応規則抽出の対象（学習データ）として無作為抽出し，残りの 10%をカタカナ表記の分からない未知語（テストデータ）とする．学習データとテストデータを合わせてデータセットとして構成する．データセットは同じ人名辞書から 10 個生成する．

5 評価方法

4 章で生成した 10 個のデータセットごとに対応規則を生成し，テストデータから正解が候補に含まれるか交差検証を行った．本研究ではあるテストデータに対して検索者が意図したカタカナ表記が含まれれば良いとした．アルファベット表記の人名入力に対して 1 つの表記を対応させる必要はなく，候補集合の中に意図したカタカナの表記が含まれていれば良いと考えることができる．以下に正解例と不正解例を示す．

正解例

テストデータ	（anselmo アンセルモ）
対応規則	（ansel アンセル）（mo モ，リモ）
候補	**アンセルモ**，アンセルリモ

不正解例

テストデータ	（batmunkh バトムンフ）
対応規則	（bat バト）（kh ク，ハ）
	（mun マン, マンゼ, ムン）
候補	バトマンク，バトマンハ
	バトマンゼク, バトマンゼハ
	バトムンク，バトムンハ

また提案手法，久保らの手法，増田らの手法によって生成された対応規則が確からしいか 3 個のデータセットを構成し，それぞれの手法で対応規則を生成した．それぞれの手法で生成した対応規則から無作為に 100 個の対応規則を取り出す作業を 3 回行った．合計 900 個の対応規則について，その対応規則が確からしいかどうか主観評価した．正解例と不正解例を以下に示す．

正解例

（t，ト）（n，ン）(kwen，クェン)

不正解例

（o，ア）（ham，ム）（walter ，レイモンド）

6 評価結果

5 章の方法でテストデータに対して正解が含まれているか評価を行った結果を表 2 に示す．また生成された対応規則が確からしいかどうか評価した結果を表 3 に示す．

7 考察

表 2 に示すように 10 個のデータセット全てにおいて増田らの手法，久保らの手法より提案手法が上回

表 2: 評価結果

	提案手法 (%)	久保ら (%)	増田ら (%)
1	95.6	94.4	61.9
2	95.2	93.9	61.9
3	95.6	93.9	62.6
4	95.4	94.2	62.8
5	95.6	94.0	62.7
6	95.1	94.4	61.4
7	95.3	93.5	64.3
8	95.7	93.9	62.5
9	95.1	94.1	63.6
10	95.4	93.8	61.2

表 3: 対応規則の正答率

	提案手法 (%)	久保ら (%)	増田ら (%)
1	14	17	66
2	12	15	61
3	13	19	69

る結果となった．符号検定を行うと提案手法が久保らの手法，増田らの手法と比べ優位であることが危険率 1%でいえる．

$$P = \frac{{}_{10}C_0}{2^{10}} = 0.0009765625 < 0.01$$

久保らの手法で用いられている EM アルゴリズムでは一度分割する箇所を誤るとそのまま誤ったまま学習が進んでしまうことがある．増田らの手法で予め出現頻度に大きな差がある強い分割点で確実に分割した対応規則にしておくことで単独の手法より向上したことが考えられる．

表 3 に示す正答率は，再現率の向上とは逆に下がっている．今回の規則は最初の検索で利用し，誤りがあっても漏れがないことを重視している状態での利用を考えており，はっきりと再現率のほうが正答率よりも重要であるため，正答率が低下するのと引き換えに再現率を高めることは価値がある．また，もし久保らの手法がパラメータを調整して，再現率と正答率のトレードオフができる手法であれば，調整結果との比較が必要であるが，久保らの手法はそのようなことができないため，提案手法で再現率が向上したことは意味があると考えられる．

8 終わりに

本研究では教師データを増田らの手法を前処理とて適用し，得られた出力を久保らの手法に適用し 2 つの手法を組み合わせた手法を提案した．教師データから提案手法で得られた対応規則を用いることで，2 章で示したように 10 回の分割交差検証した結果，10 回とも提案手法が増田らの手法，久保らの手法のそれぞれの手法より向上したた．符号検定を行うと提案手法が増田らの手法，久保らの手法と比べ再現率を重視するという観点からは，危険率 1%において優位であることが言えた．今後は言語横断検索をできるようなシステムで動作させて，正答率の低下をカバーするような検索が実現できるかを検証することが課題である．

謝辞

提案方法については，NTT 研究所 笠原 要氏との議論において，難読名のよみ付与を検討した結果がもととなっています。この結果をふまえ，名前と読みから，カタカナとアルファベットに関係を移して，言語横断検索という別のタスクで再現率を重視する設定をし，検討し実験したものが本報告です。笠原氏との有用な議論に深く感謝いたします。

また、本論文執筆にあたって日立製作所中央研究所 塩野谷 友隆氏に有益なコメントを頂きました。感謝いたします。

参考文献

[1] 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) pp.2927-2936，1999

[2] Sittichai Jiampojamarn, Grzegorz Kondrak and Tarek Sherif, Applying Many-to- Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion, Proceedings of NAACL HLT 2007, pp.372-379, 2007

[3] 久保慶伍，川波弘道，猿渡洋，鹿野清宏，多対多最小パターンアライメントアルゴリズムの提案と自動読み付与による評価，情報処理学会研究報告，2010-SLP-85 No.16，pp.1-6 ，2011