# 人名のモノルビ自動付与手法

宮崎 太郎　　加藤 直人
NHK 放送技術研究所
{miyazaki.t-jw, katou.n-ga}@nhk.or.jp

## 1　はじめに

文字につけるふりがなのことをルビという．ルビには文字 1 文字ごとにふりがなをつけるモノルビと，文字列全体にふりがなをつけるグループルビとがある [1]．例えば「一ノ瀬（イチノセ）」はモノルビで表すと「一(いち)ノ瀬(の)(せ)」となり，グループルビで表すと「一ノ瀬」となる．モノルビでは文字とふりがなの対応関係が明確になるので，日本語学習者にとっては便利である．実際，幼児や小学生向けの書籍や，留学生向けの日本語学習テキストでは，モノルビが振られていることが多い．また，モノルビは，我々が現在研究を進めている，ニュースや気象情報を対象とした日本語固有名詞の手話への翻訳でも重要な情報となる．例えば，日本語の固有名詞「中野（ナカノ）」を手話に翻訳すると，手話単語 {中} [1] と指文字[2] {ノ} で表現されるが [2]，このように手話に翻訳するためには，「中(なか)野(の)」のようなモノルビの情報が必要である．NHK のニュース原稿では，固有名詞には読みが付与されているが，通常は，読みやすさを考慮して「中野（ナカノ）」のように書かれる．このため，ニュース原稿を手話に翻訳しようとする場合，固有名詞の漢字表記と読みから，モノルビの情報を推定することが必要である．

本稿では，人名の表記と読みを入力し，そのモノルビを自動付与する手法について述べる．普通名詞であれば，そのモノルビをあらかじめ辞書に登録しておくことが可能であるが，固有名詞は数が多く，すべてを辞書に登録しておくことは困難である．しかしながら，固有名詞，特に人名では大規模なコーパスが存在するので，それを使うことでモノルビを精度よく推定することが期待できる．我々の提案手法でも，人名コーパスを利用しているが，さらにコーパスに出現しない場合にも対処している．

[1] この {} で括られた部分は手話の 1 単語を表す．
[2] 指文字は指の形を使って日本語の 50 音を表現する方法である．

表 1: 学習データの例

| 表記 | 読み |
|---|---|
| 康/弘 | ヤ/ス/ヒ/ロ |
| 三/ツ/藤 | ミ/ツ/フ/ジ |
| 長/久/保 | ナ/ガ/ク/ボ |

## 2　モノルビの自動付与手法

モノルビを付与するには，表記の漢字[3] と読みのカナとを対応付ければよい．本章では，その自動対応付けの手法について述べる．

### 2.1　DP マッチングによる対応付け（手法 DP）

漢字とカナとの対応付けは，漢字-カナ間の尤度に基づいた DP マッチングにより解くことができる．また，その尤度は，学習データとして人名コーパスを用いて推定することができる．学習データの例を表 1 に示す．ただし，尤度を推定する際には，漢字 1 文字とカナ 1 文字間のものだけでなく，漢字 1 文字とカナ複数文字間のものも対象とした．例えば，図 1 の「茂木」の例であれば，「茂（モ）」だけでなく「茂（モテ）」の場合も尤度の推定を行った．一方，漢字複数文字とカナ間の尤度は対象としていない．例えば「茂木（モテギ）」の尤度は用いない．

### 2.2　未学習データへの対応

DP マッチングでは，漢字-カナ間の尤度が推定できない場合，すなわち，未学習データの場合には対応付

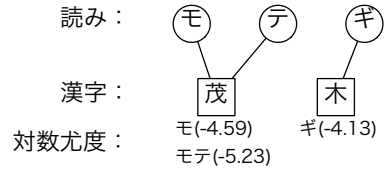


図 1: 対応付けの例

[3] 実際には人名にはひらかなやカタカナも使われるが，ここでは便宜上「漢字」と呼ぶ．

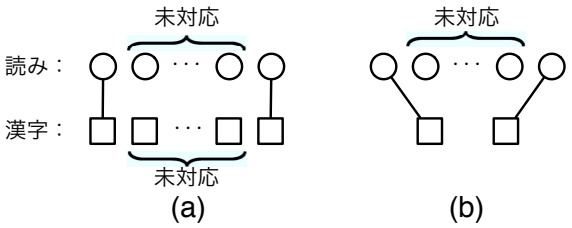

図 2: 未対応の 3 つの場合の例

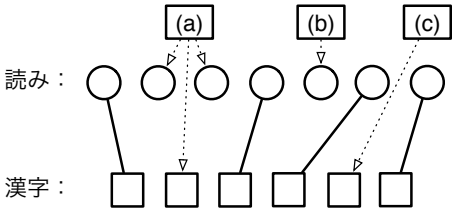

図 3: (a)〜(c) が混在する場合の例

けが行えない．もちろん未学習データに対して，なんらかの尤度を定義すれば DP マッチングを行うことができるが，その尤度を適切に定義するのは難しい．また，そもそも DP マッチングでは，前後の文字の情報，すなわち文脈を扱うことができない．

提案手法では，DP マッチングで対応付けができなかった場合に，その部分解に対して，文脈を利用して別途処理している．ここで，対応付けができない文字列とは，図 2 に示すような，次の 3 つの場合に分けられる．

(a) 漢字・カナの双方に未対応のものがある場合

(b) カナに未対応のものがある場合

(c) 漢字に未対応のものがある場合

これらは 1 つの人名に混在して現れる．例えば図 3 のようになる．

提案手法ではこれらのそれぞれの場合について，個別に処理を行う．提案手法の手順は以下のようである．

**step 0** DP マッチングによる対応付けを実施

**step 1** step 0 で得られたすべての部分解に対し，ペナルティを考慮してリランキングする

**step 2-a** 先頭の漢字から順に (a) に対する処理を実施

**step 2-b** 先頭の漢字から順に (b) に対する処理を実施

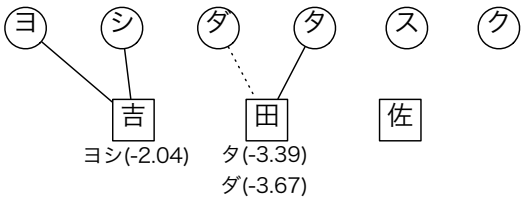

図 4: ペナルティ付与の例

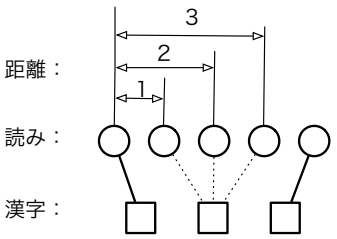

図 5: ペナルティを付与する場合

**step 2-c** 先頭の漢字から順に (c) に対する処理を実施

step 0 の DP マッチングによる対応付けは 2.1 節で述べたものである．ここで解が得られればその結果を出力し，解が得られなければ step 1 を実行する．step 1 では DP マッチングの部分解のうち，未対応となる漢字が最も少ないものを選び，それが複数ある場合は尤度和が最大になるものを対象とする．ただし，尤度和は 2.3 節で述べるペナルティの付与を行なって再計算する．

step 2-a〜c は，未対応の 3 つの場合の（a）〜（c）に対する処理である．以下では step 1 のペナルティ付与と，step 2 の各処理について説明する．

## 2.3 ペナルティの付与 (手法 PEN)

漢字と読みの対応付けは，前の方にあるものから行いたい．しかし，DP マッチングを行うと，図 4 の例では，「田」には尤度が高い「タ」が対応付けられてしまう．

この問題に対応するために，前からの距離に応じたペナルティを付与する．図 5 のように，複数のカナの候補がある場合，直前に対応付けられたカナと候補の間の距離を計算し，$(\text{距離} - 1) \times \alpha$ のペナルティを与える．今回は，予備実験により，$\alpha$ を 0.5 とした．

図 4 の例であれば，「田（タ）」は，直前に対応付けられたカナ「シ」からの距離が 2 になるので，0.5 のペナルティを与え，新しい尤度は-3.89 になる．一方，「田（ダ）」は距離が 1 なのでペナルティは 0 になり，尤度は-3.67 のままである．すると，「田（ダ）」の方が尤度が高くなるため，「田」には「ダ」が対応付けられる．

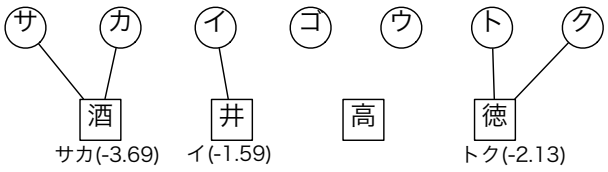


図 6: 漢字，カナの双方に未対応がある例

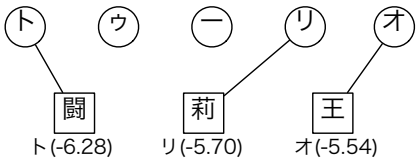


図 7: カナに未対応がある例

## 2.4 未対応部分の処理 (手法 UNK)

### 2.4.1 漢字，カナの双方に未対応のものがある場合の処理（step 2-a）

(a) の場合は，一般的には漢字側も複数の未対応ができるが，実際には 1 つの場合が多い．そこで，今回は漢字 1 つの場合のみを対象とした．

図 6 の例で説明する．この場合，漢字「高」とカナ「ゴウ」がそれぞれ未対応となっている．そこで，未対応の「高」と「ゴウ」を対応付ける．

### 2.4.2 カナに未対応のものがある場合の処理（step 2-b）

(b) の場合は，まず，前後の対応付けられた漢字-カナの尤度を比較する．その結果，尤度が高い方の対応付けは，その対応付けの確度が高いためそのまま残し，尤度が低い方の対応付けに未対応のカナを加える．

図 7 の例で説明する．この場合はカナ「ゥ」「ー」が未対応となっている．前後の対応付けられた漢字-カナの尤度（「闘（ト）」が-6.28,「莉（リ）」が-5.70）を比較すると，「闘（ト）」の方が尤度が低い．そこで，「闘」に「ゥ」と「ー」を対応付ける．その結果，「闘（トゥー）莉（リ）王（オ）」となる．

### 2.4.3 漢字に未対応のものがある場合の処理（step 2-c）

(c) の場合には，まず，前後の対応付けられた漢字-カナの尤度を比較する．その結果，尤度が高い方の対応付けは，その対応付けの確度が高いためそのまま残し，尤度が低い方のカナを分割し，一部を未対応の漢字に仮に対応付けるとともに，既対応の漢字からは除く．これにより新たにできた漢字-カナの対応付けのうち，どちらか一方でも尤度が計算できれば，その対応付けを採用する．どのように分割しても尤度が計算できない場合には，複数の漢字をまとめて読みと対応付ける．

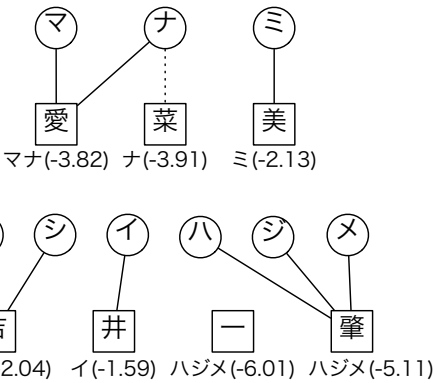


図 8: 漢字に未対応がある例

図 8 の「愛菜美 (マナミ)」の例で説明する．DP マッチングの結果，漢字の「菜」が未対応で残る．「菜」の前後の漢字-カナの対応付けの尤度を見ると，前にある「愛（マナ)」の方が低い．そこで，「愛」に対応付けられたカナである「マナ」を「愛」と「菜」に分割して仮に対応付ける．その結果，「愛（マ）菜（ナ)」となるが，「菜（ナ)」は，学習データ中に存在する対応付けなので，この分け方は採用される．また，これ以上カナを分割することができないので，ここでは「愛（マ）菜（ナ）美（ミ)」が結果となる．

図 8 の「吉井一肇（ヨシイハジメ)」の例では，DP マッチングの結果，漢字の「一」が未対応として残る．「一」の前後の漢字-カナの対応付けの尤度を見ると，後ろの「肇（ハジメ)」の方が低い．そこで，「肇」のカナである「ハジメ」を分割して「一」と「肇」に対応づける．この場合，「一（ハ）肇（ジメ)」と「一（ハジ）肇（メ)」が考えられるが，このいずれの場合も学習データに存在する対応付けではない．そのため，この場合は複数の漢字をまとめて「一肇（ハジメ)」と対応付ける．

# 3 評価実験

提案手法の有効性を確認するために，評価実験を行った．評価実験では，2 章で述べた手法 DP のみを使った対応付けをベースラインとし，手法 UNK，手法 PEN の効果を確認する．

## 3.1 実験条件

文字とカナの対応付けの学習データには，IPADIC-2.7.0 の人名辞書を用いた．学習データは 34,202 対の

表 2: 評価実験結果

| 手法 | 不正解数 | 正解率 (%) |
|---|---|---|
| DP | 536 | 94.44 |
| DP+UNK | 77 | 99.20 |
| DP+PEN+UNK | **72** | **99.25** |

表記-読み対で構成されている．学習データからの漢字とカナの対応付けの学習には giza++[4] を用いた．

評価データは 2010 年 1 月から 2012 年 3 月までの NHK のニュース原稿から抽出した 9,633 の人名を用いた．ニュース記事からの人名抽出には CaboCha[5] を用いた．NHK のニュース原稿では，人名には必ず読みが付与されているため，その読みを使い，人手で正解データを作成した．評価データの人名は名字だけでなく，フルネームのものや，名前のみのものも含まれている．

### 3.2 評価実験結果

評価実験の結果を表 2 に示す．表中の DP，PEN，UNK は，それぞれ 2.1 節，2.3 節，2.4 節の手法を表す．ベースラインとなるのはを DP のみを用いた手法（DP）であり，提案手法は 3 つの手法を組み合わせた手法（DP+PEN+UNK）である．正解率は単語に現れるすべての漢字-カナについてモノルビの付与に成功した単語の割合である．

表 2 を見ると，手法 UNK は，単独で手法 DP に加えても性能の向上に効果があることがわかる．また，2 つの手法を組み合わせて使うことで，さらに性能が向上することがわかる．提案手法では対応付けの正解率が 99.25%と，高い性能を得ることができた．

## 4 考察

手法 DP+UNK でモノルビ付与に成功し，提案手法で失敗した単語は一つのみであった．「池本賞治（イケモトショウジ）」の例である．これは，「賞」の文字が学習データに現れず，尤度が計算できなかったことと，「治」の文字の読みの候補に「ジ」と「ウジ」があり，その 2 つの尤度差が小さかったことにより，ペナルティ付与を行うと「治」に「ウジ」のカナを対応付けてしまったことが原因であった．

提案手法でモノルビの付与に失敗したものの中には，「羽生（ハニュウ）」のように，人手でも漢字とカナの対応付けを明確に行うのが難しいものがあった．このような単語は，そもそも正解を作ることが難しい．今回は「羽生（ハニュウ）」を正解としているが，提案手法の出力は「羽（ハ）生（ニュウ）」であった．このように，2 つの文字の読みが融合したような読みを持つ単語などは，人手で辞書に登録する必要があるだろう．

また，他の失敗した例として，「山内（ヤマノウチ）」のように，2 つの漢字の間に「ノ」などが挿入されたような読みを持つものがあった．今回は，「山（ヤマ）［ノ］内（ウチ）」を正解として評価した．ここでの［ノ］は，「ノ」のカナがどの漢字にも対応付けられていない状態を表す．しかし，今回の提案手法では，未対応のものがある状態では終わることができないため，正解を出力することができない．2 つの漢字の間に「ノ」の読みが入る場合の処理については，個別に対応する必要があるだろう．

他には，「経（ノブ）惟（ヨシ）」のように，学習データ中に現れない未知の読みが一つの単語に複数現れるものは，未対応の解消ができず，対応付けに失敗することが多かった．これらについては，学習データを拡充するなどして対応する必要がある．

## 5 おわりに

本稿では，人名を対象として，モノルビを自動で付与する手法について述べた．文字の文脈を利用することにより，人名を対象としたモノルビの自動付与の正解率は 99.25%となった．DP マッチングによる手法との比較では，86.5%の誤り率の削減ができた．

今後の課題は，人手でも漢字とカナの対応付けが難しいような単語について，単語全体で一つの読みを与えるようなルールを作成することや，wikipedia などから，大規模な人名コーパスを作成し[6]学習データとすることが挙げられる．

## 参考文献

[1] 一般社団法人 日本電子書籍出版社協会，“組版表現説明書，” Mar. 2012.

[2] 宮崎太郎，加藤直人，金子浩之，井上誠喜，梅田修一，清水俊宏，比留間伸行，長嶋祐二，“日本語から手話への地名の機械翻訳，” 言語処理学会第 18 回年次大会, E3-1, pp665–668, 2012.

[4] http://code.google.com/p/giza-pp/
[5] http://code.google.com/p/cabocha/
[6] wikipedia 日本語版には 20 万〜30 万程度の人名が登録されており，その多くに読みが付けられている．