2009 年度卒業研究概要

日本人名を対象とした名前読み規則導出システム

大谷 紀子 研究室
0632063 岸田武朗

1. 研究の背景と目的

日本人の名には，非常に多くの漢字表記が用いられ，読み方も様々なものが存在し，日本人の名の種類は約 40 万近くになるといわれている.日本人が名の読み方を付ける際に関して法的な制限は特に存在していない.ゆえに,漢字表記に対して関連性のない当て字と呼ばれる読み方も許容されるため,自由性に富んでいることが特徴であると考えられる.

新生児の誕生につれ，これまで名に使用される機会のなかった漢字表記と読み方とを伴った名が増え続けており，日本人の名は複雑かつ多様化の様相を呈している．複雑多様化の傾向の中で，日本人の名の漢字表記を見て正しい読み方を導き出すのは容易ではない．先行研究において，英語圏人名のアルファベット表記と読みであるカタカナ表記との 2 つの情報を持つ日本人向けの人名辞書を使用して，アルファベット表記に対応した読み方の規則を導出するアルゴリズムが提案されている[1]．発音記号のような音情報を用いず，特定の言語に依存しない手法である.

本研究では，先行研究のアルゴリズムを利用し，日本人の名の漢字表記に対して，もっとも読まれうる一般的な読み方を導き出すことを目的とする．与えられた日本人の人名辞書から，漢字表記と読み方であるカタカナ表記との対応規則を自動的に導出するシステムを構築する.

2. システムの処理方法

本システムは日本人の人名辞書をデータベースとし，漢字表記の部分文字列とカタカナ表記の部分文字列の組の出現頻度の変化を用いて，漢字表記とカタカナ表記の対応規則を導出する．多くの対応規則を導出するため，漢字表記文字列とカタカナ表記文字列をそれぞれ分割した語頭文字列の組と語尾文字列の組に分けて処理する.

対応規則を抽出する際に使用される 2 つのパラメータを以下に記述する.

① 信頼限度（R）：信頼にいたる出現頻度の下限値

② 閾値（T）：変化点における出現頻度の比の上限値

対応する漢字表記とカタカナ表記の部分文字列の組が出現した数を出現頻度数とする.出現頻度数が信頼限度(R)=7 より大きく，次に出現する部分文字列の組の出現頻度数との比が閾値(T)=1/3 以下のときに対応規則が導出される.

表.1 部分文字列組の出現頻度数

語頭	ケ	イ	ジ
啓	**27**	**26**	2
ニ	2	2	2

語尾	ジ	ロ	ウ
ニ	24	24	24
朗	46	**342**	**342**

処理の例として，(啓ニ朗，ケイジロウ)のとき，部分文字列の組の出現頻度数を表 1．に示す．そ

れぞれの文字列が分割した際に語頭と語尾が同じ文字数になるようにすると,語頭の文字列の組が(啓二，ケイジ）語尾の文字列の組が（二朗，ジロウ）となる．表 1.の語頭における部分文字列の組（啓，ケイ）の出現頻度数は 26 で，次に出現する部分文字列の組（啓，ケイジ）の出現頻度数は 2 である．出現頻度数の比は 2/26 で閾値(T)=1/3 以下となり，出現頻度数は信頼限度(R)=7 以上で 2 つの条件を満たす．よって（啓，ケイ）が対応規則として導出される．（啓，ケイ）が対応規則ならば全体の文字列の組（啓二朗，ケイジロウ）から差し引いた残りの（二朗，ジロウ）も対応する読みの規則としてみなすことができ，対応規則として導出する．語尾文字列の組においても同様の処理を施し，対応規則は（朗，ロウ）と（啓二，ケイジ）となる．

3. 評価実験

日本人の人名 23975 個をデータベースとして対応規則を導出する．本研究で用いたパラメータの値は先行研究とのデータベースの量の違いから信頼限度（R）=10 のところを（R）=7 としたが，閾値に関しては先行研究のパラメータと同じ（T）=1/3 を用いてシステムを動作させた．

語頭の処理から得られた対応規則が 4389 個，語尾の処理から 2320 個の対応規則が得られた．対応規則の精度を確かめるため男性と女性の人名をそれぞれ 100 個用意し，人名の漢字表記が入力された際に対応規則から読みの再現を試みたところ，正答率は男性の人名が 61%，女性の人名が 72%で，平均 66.5%であった．また，本人の読み方と違っていても実際に読む機会のある読み方を含めると，正答率は男性の人名 72%，女性の人名 82%で平均 77%の精度が得られた．

4. 考察

先行研究において読みの正答率は 48.8%であり，本研究は先行研究の正答率を大きく上回ることとなった．日本人名の漢字表記文字列は，先行研究の対象としたローマ字綴りを用いる外国人名のアルファベット表記文字列より短いことが要因として考えられる．また，先行研究で導出された語頭と語尾の対応規則はそれぞれ 11696 個，8608 個であった．先行研究よりデータベースの量が少ないとはいえ対応規則の導出数に大きな差が出ていることから，日本人の人名は表記の数こそ多いが英語圏の人名と比べた際に，読み方は使われやすい読みに収束する性質を持っていると考えることができる．

本システムによって導出された対応規則の正答率は先行研究を上回ったものの，全面的には信頼しがたく不安を残す結果となった．正答率を下げた原因として，人名に用いられやすい漢字表記の中で，（裕，ヒロ），（裕，ユウ）や（俊，シュン），（俊，トシ）のように複数の読み方を持つものが多く存在することがわかった．両方の読み方とも人名に用いられるケースが多く，どちらか一方の読み方に偏らない．したがって，複数の読み方を持つ漢字表記において，使われやすい読み方も複数あるという場合には本研究の手法は相応しいとはいえない．さらに一部の漢字表記において（佳，ヨシ），（佳，カ）のように男性，女性の性別によって用いられる読み方の異なるケースが存在した．

結果として本研究は多くの課題を露呈することとなった．しかし，未だ改善の余地を残している．人名のデータベースに新たに性別の情報を加え，性別ごとに望まれやすい対応規則の導出を可能にするなどして，高い有用性を示すためにはさらなる改良が必要である．

参考文献

[1] 増田恵子，梅村恭司，“人名辞書から名前付与規則を抽出するアルゴリズム,” 情報処理学会論文誌，Vol．40，No．7，pp．2927-2935，1999.