検索エンジンを用いた人名の読み仮名抽出

Acquisition of Kana Person Names from Using Web Search Engine

酒巻智宏[1]　丹英之[2]　大向一輝[3]　武田英明[3,1]
Tomohiro SAKAMAKI[1]　Hideyuki TAN[2]　Ikki OHMUKAI[3]　Hideaki TAKEDA[3,1]
[1]東京大学　[2]株式会社アルファシステムズ　[3]国立情報学研究所
[1]The Univercity of Tokyo　[2]Alpha Systems inc.　[3]National Insutitute of Informatics

1. はじめに

近年、SNSやミニブログのように人にフォーカスしたサービスが注目されている。人物を扱う際に人名は重要な属性である。日本人の人名については、漢字名はWeb上で現れることが多いが、その読みがわからないことがある。これらを踏まえて、本研究では漢字名と共に振り仮名が与えられているような情報を効率的に抽出することを目的とする。

2. パターンマッチングによる人名読み候補の取得

まず、検索エンジンから人名読み候補が取得できるかどうかについて検討する。テストデータは、Wikipedia[1]に存在する人物の漢字名と読みのペアを2817件用意した。

次に、どのようなパターンで検索結果スニペットから人名読みを抜き出すことができるのかを調べる。テストデータを用いて、「人名漢字 and 人名読み」というクエリで検索を行い、検索結果のスニペットを取得する。このとき、検索エンジンとしてGoogle Ajax Search API [2]を用い、29516件のスニペットを取得した。

スニペットから人名読み候補を抜き出すためには、計4つのパターンが必要になる。

田中 [1] 太郎 [2] たなか [3] たろう [4]　(1)

式1で、それぞれ[1]が「漢字の苗字と名前間のパターン」、[2]が「読みの苗字と名前間のパターン」、[3]が「漢字と読み間のパターン」、[4]が「読み終端のパターン」となる。これらの位置に、どのような記号や文字列が現れるかを出現頻度順に並べたものを、表1に示す。

表1: 出現するパターンの例

[1]	出現数	[2]	出現数	[3]	出現数	[4]	出現数
なし	16395	space	12592	（	11284	）	10821
space	12474	なし	12157	,	3184	)	3931
space3	248	・	3287	(	2812	、	3677
…	…	…	…	…	…	…	…

[1][3]には、何も入らない場合やスペースが一つだけ入る場合が上位に、[2]には、"("や"（"といった記号が、[4]には、")"や"）"、"といった記号が上位に現れるという結果になった。

ここまで人名読みを抽出するパターンを取得してきたが、パターンを使えば使うほど、人名読み以外の文字列を抜き出してしまう可能性も高くなる。ここで、パターン数による人名読みの抽出の性能評価を行う。なお、パターンは、出現頻度順に並べたものを用いる。評価基準は、適合率 (precision) を用いた。

表2: パターンの使用数と適合率の関係

パターンの使用数	6	7	8	9	**10**	11	12
適合率 (%)	90.6	92.1	92.9	92.8	**93.2**	92.3	92.0

表2によれば、パターン数が10の時が最も高い適合率で人名読みを抜き出すことができることがわかる。

3. DPマッチングによるフィルタリング

人名読みを取得する最初の段階として、人名漢字をクエリとして検索エンジンで検索を行う。検索エンジンには、Yahoo! Search BOSS [3]を使用した。本研究では、検索結果の上位500件を解析対象とする。得られた検索結果から、式1の形式によって人名読み候補を抜き出していく。

取得された読み候補の中には、明らかに人名読みでないものが含まれる。本研究では、DPマッチング [2] を用いてフィルタリングを行う。先行研究 [1] を参考に、DPマッチングのパラメータの決定やスコア付けを行う。

はじめに、表3のように、人名漢字を漢和辞典、人名辞書を用いて分解し、これらを組み合わせて、その人名漢字の辞書的に可能な読み方を列挙する。なお、本研究では、漢和辞書として「Infoseekマルチ辞書漢和辞典」[4]。人名事典として日外アソシエイツの「苗字8万よみかた辞書」と「名前10万よみかた辞典」を用いた。

表3: 人名漢字から生成される文字列

	田	中	太	朗
1	た	なか	た	ろう
2	でん	なか	た	ろう
…	…	…	…	…
25	たなか		ふと	ろう
…	…	…	…	…

次に、表3で得られたすべての可能な読みとスニペットから得られた読み候補に対してDPマッチングを適用

[1] http://ja.wikipedia.org/
[2] http://www.google.com/
[3] http://www.yahoo.com/
[4] http://dictionary.infoseek.co.jp/

する。一つの読み候補について、辞書的に可能な読み方すべてと DP マッチングを行い、そのうち最もスコアが良かったものをその読み候補のスコアとする。

また、DP マッチングによるフィルタリングを行う際に、スコアの閾値を決める必要がある。

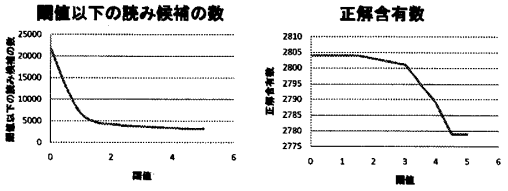

図 1: 閾値による DP マッチングの結果の変化

図 1(左) は閾値を変化させた際の閾値以上の読み候補の数の変化をグラフにしたものである。また、図 1(右) は閾値以下の読み候補中に正解読み候補が含まれている数の変化をグラフにしたものである。図 1 を見ると、閾値が 2 の時がフィルタリングの性能、正解データの損失ともに最もバランスのとれる結果を得られるといえる。

4. 複数の人名読み候補がある場合

最後に、DP マッチングによるフィルタリングを通過した人名読み候補の中から、最も確からしい人名読みを決める。本研究では、人名読みの決定方法として、人名漢字と人名読み候補との共起度を、検索エンジンの Hit 数を用いることで計る方法をとる。共起度の尺度には、jaccard 係数を用いた。jaccard 係数は、以下の式で表される。

$$jaccard\ 係数 = \frac{人名漢字と人名読み候補の\ and\ 検索の\ hit\ 数}{人名漢字と人名読み候補の\ or\ 検索の\ hit\ 数} \quad (2)$$

すべての読み候補に対して jaccard 係数の計算を行い最も共起度が高かった人名読み候補を人名読みと決定する。最終的に、テストデータのうち 93.5% に対して正しい人名読みを付与することに成功した。

5. 評価実験と考察

5.1 評価方法

提案手法を用いて、CiNii [5] の著者データをもとに、大規模なデータでの評価実験を行う。CiNii は国立情報学研究所の提供する論文データベースである。CiNii が保持する著者データのうち 30704 件を抽出して元データとする。CiNii では著者データが人名漢字と人名のローマ字表記のペアで格納されている。本研究では、ローマ字を平仮名に変換して正解データとした。

5.2 実験結果

実験結果を表 4 に示す。

表 4: 実験結果

実験結果	値
元データの数	30704
スニペット中に正解が含まれる	24235
正しい人名読みを付与できた	19624

本研究の手法では、そもそも検索結果のスニペットに正しい人名読みが存在しない場合は人名読みを抽出することができない。スニペット中に正しい人名読みが含まれるものは、24235 件という結果であった。この中から、19624 件の人名について正しい人名読みを付与することができた。また、先行研究 [1] 参考にして、一般性 (generality) と適合率 (precision) を評価尺度として用いた。結果、一般性が 0.6391、適合率が 0.6895 となった。なお、一般性と適合率は以下の式で表される値である。

$$一般性 = \frac{正しい読みが得られたキーワード数}{すべてのキーワード数} \quad (3)$$

$$適合率 = \frac{正しい読みが得られたキーワード数}{読み候補が抽出されたキーワード数} \quad (4)$$

6. 考察と今後の課題

評価実験では全体の 63.9%に対して読みの付与に成功した。読みの付与に失敗した理由としては、1. スニペット中に正解読みがない、2. パターンによって正解読みを取得できない、3.DP マッチングで正解が除外される、の 3 点が挙げられる。また、全体のうち 1 割はローマ字を平仮名に変換する時点で正しく変換されなかったことが原因である。精度のさらなる向上のためにはこれらの箇所でパラメータの調整を行うことが考えられる。また、今回の研究では DP マッチングフィルタリングを通過した候補から一つに答えを絞り込んだが、本来 DP マッチングを通過した読み候補は何らかの人名読み候補である可能性が高い。そのため、複数の読み候補に対して信頼度を付与した形でアウトプットを行うことも考えられる。これらは、今後の課題である。

参考文献

[1] 三宅純平, 竹内翔大, 川波弘道, 猿渡洋, 鹿野清宏. 括弧表現に基づく web テキストマイニングを用いた流行語への自動読み付与の提案. 電子情報通信学会技術研究報告. SP, 音声, Vol. 108, No. 422, pp. 1–6, 2009.

[2] 内田誠一. Dp マッチング概説 : 基本と様々な拡張. 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 106, No. 428, pp. 31–36, 2006.

[5] http://ci.nii.ac.jp/