生起確率の差を用いた人名判定

開出 紗代子 †　　佐 藤　理 史 ‡

† 名古屋大学 工学部 電気電子・情報工学科　　‡ 名古屋大学大学院 工学研究科

1. はじめに

英語のテキストを日本語に翻訳する際，そこに現れる外国人の人名は，その発音に基づいてカタカナ表記される (翻字される) のが普通である．もし，標準的な既訳が存在しているのであれば，それに倣う必要があるため，翻訳者は，まず，外国人名対訳辞典，関連する既訳書籍・文書等を調べ，それらに既訳が見つからなければさらにウェブへと探す対象を広げていくのが普通である．

我々は，外国人名の既訳を探す作業を効率化するため，ウェブから外国人名対訳辞書を自動編纂することに取り組んでおり，昨年度までに，19 万件を収録した対訳辞書の自動編纂を実現した [1]．この辞書が収録している対訳の精度はおおよそ 90%であるが，実用に供するためには，さらなる精度の向上が不可欠である．調査の結果，誤りの多くは「人名以外」の対訳の混入であり，人名か否かの判定精度の向上を図る必要がある．

このような背景より，我々は，昨年度提案した人名判定法を見直し，その基本的な考え方を踏襲しつつ，さらなる改良を行なうことにより，高精度・高再現率の人名判定法の実現を目指した．本論文では，この内容について報告する．

2. 人名の表記条件

本節では，まず，準備として，本研究が対象とする人名 (フルネーム，および，人名コンポーネント) の，文字列としての表記条件を明確にする．すなわち，我々が扱う人名はすべて，ここに示す条件を満たす文字列として表記される．但し，逆は成り立たない．すなわち，表記条件を満たす文字列が，すべて人名であるわけではない．以下，本論文では，表記条件を満たす文字列を，人名候補文字列と呼ぶ．

人名候補文字列は，以下の 5 つのコンポーネント (構成要素) から構成されているとする．

Prefix　**First**　Middle　**Last**　Suffix

これらのコンポーネントは上記の順序に出現し，First と Last は必ず存在しなければならない．それ以外のコンポーネントは省略可能である．コンポーネントの区切りは，英語ではスペース，日本語では「・」とする．

英語，日本語のそれぞれのコンポーネントの表記条件は次のとおりである．

英語のコンポーネントの表記条件

Prefix　「Sir」

First　イニシャルまたは大文字から始まる文字列

Middle　イニシャル，大文字から始まる文字列，または「von」,「de」,「da」,「van」のいずれか 1 つ

Last　大文字から始まる文字列

Suffix　「Jr」または「Sr」

日本語のコンポーネントの表記条件

Prefix　「Ｓｉｒ」または「サー」

First　イニシャルまたはカタカナ文字列

Middle　イニシャルまたはカタカナ文字列

Last　カタカナ文字列

Suffix　「Ｊｒ」,「Ｓｒ」,「ジュニア」,「シニア」のいずれか 1 つ

ここで，英語と日本語それぞれのイニシャル，文字列は以下のとおりとする．

英語のイニシャル　アルファベット大文字 1 文字の後にピリオドの文字列

大文字から始まる文字列 (英語)　アルファベット大文字から始まり，以下アルファベット大文字，アルファベット小文字,「-」(ハイフン),「'」(アポストロフィー) のいずれかが 1 文字以上続く文字列．但し，アルファベットにはアクセント記号が付いていてもよいとする．

日本語のイニシャル　アルファベット全角大文字 1 文字

カタカナ文字列 (日本語)　カタカナ文字 (小さなカタカナを除く) から始まり，以下カタカナ文字,「ー」(長音記号) のいずれかが 0 文字以上続く文字列

3. コンポーネントに対するスコアの付与

3.1 生起確率の差を用いたスコアの付与

人名候補文字列の 5 つのコンポーネントのうち，First と Last の 2 つにスコアを付与する．ここでは，First に対してスコアを付与する方法について説明する．Last についても同様の方法でスコアを付与する．

まず，準備として，次の 3 つの条件を満たす 2 つの集合 D と S を用意する．

(1)　集合の要素は，すべて，人名候補文字列である．

(2)　D は高い割合で人名を含み，S はそれより低い割合で人名を含む．

(3)　D は S に完全に含まれている．

次に，それぞれの要素から First の位置にある語を取り

出し，これを集めたリストを作成する (このリストには，要素の重複がある)．2 つの集合 D と S からこうして作られるリストを，それぞれ D_f と S_f と表す．

D は，S より高い割合で人名を含んでいる．このため，ファーストネームとしてよく使われる語 (たとえば，John) は，S_f よりも D_f により高い割合で含まれていることが期待できる．このような考え方に基づき，語 w のファーストネームらしさの指標を次式で定義する [1]．

$$d(w) = \frac{freq(w, D_f)}{|D_f|} - \frac{freq(w, S_f)}{|S_f|} \tag{1}$$

ここで，$freq(w, D_f)$，$freq(w, S_f)$ は，それぞれリスト D_f，S_f における w の出現回数を，$|\cdot|$ はリストの要素数を表す．この式から明らかなように，$d(w)$ は，リスト D_f，S_f における w の出現確率の差を表す．

式 (1) を $|D_f|$ で正規化すると次式が得られる．

$$\begin{aligned} s(w) &= freq(w, D_f) - \frac{freq(w, S_f)}{|S_f|} \cdot |D_f| \\ &= freq(w, D_f) - e(w) \end{aligned} \tag{2}$$

第 1 項は，リスト D_f における w の出現回数，第 2 項 $e(w)$ は，リスト S_f における w の出現確率から計算した，サイズ $|D_f|$ のリストにおける w の出現回数の期待値となっている．本研究では，w のファーストネームらしさのスコアとして，この $s(w)$ を採用する．

3.2 人名コンポーネントの 3 値判定

リスト S_f の各要素にスコアを付与した後，閾値を用いてファーストネームかどうかを判定すれば，ファーストネームの辞書が作成できる．

ここでは，次の式を用いて，ファーストネームかどうかを 3 値で判定する．

$$j(w) = \begin{cases} +1 & \text{if } s(w) > 0 \\ 0 & \text{if } -1 < s(w) \leq 0 \\ -1 & \text{if } s(w) \leq -1 \end{cases} \tag{3}$$

すなわち，

- 期待値 $e(w)$ よりも実際の出現数 $freq(w, D_f)$ が大きい場合，w をファーストネーム (+1) と判定する．
- 期待値 $e(w)$ よりも実際の出現数 $freq(w, D_f)$ が小さい場合，
  - もし，その差が 1 以上であれば，ファーストネームではない (−1) と判定する．
  - その差が 1 より小さい場合は，不明 (0) と判定する．これは，2 つのリストのサイズに $|D_f| < |S_f|$ という関係があり，期待値の小数点以下を丸めると，差が 0 となるからである．

この判定法を用いて，どのくらいの大きさのコンポーネント辞書が作成できるかを調べた．ここでは，集合 D および S として，以下のものを用いた．

- 英語 (異なり数については表 1 参照)
  - D 昨年度収集した人名対訳のうち，日･英ともに人名の表記条件を満たす対訳の英語側 (236,459 件)．
  - S 英語版 Wikipedia の abstract から，大文字で始まる単語が連続する単語列を抜きだし，そのうち人名の表記条件を満たすものを抽出する．これに D を加えたリスト (1,354,950 件)．ただし，アクセント記号は落とす (ex. $\ddot{a} \rightarrow a$)．
- 日本語
  - D 昨年度収集した人名対訳のうち，日･英ともに人名の表記条件を満たす対訳の日本語側 (236,459 件)．
  - S 新聞 (毎日新聞 15 年分 (1991-2005)) と河原ら [2] が収集したウェブコーパスから人名の表記条件を満たすものを抽出する☆．これに D を加えたリスト (ファーストネームの異なり数: 415,880，ラストネームの異なり数: 637,698)．

表 1 ファーストネーム，ラストネームの判定結果 (英語)

	S		D		Census
	異なり数	%	異なり数	%	
last	329,589	1.000	72,151	1.000	
+1	**68,131**	0.207	68,131	0.944	88,799
0	252,207	0.765	1,742	0.024	
-1	**9,251**	0.028	2,278	0.032	
first	205,924	1.000	19,214	1.000	
+1	**15,665**	0.076	15,665	0.815	5,163
0	179,877	0.874	1,342	0.070	
-1	**10,382**	0.050	2,207	0.115	

英語の判定結果を表 1 に示す．比較のため，1990 年のアメリカの国勢調査 (Census) の結果の一部として公開されているファーストネーム，ラストネームのリストの件数も併記した．この表より，ファーストネームに関しては，+1 のエントリー数が Census の収録数を大きく上回っていることがわかる．一方，ラストネームに関しては，Census の収録数を下回っている．

作成したコンポーネント辞書が既存のコンポーネント辞書 (例えば，Census リスト) と大きく異なる点は，「ファーストネームではない」というエントリーを含んでいる点である．すべてのファーストネームを網羅することは事実上不可能であるので，このような否定情報を収録することは，人名判定の精度向上に役立つと考えられる．

3.3 既存の辞書を用いたスコアの補正

上記のファーストネームの判定法には弱点がある．

- w が D_f に含まれない場合は，かならず $j(w) \leq 0$ となる．つまり，D_f に含まれない w は，絶対にファーストネームとは判定されない．
- 一般によく使われる語で，ときどき人名のファーストネームとして使われるもの (たとえば，Rose) は，スコアがマイナスの大きな値となり，ファーストネームでないと判定される．

これらの弱点を補うために，既存のファーストネーム辞書を用いてスコアを補正する．具体的には，既存辞書に

☆ 昨年度抽出したデータをそのまま流用し，First と Last を取り出した後のリストを使用した．

w が存在するとき，そのスコアを以下のように補正する．

$$s'(w) = \begin{cases} s(w)+1 & \text{if } s(w) > -1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

ここで，$s(w)$ は補正前のスコア，$s'(w)$ は補正後のスコアを表す．

この補正により，w が既存辞書に含まれていれば，判定「不明 (0)」が「ファーストネーム (+1)」に，判定「ファーストネームではない (−1)」が「わからない (0)」に変更される．すなわち，D_f に出現しないような珍しいファーストネームや，ときどきファーストネームとしても使用される一般語が既存辞書に登録されていれば，誤判定の数が減少すると考えられる．

4. 人 名 判 定

4.1 人名候補文字列に対するスコアの付与

3 節で定義した First，Last に対するコンポーネントのスコア $s(w_f)$，$s(w_l)$ を用いて，人名候補文字列 t のフルネーム (人名) らしさのスコアを計算する．スコアの計算には，次式を用いる

$$score(t) = score(w_f, w_l) = \frac{s(w_f) + s(w_l)}{2} \quad (5)$$

このスコアは，2 つのスコアの平均値として定義されるため，$s(w_f)$，$s(w_l)$ のいずれかのスコアが非常に高いと，$score(t)$ はその影響を強く受けてしまう．(いずれかのスコアがとても低い場合も同様である．) これを回避するために，フルネームのスコアを計算する前に，コンポーネントスコアに上限・下限を設ける．具体的には，次式でコンポーネントのスコアを補正する (これを区間補正と呼ぶ)．

$$s''(w) = \begin{cases} +m & \text{if } s(w) \geq +m \\ -m & \text{if } s(w) \leq -m \\ s(w) & \text{otherwise} \end{cases} \quad (6)$$

ここで，$s(w)$ はコンポーネント w に対する区間補正前のスコアを，$s''(w)$ は区間補正後のスコアを，$\pm m$ は区間の上限・下限を表す．このような補正を行うことにより，コンポーネントのスコアは，$+m$ から $-m$ の間の値をとることになる．

4.2 人名の 5 値判定

以上の準備を経て，人名候補文字列 t が人名であるかどうかを次のように判定する．

$$judge(t) = judge(w_f, w_l) = \begin{cases} +2 & \text{if } score(t) > 0 \wedge j(w_f) = j(w_l) = 1 \\ +1 & \text{if } score(t) > 0 \wedge j(w_f) \neq j(w_l) \\ 0 & \text{if } -1 < score(t) \leq 0 \\ -1 & \text{if } score(t) \leq -1 \wedge j(w_f) \neq j(w_l) \\ -2 & \text{if } score(t) \leq -1 \wedge j(w_f) = j(w_l) = -1 \end{cases} \quad (7)$$

ここで，判定値+2 は「高い確信度で人名である」，+1 は「人名である」，0 は「不明」，−1 は「人名ではない」，−2 は「高い確信度で人名ではない」を意味する．なお，$j(w_f), j(w_l)$ は，コンポーネントの 3 値判定結果 (式 (3)) の値を表す．

4.3 人名判定実験

実験では，以下のデータを用いた．なお，D と S は 3.2 節で用いた D，S をそのまま用いた．

人名判定テストデータ　人名判定実験を行うために人名判定テストデータを作成した．このデータは，Wentland ら [3] が作成した HeiNER 辞書から作成した．HeiNER 辞書は，様々な言語の Wikipedia から固有名詞の対訳を取り出した辞書である．HeiNER 辞書に含まれる日・英対訳ペアのうち，日・英ともに人名の表記条件を満たすペアを無作為に 1,551 件抽出し，人手で人名か否かのラベルを付与した．テストデータは，人名対訳 1,204 件 (77.6%)，人名の対訳以外 347 件 (22.4%) から構成されている．

コンポーネントのスコア補正用の既存辞書

- **英語**　3.2 節で比較に用いた Census のファーストネーム，ラストネームのリスト (ファーストネーム: 5,163 件，ラストネーム: 88,799 件)
- **日本語**　Census のファーストネーム，ラストネームのリストの一部にカタカナ訳を付与したもの (ファーストネーム: 2,180 件，ラストネーム: 6,236 件)

人名判定方法として，以下の 6 種類を試みた．

(1) 補正なし
(2) 既存辞書でコンポーネントのスコアを補正
(3) 既存辞書でスコアを補正後，さらに $\pm m = 10$ で区間補正
(4) 既存辞書でスコアを補正後，さらに $\pm m = 5$ で区間補正
(5) 既存辞書でスコアを補正後，さらに $\pm m = 3$ で区間補正
(6) 既存辞書でスコアを補正後，フルネームのスコアを算出せずにコンポーネントの判定結果の和 $(judge(w_f) + judge(w_l))$ で人名判定を行う

この他に，比較のため，既存辞書 (スコア補正用に用いるものと同一) のみを使用して人名判定した場合の結果を示した．この場合は，次の判定法を用いた．

$judge = +2$　w_f, w_l ともに既存辞書に含まれる
$judge = +1$　w_f, w_l のいずれかが既存辞書に含まれる

人名判定テストデータ (対訳データ) の英語側，日本語側のそれぞれに対する実験結果を表 2，表 3 に示す．なお，英語側の文字列に含まれるアクセント記号は，すべて削除して実験を行った．これらの表において，提案手法 (2)〜(6) で $judge = +2$ の判定結果は，同一である．これは，区間補正を適用しても，コンポーネントのスコアの正・負は変化しないことによる．

今回，我々が目指した人名判定精度は 95%以上である．判定基準 $judge \geq 0$ は，要求精度を満たさない．一方，

表 2 人名判定結果 (英語)

	$judge = +2$			$judge \geq +1$			$judge \geq 0$		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
(1) スコアの補正なし	0.993	0.781	0.874	0.963	0.909	0.935	0.917	0.965	0.941
(2) 既存辞書補正	0.992	0.806	0.889	0.953	0.925	0.939	0.895	0.979	0.935
(3) 既存辞書補正, 区間補正 $\pm m = 10$	0.992	0.806	0.889	0.954	0.924	0.939	0.893	0.981	0.935
(4) 既存辞書補正, 区間補正 $\pm m = 5$	0.992	0.806	0.889	0.957	0.924	0.940	0.890	0.983	0.934
(5) 既存辞書補正, 区間補正 $\pm m = 3$	0.992	0.806	0.889	0.959	0.922	0.940	0.884	0.985	0.932
(6) コンポーネントの判定結果の和で判定	0.992	0.806	0.889	0.960	0.920	0.940	0.893	0.983	0.936
既存辞書のみで判定	0.991	0.468	0.636	0.882	0.826	0.853	-	-	-

表 3 人名判定結果 (日本語)

	$judge = +2$			$judge \geq +1$			$judge \geq 0$		
	精度	再現率	F 値	精度	再現率	F 値	精度	再現率	F 値
(1) スコアの補正なし	0.986	0.772	0.866	0.949	0.921	0.935	0.911	0.974	0.941
(2) 既存辞書補正	0.986	0.777	0.869	0.946	0.928	0.937	0.895	0.980	0.936
(3) 既存辞書補正, 区間補正 $\pm m = 10$	0.986	0.777	0.869	0.949	0.919	0.934	0.893	0.981	0.935
(4) 既存辞書補正, 区間補正 $\pm m = 5$	0.986	0.777	0.869	0.952	0.918	0.934	0.891	0.983	0.935
(5) 既存辞書補正, 区間補正 $\pm m = 3$	0.986	0.777	0.869	0.955	0.914	0.934	0.888	0.985	0.934
(6) コンポーネントの判定結果の和で判定	0.986	0.777	0.869	0.957	0.909	0.933	0.889	0.983	0.933
既存辞書のみで判定	0.985	0.280	0.436	0.921	0.757	0.831	-	-	-

判定基準 $judge = +2$ では，再現率が低すぎる．以上により，判定基準 $judge \geq +1$ を採用するという方針が導かれる．

この表から，判定基準 $judge \geq +1$ において，次のことがわかる．

(a) 提案手法 (1)〜(6) は，既存辞書のみを利用した判定方法と比べ，精度，再現率，F 値のすべてにおいて優れている．大規模な既存辞書が利用可能な英語の場合においても，F 値で比較すると，提案手法の方がかなり良い値となっている．

(b) 提案手法の (1) と (2) を比較すると，既存辞書を用いた補正により，再現率が向上するが，その一方で精度が若干低下することがわかる．F 値は向上しているので，この補正は有効とみなせる．日・英を比べると日本語の方が効果が限定的なのは，使用した既存辞書のサイズによると考えられる．

(c) 提案手法の (2) と (3)〜(5) を比較すると，区間補正 (上限・下限の導入) により，精度が向上するが，その一方で再現率は低下し，F 値はほぼ変わらないことがわかる．上限・下限の値を変えても F 値はほとんど変わらない (区間を狭めると精度は向上するが，再現率は低下する)．

(d) 提案手法の (3)〜(5) と (6) の性能 (F 値) は大差はない．(6) の方が精度は高いが，再現率は低い．

これらの結果を受け，我々は提案手法 (4) を採用する．その性能は，英語で精度 0.957，再現率 0.924，日本語で精度 0.952，再現率 0.918 となる．

5. おわりに

本論文では，2 つのリストの生起確率の差を利用した人名判定方法を提案し，その精度・再現率を実験的に明らかにした．提案方法は，英語 (ラテン文字列) に対して F 値で 0.940，日本語 (カタカナ文字列) に対して 0.934 となり，日・英ともにかなり良い精度・再現率で人名か否かの判定が可能であることがわかった．

提案方法の性能は，用いる 2 つの集合 D と S の大きさと質に依存する．より大きな D と S，より高い割合で人名を含む D が入手可能となれば，人名判定性能はさらに向上すると考えられる．これを実現する一つの方法として，ブートストラップを用いた方法が考えられる．実際，本研究では，昨年度収集した人名対訳の英語側，日本語側の文字列を集合 D として使用している．この対訳収集には，昨年度の人名判定方法 (日本語のみ) を利用しており，ある種のブートストラップ法となっている．より直接的なブートストラップ法として，S に対して提案手法を適用し，人名と判定されたものを新たに D に追加した後，再度スコアを計算し直す方法が考えられる．今後，方法の有効性を調査していく予定である．

謝辞 本研究は，栢森情報科学振興財団の助成 (「ウェブを利用した対訳辞書の自動編纂」) を受けて遂行された．

参 考 文 献

1) 榊原洋平, 佐藤理史: 外国人名対訳辞典の大規模化ー15 万件の自動編纂ー, 言語処理学会第 14 回年次大会発表論文集, pp. 833–836 (2008).

2) 河原大輔, 黒橋禎夫: 高性能計算環境を用いた Web からの大規模格フレーム構築, 情報処理学会研究報告, Vol. NL-171-12, pp. 67–73 (2006).

3) Wolodja Wentland, Johannes Knopp, C. S. and Hartung, M.: Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* ((ELRA), E. L. R. A.(ed.)), Marrakech, Morocco (2008).