増深解析法による
日本語の固有名の認識(1)

舘野昌一(富士ゼロックス株式会社、tateno.masakazu@fujixerox.co.jp)
李　黎陽(中国北方交通大学、leeliyang@yahoo.com.cn)

要旨
日本語のテキストから人名を認識する方法を提案する。人名の認識は固有名認識の中でも重要なものの一つである。本方法は日本人の人名用字の統計情報を用いる。その典型的な方法を例とともに示した。評価データにより得たFスコアは90.8%であった。

1 はじめに

固有名は、形態素解析でうまく解析されにくいという、言語処理における障害物としての側面と、うまく解析されればその情報量が大きく、構文解析や意味解析への貢献が期待できるという二つの側面があり、その精度を向上させることは重要である。Message Understanding Conference の情報抽出のサブタスクでは、人名、組織名、地名、日付、時刻、金額、パーセントの7種類を固有名としている。この中で人名の抽出は、個人を特定することや逆に隠蔽することなど、固有名を応用する観点からはきわめて重要である。そこで、本稿では、日本語のテキストに現れる人名に絞り、認識する方法を述べる。

2 人名の種類

人名は日本人の人名、カタカナで書かれる主に欧米人の人名、漢字で書かれる中国人など日本人以外の人名に分けられる。このような人名が現れるパターンは多様であり、本タスクの難しさを示している。ここで、日本人の姓と名は、ともに多くの種類があり、また、この中にはめずらしいものが含まれるので、すべての姓と名をあらかじめ登録しておくことは困難である。形態素解析だけではいい結果が得られない。そのため、エントロピー最大法・決定リスト学習法[1]、サポートベクターマシン[2]などの統計的方法や、言語的な規則による方法により、認識する方法が提案されているが、本稿では、増深解析法[3]による方法で行う。なお、アメリカ、イギリスなどの片仮名で表される人名と、中国などの漢字で表される人名を抽出する規則については紙幅の都合上、説明を割愛する。

3 本方法の概要

形態素解析には、Chasen2.0[4]を用い、人名の認識にはXerox Incremental Parser (XIP、[3])の規則を記述し実行した。また、固有名認識後の不確かな結果を検査し処理するため、Perlのスクリプトを記述し実行した。

4 日本人名の認識

日本人の姓と名は、それぞれ概ね1文字から3文字の漢字であるが、4文字以上の場合もある。したがって、日本人の姓名は2文字から6文字またはそれ以上であり、人名を認識するということは、その左境界と右境界を見つけることである。形態素解析ソフトは、その辞書部に、姓と名の部分集合を含んでいるだけなので、細かく分割されてしまうことがよくある。以下に例を示す。

FNAME	GNAME	NOUN	NOUN
依田	紀	基	名人
FNAME	NOUN	GNAME	
嶋	山	ハル子	

そこで、そのような断片を適切につなぎ、人名の境界を明確にする。最初の例では名である「紀基」が、2番目の例では姓である「嶋山」が誤って分割されている。

5 前処理

このような現象を処理するため、あらかじめ次の準備を行った。つまり、約 24,000 の姓の先頭の漢字と末尾の漢字を収集した。その数はそれぞれ 1,990 と 1,600 であった。同様に、約 59,000 の名についても行うと、それぞれ 5,550、4,030 であった。そこで、それぞれの文字集合に素性として姓の先頭になりうるか(flb)、末尾になりうるか(frb)、名の先頭になりうるか(glb)、末尾になりうるか(grb)、を割り振っておく。

これを、認識時の形態素解析後に各形態素の先頭や末尾の文字に関して確認し、上記4種類の素性の文字であれば、その素性を付与する。一つの形態素が、これらの素性を複数持つことはもちろんある。

前述の例では、XIP の塊化規則により、「紀」は素性 glb を持ち、「基」は素性 grb を持ち、さらに、名人が素性 title を持つので、「紀基」は名であると認識される。しか

し、以下の例では、「戦」は素性 grb を持たないので、名の一部ではないと判断される。

そのための規則の概要は次のとおりである。

FNAME	GNAME	NFCOM	PUNC
金子	金五郎	戦	。

name = | left-boundary | name-body [flb:+], name-body+, given-name | right-boundary |.

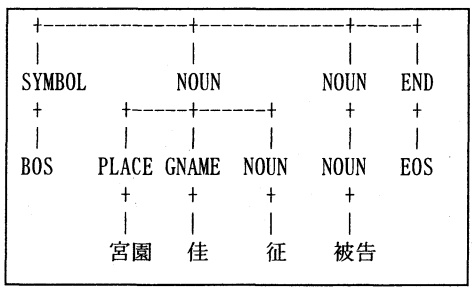
ここで、| left-boundary |と| right-boundary |は、人名としてまとめ上げる際に条件として用いるが、人名の一部としては用いないことを示す。また、name-body+は、1個以上の1文字漢字の語を示す。この規則により、素性 flb を持つ1文字漢字の後に、1個以上の1文字漢字があり、その後に名があり、この列全体が、特定のコンテクストで囲まれていれば、人名と認識される。もう一つ、形態素解析結果が地名となる人名の場合を説明する。「宮園佳征被告」は、次のように形態素解析される。

宮園　+0+4+名詞-固有名詞-地域-一般
佳　+4+6+名詞-固有名詞-人名-名
征　+6+8+名詞-サ変接続+grb+…
被告　+8+12+名詞-一般+L2

ここで、姓である宮園は地名としてタグ付けされている。これを修正するために次のような規則が必要となる。

name = | left-boundary | family-name (place-name), name-body+, name-body[grb:+] | right boundary |.

ここで、family-name (place-name)は、姓または地名のどちらかが可能であることを示す。その結果、以下のような結果が得られる。



なお、一つの規則が別の規則が抽出する箇所の一部分だけを抽出する場合は、それを削除した。また、姓名が合わせて抽出されているのと同一記事内で出現する姓または名と一致する形態素は、それぞれ姓、名として認識する処理を行った。

5. 実験および結果

'95 年版の毎日新聞の1月1日、10日、14日、15日の4日分のうち、1日と15日の記事を用いて規則を記述し訓練し、残りの2日分で検査を行った。人名としての集計を行うと、訓練コーパスでは、適合率 89.5%、再現率 90.6%、F スコア 90.0%、検査コーパスでは、適合率 87.4%、再現率 88.9%、F スコア 88.2%となった。また、'99 年の毎日新聞の記事に基づく IREX の NE タスクの中の 69 件の記事を用いて、評価も行った。その際に、"(みょうじ・なまえ)"のようにひらがなで表記されたものに対応する、人名の読み仮名に関する規則を書き足した。その結果、適合率 90.6%、再現率 91.0%、F スコア 90.8%を得た。なお、固有名を認識するための処理速度は、9.4KB/秒(Linux(RedHat8.0)、Pentium4-166GHz)である。この中には、組織名など他の固有名を認識する処理も含まれている。

6. 比較

我々の結果は、言語的方法によるものとしては、高い精度のものとなった。しかし、統計的方法を用いた磯崎・賀沢[2]によるものに及ばなかった。

方法		F-score (%)
言語規則(XIP)	舘野、李	90.8
サポートベクターマシン	磯崎、賀沢	94.4
決定リスト学習と最大エントロピー	宇津呂、颯々野、内元	86.15

7. 結論と将来の活動

構文規則を記述し、増深解析法により人名の認識を行う方法により、統計的方法とほぼ同等の結果が得られることを示した。このような言語的な方法で行うことの利点は次のとおりである。第一に、言語現象を捉えた規則は、その言語内で一般性がある。第二に、規則は、人間が言語現象として理解できるので、合理性があり編集可能である。第三に、固有名の誤認識は、どの規則が適用されたかを見ることにより明白になるので、修正しやすい。今後は、さらに高い精度を実現することを目指すとともに、本方法を他の固有名にも適用して、再現率と適合率の向上を目指す。

参考文献

[1] 宇津呂 武仁, 颯々野 学, 内元 清貴, "正誤判別規則学習を用いた複数の日本語固有表現抽出システムの出力の混合", 自然言語処理, 第9巻, 第1号, pp.65-100, January 2002
[2] Hideki Isozaki and Hideto Kazawa, Efficient Support Vector Classifiers for Named Entity Recognition, Proceedings of COLING-2002.
[3] Salah Aït-Mokhtar, Jean-Pierre Chanod, and Claude Roux. "Robustness beyond shallowness: incremental deep parsing". In Natural Language Engineering, 8(2):121--144, 2002
[4] http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html.