決定木学習を用いた人名情報抽出のための規則生成

荻田　泰宏†　長崎　英紀†　古宮　嘉那子†　柴原　一友†　但馬　康宏†　小谷　善行†
†東京農工大学工学部情報コミュニケーション工学科

1.はじめに

近年，Web 上でのやり取りは常に増加傾向にあり，それに伴って個人情報の流出問題も増大している．その個人情報漏洩を防ぐために，文書中から人名，地名，住所などの個人情報と思われる単語を捜し出し別の単語や黒丸などの記号に変換する個人情報マスキングの研究がなされている．人名抽出を研究することは個人情報の代表格の一つである人名を文書中から捜し出し，個人情報漏洩を防ぐ手段に繋がる．

人名抽出の関連研究としては，辞書と共起情報を用いて新聞記事から人名を獲得する研究[1]や，形態素解析器の品詞情報によって Web 上のテキストデータから人物情報を取得する研究[2]などがある．[2]では Web 上のテキストデータは多様で統制されておらず，人名が出現する典型的なパターンを調べあげることが難しいという理由から形態素解析器によって人名と同定された形態素の並びだけを人名として抽出している．ただし[2]は人名抽出だけでなく Web 上の人物同士の関係を可視化することを目的としている．

個人情報マスキングの観点から見ると[2]の抽出方法では形態素解析の結果に依存してしまい人名をマスキングしきれないと思われる．特に Web 上ではハンドルネームという個人情報も存在する．しかしハンドルネームも含めて人名を抽出する研究はなされていない．

そこで本研究では Web 上の主要なテキストデータである blog データを対象にハンドルネームを含めた人名を高再現率で抽出することを目的とし，決定木学習による人名情報の抽出を行うシステムを作成した．

2.人名情報抽出の方法

2.1 人名の定義

本研究では，人名を「特定の個人を指すもの」として定義する．姓や名などは勿論，blog データ上では頻出する「ハンドルネーム」も人名に含まれるものとする．

2.2 決定木学習

決定木学習を用いる．決定木は，属性と結果で構成される木であり，ひとつのノードは属性によって結果を分類する．決定木学習とは，この木の性質を用いた機械学習であり，学習データを与えて木を生成し，生成した決定木をルートノードからトップダウンで辿ることで，適切な結果を選択することができるようにする手法である．人名情報抽出システムで用いた決定木作成のアルゴリズムは C4.5 であり，生成する決定木は二分決定木である．

2.3 決定木学習に用いる要因

要因には，表記上の情報や，形態素解析器「茶筌」，係り受け解析器「Cabocha」によって得られる情報を用いた．その詳細を表１に示す．これらの要因は決定木学習の際に与えられる学習データの一部となる．

表１　人名情報抽出システムの要因一覧

対象語の 4 階層に分けた品詞情報	4 種類
ひらがな，カタカナなどの表記情報	1 種類
周囲の品詞情報　6 単語×４階層＝２４	種類
係り先の単語情報と品詞　1＋4＝5	種類
係り先の単語の活用形と活用型	2 種類
特定の助詞がついた単語情報	1 種類
特定の助詞がついた単語の品詞情報	4 種類
人名に頻出する漢字の有無	1 種類

3.人名情報抽出システムの概要

人名情報抽出システムの概略図を図１に示す．

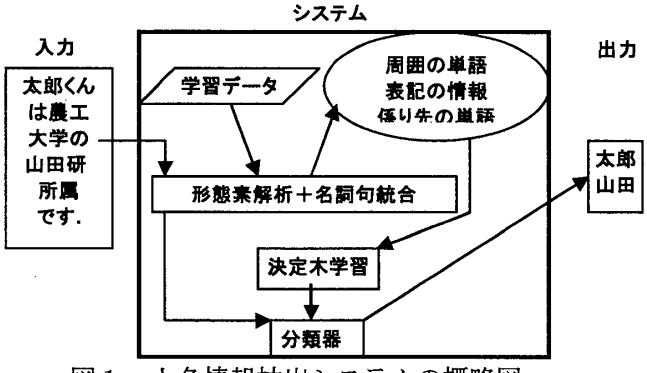

図１　人名情報抽出システムの概略図

Rule Generation for Extraction of Personal Names using Decision Tree Learning
Yasuhiro Ogita † Hideki Nagasaki † Kanako Komiya † Kazutomo Shibahara † Yasuhiro Tajima † Yoshiyuki Kotani †
† Department of Computer, Information and Communication Science, Tokyo University of Agriculture and Technology

入力は blog データの文章である．システムは文章を一行ごとに読み込み，形態素解析にかけて形態素に分割する．そこで分かれて解析されてしまった人名を結合するため，名詞句の統合をする．その方法は，同文節内の形態素を助詞や用言が出現するまで統合するというシンプルな方法をとっている．誤って統合されてしまったものは無視する．そうしてできた名詞句をシステムが学習データから決定木学習によって生成された規則に当てはめ，人名かどうか判定し抽出を行う．

4.決定木性能評価実験

4.1 実験方法

blog データを用いて，決定木学習による人名情報抽出の性能を測定した．決定木作成の終了条件は情報利得比がなくなった時点で終了とした．対象データは名詞句 10000 件で，正例として人名が 980 件，負例として非人名が 9020 件である．評価には4分割のクロスバリデーション法を採用している．

またベースラインとして，[2]で用いられている方法の，品詞が「名詞—固有名詞—人名—姓」の直後に「名詞—固有名詞—人名—名」が連続しているものだけを正解としたときの精度と，形態素解析器が「名詞—固有名詞—人名」と判定したものだけを正解としたときの精度を測定する．

4.2 実験結果

人名，非人名それぞれに対する適合率と再現率，F値およびその平均を表2に示す．表2から人名に関しての再現率は 50.51%，適合率は 49.01%であることがわかる．

表2　決定木の性能測定結果

(値)%		Data1	Data2	Data3	Data4	平均
非人名	適合率	93.83	92.68	95.17	91.06	93.22
	再現率	95.46	93.93	96.22	92.60	94.59
	F値	94.63	93.30	95.70	91.82	93.90
人名	適合率	49.34	48.37	48.06	50.18	49.01
	再現率	50.45	49.58	50.61	51.28	50.51
	F値	49.88	48.97	49.30	50.72	49.75
全体	適合率	89.98	88.29	90.43	86.48	88.82
	再現率	91.58	89.58	91.79	88.00	90.27
	F値	90.78	88.93	91.10	87.23	89.54

「名詞—固有名詞—人名—姓」と「名詞—固有名詞—人名—名」連続するものは 7 件．そのうち 5 件は正解．形態素解析結果が人名のものは 899 件．そのうち 410 件は正解．よって表3のように示される．

表3　ベースラインの精度

(値)%		人名連続	形態素結果
人名	適合率	71.43	45.61
	再現率	0.51	41.84
	F値	1.01	43.64

5.考察

表2では，全体の平均が F 値 89.54%と高い精度に見えるが，本研究の最も重要な評価は人名の再現率にある．人名の再現率は平均 50.51%と半分ほどの正解率である．表3の形態素解析器による再現率 41.84%より 8.67%上回る結果となったが，半分近くの人名を取りこぼしてしまうため，あまりよい精度とはいえない．その理由には，blog データには顔文字や特殊な記号，明らかにわざと間違えている単語など，抽出精度を高めるために妨害となる要素がいくつもあることから，表1で用いた要因だけでは多様なパターンを網羅できなかったことや，実験データ 10000 件では事例が少なすぎたことが考えられる．また，負例と正例のバランスが悪いことも精度を下げる一因になっていると思われる．名詞句は人名でないものの方が圧倒的に多いので，形態素解析結果から人名を効率よく統合できる方法が必要である．

6.おわりに

本研究では，blog データから決定木学習を用いてハンドルネームを含めた人名を抽出する規則を生成した．適合率 49.01%，再現率 50.51%，F 値 49.75%という成果を得た．blog データのパターンルールを見つけ出すことは困難であるがデータを増やしていけば可能性はあると考えられる．

参考文献

[1] 久光徹，丹羽芳樹："辞書と共起情報を用いた新聞記事からの人名獲得" 情報処理学会研究報告,NL-118-1, pp.1-6, 1997.

[2] 原田昌紀，佐藤進也，風間一洋："Web上のキーパーソンの発見と関係の可視化(テキストマイニングの応用(1))" 情報処理学会研究報告,DBS-130-3, pp.17-24, 2003.

[3] 吉田祐樹，古宮嘉那子，但馬康宏，小谷善行，"決定木学習を用いた片仮名複合語の略語生成システム",情報処理学会第68回全国大会,2006