1N-8

決定木学習を用いたカタカナ複合語の略語生成システム

吉田　佑樹　　古宮　嘉那子　　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

略語の関連研究としては，任意の日本語略語に対して新聞記事コーパスや広辞苑の語句といくつかの復元規則を用いて復元する研究[1]と，略語とその元の語の付近に現れる単語が類似しているという仮定を元に文章の類似度から略語を元の語に復元する研究[2]などが行われているが，略語を作成する研究は行われていない．そこで本研究では，カタカナ複合語の表面的な特徴だけを用いて単語の省略規則を導き出すシステムを作成した．

２．略語

2.1 略語とは

人は単語を与えられると，略語を作る事が可能で，傾向も似通っている．したがって，略語生成には，ある程度の規則性が存在すると考えられる．略語とは，「もとの語形の一部分を省略して簡略にした語」(大辞林 第二版より)の事である．その特徴として

・正式な省略の取り決め・規則が推察し難い
・音韻特徴などにより決定される事がある
・使う人，時期によってその形が異なる

などがあげられる．そして本システムでは，略語のうち，カタカナの略語だけを扱う．

2.2 略語生成の方法

今回は，単語に適応する省略規則を 15 に定めた．規則の内容は表１に示す．この表１の規則のうちから，一つを選び，その規則にしたがって単語に処理を行い，略語を生成する．適応する規則の選択は計算機が決定木学習を行った結果を用いて行う．決定木学習アルゴリズムは C4.5 を用いた．

表１　単語に適応する規則と適用例

規則	適応例
頭1モーラ取得	セントラル→セ
頭2モーラ取得	コンピューター→コン
頭2モーラ取得後拗音削除	オフィス→オフ
頭3モーラ取得	コスメチック→コスメ
頭3モーラ取得後拗音削除	コーディネート→コーデ
頭3モーラ取得後促音削除	ネット→ネト
頭3モーラ取得語長音削除	パーソナル→パソ
頭4モーラ取得	ハイオクタン→ハイオク
頭5モーラ取得	アポイントメント→アポイント
後1モーラ取得	ウェブ→ブ
後2モーラ取得	ビオロンチェロ→チェロ
後3モーラ取得	アルバイト→バイト
後4モーラ取得	ハンバーガー→バーガー
単語消去	アイス コーヒー→アイス
何もしない	アイス コーヒー→アイス

2.3 適応規則選択の判定要因

判定要因は全て単語の表面的な情報に限定し，辞書を用いなければ理解できない，意味や品詞情報などは用いない事とした．この判定要因は，決定木作成の際に与えられる学習データの一部となる．略語規則の判定要因の候補を，表２に示す．

表２　略語生成システムの判定要因一覧

対象語のモーラ数
対象語の各音節特徴
共起語の有無
共起語のモーラ数
共起語の各音節特徴
共起語に適応された処理
共起語省略後のモーラ数

３．システムの概要

システム概略図を以下の図１に示す．

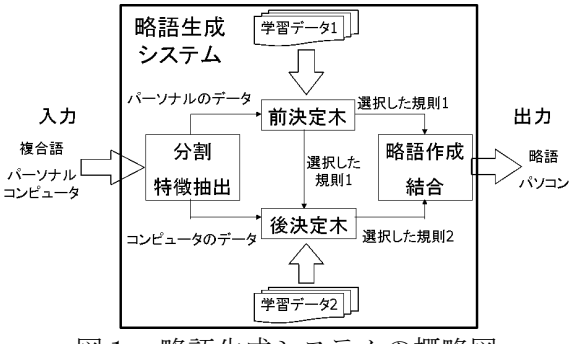

図１　略語生成システムの概略図

Abbreviation Word Generator for Compound Katakana Words by Decision Tree Learning

†Yuki Yoshida Department of Computer, Information Sciences, Tokyo University of Agriculture and Technology

入力は２単語までのカタカナ複合語とし，出力は略語１語とする．システムに語を入力すると，入力語が複合語か単一語かを判断し，複合語なら，前の単語と後ろの単語に分割する．次にそれぞれ判定要因を抽出し，前決定木と後決定木に渡す．前決定木は前の単語に適応する規則を選択し，結果を後決定木に渡す．後決定木は後の単語の判定要因と，前決定木の出力を用いて，後の単語に適応する規則を選択する．前決定木と後決定木の出力を元に，略語を作成し，出力する流れとなっている．入力が単一語の場合は前決定木が選んだ規則を元に略語を生成し，出力する．

４．決定木性能評価実験

[4]から手作業で集めた 1390 語のデータを用いて，決定木学習による規則選択の性能を測定した．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した．評価は前決定木と後決定木で別々に行い，前決定木の対象データは 802 件，後決定木の対象データは 588 件となっている．評価には，双方とも 4 分割のクロスバリデーション法を採用している．末端の葉に，適応される省略規則が複数ある場合，全て正解として出力した場合と，複数の規則の中から最も件数が多い規則を正解とした場合で，適合率と再現率をそれぞれ測定した．なお，答えを一つに限定する時において，件数が等しい場合に，表１上で，より下の規則を選択することとした．

５．結果

前決定木の測定結果を見ると，正解率は，答えを一つにした場合，52.37%で，複数出した場合は 69.70%となっている．詳しい結果は表 3 に示す．

表 3　前決定木の測定結果

		data1	data2	data3	data4	平均
答え複数	適合率	38.23%	39.46%	36.11%	40.96%	38.69%
	再現率	68.66%	72.64%	65.00%	72.50%	69.70%
答え一つ	適合率	48.26%	58.20%	46.00%	57.00%	52.37%

後決定木の測定結果を見てみると，正解率は，答えを一つに絞った場合，63.10%で，複数出した場合は 81.63%となっている．詳しい結果は表 4 に示す．

表 4　後決定木の測定結果

		data1	data2	data3	data4	平均
答え複数	適合率	40.68%	39.93%	42.14%	45.15%	41.98%
	再現率	81.63%	76.87%	85.71%	82.31%	81.63%
答え一つ	適合率	61.22%	55.78%	68.71%	66.67%	63.10%

６．考察

前決定木は，答えを一つに絞ると，約半分は不正解となる．これは，適応される規則が一つとは限らず，後ろの語や，作られた時期によって，複数存在するためと考えられる．そこで，出力する答えを複数にした所，再現率は 70%程度に向上した．正解が出せない 30%の理由として，集めたデータに偏りがある為，テストデータと学習データに分けると，学習されない規則が存在する事に原因があると考えられる．よって，そのような規則の判定はどうしても難しくなってしまう．実際、クローズドデータによるテストでは、再現率 100.00%であるので、サンプルの確保さえ上手くいけば、分類は可能であると考えられる。

後決定木は，答えを一つに絞ると 63.10%，複数出すと 81.63%と，前決定木に比べて高い確率で正解を求める事ができているが，これは，前の単語の省略された形が決まっていれば，後ろの単語に適応する規則も大方決まってくるため，それらの情報が使える後ろの単語は性能が高めに出るためだと考えられる．予備実験として，前決定木の出力を用いて後決定木を作成せず，実際の略語に適応された正解データを用いて後決定木を作成してみたところ，オープンデータによるテストでも，再現率は 86.4%，適合率は 66.9%と，主に適合率の面で飛躍的に性能が向上することが確認された．

７．おわりに

今回定義した 15 の省略規則において，二つの決定木を作成して，その性能を計測した．その結果，前決定木のほうでは適合率 38.69%，再現率 69.70%，後決定木のほうでは適合率 41.98%，再現率 81.63%という成果を得る事ができた．後決定木は前決定木が正解を出力できれば，略語作成の性能は十分であるといえる．しかし，前決定木のほうは，さらに効果的な判定要因を探して改善していく必要がある．

参考文献

[1]石井直樹,平石智宣,延澤志保,斎藤博昭,中西正和：”日本語略語の自動復元”,情報処理学会,2000-NL-137 pp61-68,2000

[2]Terada Akira and Tokunaga Takenobu, : “Automatic disabbreviation by using context information”, Information Processing Society of Japan, 2001-NL-144-6 pp39-45. , 2001

[3]古宮嘉那子、小林明子、乾伸雄、小谷善行，決定木学習による敬語の選択ルールの生成，情報処理学会　第 67 回全国大会,2004

[4]goo 国語辞典: http://dictionary.goo.ne.jp/index.html?kind=jn&mode=0&kwassist=0