方言に起因する形態素解析の区切り誤りを自動検出する手法の試作

久留間嵩之* 乙武北斗** 吉村賢治**
(福岡大学 *大学院工学研究科 **工学部)

1 はじめに

日本語を計算機で処理する際には，文を語の列として分割するために形態素解析という処理を行う．既存の形態素解析器は精度の高い解析が可能であるが，方言などの表現が含まれる発言を解析した場合，品詞推定の誤りや語の区切り位置の誤りなどの解析誤りを起こす事がある．

形態素解析誤りは日本語文書データを自動で解析・処理する上において大きな影響を与える．事前に検出を行うことができれば，その後の処理への影響を最小限に抑えることができる．解析誤りを検出する先行研究[1]は存在するが，再現率の面などに課題が存在する．本稿では分類器を用いた区切り位置誤りの検出手法を提案する．また区切り位置誤りを発生させうる方言を含む地方議会会議録を使用し，提案手法の評価実験を行った．

2 提案手法

本実験では先行研究[2]によって，方言を含む発言が確認された兵庫県の 8 人の発言者の 1979～2011 年における発言の一部(428 発言)を使用する．これらを元に方言に起因する解析誤りが発生する単文を抽出し,**図 1** のようなインスタンス群を生成する．これらインスタンス群の集合を学習データセットとし，最大エントロピーモデルによる二値分類を行うことで，区切り誤りの推定を行う．

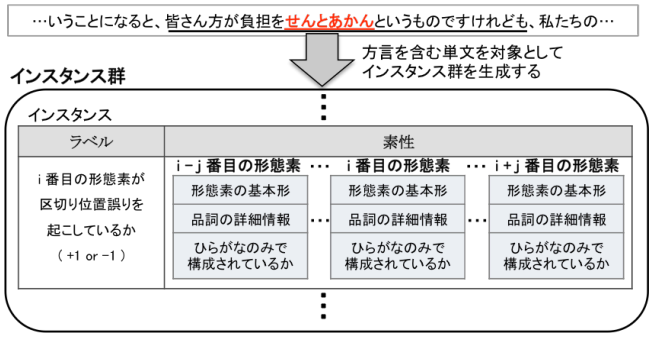

図 1 ．データセットの構成

3 実験手順

本実験は，まず地方議会会議録の一部に対して人手によって方言を含む発言の抽出を行い，解析器で形態素解析を行った．解析器には IPA 辞書を用いた MeCab[3]を使用した．その後，形態素解析結果に対し人手によって解析誤り部のアノテーションを行い，解析誤りデータを作成した．そして解析誤りデータを元にインスタンス群を生成し，学習用データセットを作成した．

分類器の実装としては，Classias[4]を用いた．また評価方法は Leave- one-out 法を採用している．解析誤りである形態素を正例，そうでない形態素を負例とした場合，テスト用データセットの内訳は，正例 253 例，負例が 65,924 例となっている．

4 実験結果

4.1 最適な前後の形態素参照数 j の探索

対象形態素前後の形態素参照数 j の値の変化による評価値の変化は**表 1** のようになる．これらの値から j=0，つまり前後の形態素を素性として考慮しない場合は他の場合と比べ著しく適合率，再現率が下がることが分かった．

また j = 1,2 の時に F 値が最も高くなり，本実験においては最適な値であるということが考えられる．

表 1．j の値が変化した時の評価値の変化

	適合率	再現率	F 値
j=0	1.38%	3.45%	1.97%
j=1	7.59%	45.2%	13.0%
j=2	7.52%	47.8%	13.0%
j=3	6.83%	45.1%	11.8%
j=4	5.82%	43.9%	10.2%

4.2 ひらがなフラグの有効性

方言を含む発言は方言を含まない発言と比較してひらがなが占める割合が 10%ポイント以上高いという研究報告があり[2]，ひらがなと方言は密接な関係にあると考えることができる．本実験では方言に起因する区切り位置誤りの検出を目的としているため，方言部分の検出を目的とした素性として，形態素がひらがなのみで構成されているかのフラグを取り入れ，再現率の向上を図っている．この素性を導入することによる評価値の変化は**表 2** のようになる．F 値のみを参照するとひらがなフラグ無しの方が高くなっている．しかし再現率のみを考えると 6.7%ポイント(253 例中 17 例)上がっており，漏れなく方言に起因する区切り誤りを検出したい場合においては，ひらがなフラグは有効である素性だと考えられる．

表 2．ひらがなフラグの有無による評価値の変化

	適合率	再現率	F 値
フラグ有り	7.52%	47.8%	13.0%
フラグ無し	7.98%	41.1%	13.4%

5 まとめ

本稿では方言に起因する区切り誤りの自動検出を最大エントロピー法による二値分類を用いて試作した．しかし検出率は低く，再現率も決して高いとは言えない結果となった．このような結果になった理由の 1 つとして，前後の誤り情報に関しての考慮が十分でないという点が挙げられる．この点に関して，今後 CRF を用いた分類器での実装を行い，評価値の向上を目指したいと考えている．

参考文献

[1] 黒澤 義明,坂本 裕二,市村 匠,相沢 輝昭:類似度を用いた形態素解析誤り検出尺度，第 21 回ファジィシステムシンポジウム講演論文集 pp.143-146

[2] 地方議会会議録の方言を含む発言における形 態素解析誤りの分析:乙武 北斗,折舘 直樹,吉村 賢治,第 30 回ファジィシステムシンポジウム

[3] MeCab:YetAnother Part-of-Speech And Morphological Analyzer, http://taku910.github.io/mecab/

[4] NaoakiOkazaki,Classias: a collection of machine-learning algorithms for classification, http://www.chokkan.org/software/classias/