3Q-01

文書分別のための未知語の索引化手法の評価

大西 周† 山田 剛一† 絹川 博之†

東京電機大学大学院 未来科学研究科†

1．はじめに

文書分別のシステムを作成する際，索引作成のため形態素解析エンジンを活用することが多い[1]．しかし，形態素解析エンジンを用いると，ツール内辞書に存在しない単語に対し，正しい解析ができない．その結果，未知語が生じ，索引付けに支障が出てしまう場合がある．現在，未知語が出現した際は，新たな単語を辞書に登録し，次回の解析の際に未知語を発生させないようにしている[2]．しかし，新語は次々に生まれるため，辞書への登録を逐次しなければならない．

本研究では，汎用的な文書分別システムの制作を最終目的とする．未知語を含む文書を形態素解析し，結果の形態素の中の未知語を索引化する手法を提案するとともに，提案手法を評価する．

2．形態素解析ツールにおける未知語

形態素解析ツールは複数存在し，それぞれ処理の手法や解析の結果が異なる．本研究ではKyTea[3] (Ver 0.4.2, 高性能SVM使用)を用いる．

2.1 未知語とは

未知語とは，ある文書に対し形態素解析ツールを用い形態素解析した際の辞書未登録である語を指す．未知語は主に新語や略語の発生により生ずるため，永久に増え続けるものと考えられる．

2.2 未知語に対する形態素解析の処理パターン

形態素解析の際，辞書未登録語があった場合の解析結果には2つのパターンがある．

(1) 未知の語が誤って分割され，その断片が別の語として認識される場合
(2) 形態素解析処理された結果の形態素をツールが未知語とする場合

(1) の未知語は，ツールは未知語として処理せず，語を分割し別の語として認識する．そのため，後から機械的にその未知語を検出することは困難である．(2) には，辞書未登録語が語として正しく認識され，未知語として処理される場合と，誤って他の形態素と連結されたものや，語の断片をツールが未知語処理する場合がある．

2.3 KyTea を用いる利点

索引は文書の特徴語であり，索引となる語は，文章内で現れる語の形のまま正しく認識されている必要がある．語として正しく認識されていれば，未知語でも索引語として用いることができる．よって，形態素解析による語の区切りの誤りが少ないほど，文書分別に適するツールであると言える．調査の結果，KyTea が他の形態素解析ツールと比べ，語の区切り誤りが少なく，文書分別の際の索引付与に適しているという結論に至った[4].

3．文書分別のための索引生成

文書分別をするにあたり，文書に索引を付与することにより文書の特徴を表すことが多い．そのため，すべての形態素を索引語とするのではなく，文書の特徴を表す語のみを索引語として文書に付与する必要がある．

3.1 索引語とする品詞

以下に，形態素解析の結果処理された形態素の中から，索引語とする品詞を示す．

(1) 名詞（「こと」「もの」除く）
(2) 動詞（補助動詞・存在動詞を除く）
(3) 形状詞（形容動詞と同一．平仮名のみで構成されるものを除く）
(4) 形容詞（平仮名で構成されるものを除く）
(5) 接尾辞

主に上記の品詞の語が文書の特徴を表すため，索引語とする．

3.2 結果未知語からの索引生成

2.2 節の (2) にて述べた，ツールが未知語とした形態素は，KyTea の機能により品詞が推定される． KyTea による品詞推定の結果「名詞」「形状詞」「形容詞」「動詞」「副詞」となる未知語の多くが文書の特徴を表す語であったため，本研究ではこれを索引語とした．また，品詞推定の誤処理を考慮し，未知語内の文字種も考慮した索引付与を行う．顔文字等，文書の特徴とならない記号列は索引語として相応しくないため，語を構成する文字の半数以上の文字が記号のものを索引から除外した．

3.3 形態素の連結による索引生成

2.2 節の (1) にて述べた未知語は，形態素解析処理の結果，語の区切りを誤り，複数の語として処理されている．分割された形態素を連結することによって正しい語の形に直し，索引語とする．以下に連結する条件を示す．

(1) 名詞が連続している場合
(2) 英数字や記号が連続している場合
(3) 名詞の直後に接尾辞が存在する場合

過剰な索引生成を避けるため，1つの形態素を起点とした場合の連結長を 5つの形態素までに制限する．

Indexing from Unknown Words for Text Discrimination
Itaru Onishi†, Koichi Yamada†, Hiroshi Kinukawa†
Graduate School of Science and Technology for Future Life, Tokyo Denki University†

3.4 分別処理速度向上のための索引除去

これまで述べた索引生成の全てを分別器生成の際に用いた場合，索引語数が多く，処理にかかる時間が増大する．文書頻度の低い索引語は分別器（決定木）に現れないことから，あらかじめ除去する．

4．文書分別の実験評価

人間が収集できる正解データ数には限度があるため，少数の学習データで多量の分別が行えることが好ましい．実際の利用状況を想定した実験をする．

4.1 実験データと実験方法

分別対象の文書は，文書の内容や体裁が多様であるため，ヤフージャパン株式会社の運営する Yahoo! ブログ [5] に投稿された新着記事を無作為に収集したものの中から，未知語を含む特定のテーマについて記述されている文書を正解データ，そのテーマに関連する語を含む文書を不正解データとし，それぞれ収集した．本研究では，以下の2つのデータセットを用意した．

(A)　正解：H7N9鳥インフルエンザの感染状況(160件)
　　不正解：「インフルエンザ」を含む(1,100件)

(B)　正解：PM2.5による被害状況(250件)
　　不正解：「環境汚染」「大気汚染」を含む(1,000件)

データセットごとに，分別器生成用の学習データとテスト用データに分け，分別実験を行う．また，実際の利用状況を想定し，一般的な交差検定における学習とテストのデータ量比を逆転させた形での交差検定を行う．本研究では，人間による収集が可能な学習データ量の範囲が40から60程度であると想定し，4交差での実験を行った．

特徴的な索引語が分別の指標となりやすく，分別に有効な索引語が明確であることから，決定木を利用した分別を行う．統計ソフト R[6] を用い，CARTアルゴリズムにより生成される決定木を用いる．

また，**3.4** 節にて述べた索引除去に関して，本研究では文書頻度が 3 以下のものを対象とした．

4.2 評価方法と比較対象

前述の特殊な形での交差検定を行い，以下の項目にて精度，再現率，F値を比較し，それぞれの必要性を示す．

(1)　**3.2** 節, **3.3** 節 による未知語からの索引語の有無
(2)　決定木の過学習抑止目的の剪定の有無

4.3 実験結果

実験結果の内，**4.2** 節の (1) の比較を **表１** に，**4.2** 節 (2) の比較を **表２** に示す．

表１．未知語からの索引語の有無による分別性能の比較

データ	未知語	精度	再現率	F値
A	有	75.2%	79.9%	76.5%
	無	75.9%	76.1%	75.8%
B	有	73.9%	76.1%	73.8%
	無	73.5%	67.7%	69.2%

表２．決定木の剪定の有無による分別性能の比較

データ	剪定	精度	再現率	F値
A	有	76.7%	78.9%	76.8%
	無	74.4%	77.2%	75.5%
B	有	76.1%	71.5%	71.5%
	無	71.3%	72.3%	71.5%

4.4 考察

実際の利用を考慮すると，正解の文書が漏れることは避けるべきであり，再現率が重要であると言える．

両データセットにおいても，未知語からの索引語を加味した状態での分別の方が再現率・F値共に良好であった．未知語が分別の結果に影響していることが分かる．データセット(A)における実験では，未知語である「H7N9」という語が分岐条件に関わる決定木が生成されるなど，未知語処理の重要性が明らかとなった．

決定木の剪定については，剪定の有無による性能の差は小さかった．幾度も分岐する決定木を，分岐を1度のみにまで剪定した状態で分別実験を試みても，元の決定木による実験の際と結果の差が小さい．これは語による検索でも，ある程度の文書分別が可能であることを示している．

5．おわりに

5.1 成果のまとめ

本研究では，未知語からの索引生成を行い，付与した索引を用いての文書分別を行った．未知語からの索引生成を行った際の評価が良好であった．

5.2 今後の課題

今後は他のデータセットでも分別実験を試みると共に，文書の検索を行う上での索引付与の手法を考案する．

謝辞

本研究に際して，使用させていただいた KyTea，R の開発者の方々に深く感謝いたします．

参考文献

[1] 後藤正幸，石田崇，鈴木誠，平沢茂一，“高次元ベクトル空間モデルによるテキスト分類問題について：分類性能と距離構造の漸近解析”，日本経営工学会論文誌，Vol. 61，No. 3，pp. 97-106，(2010)
[2] 村脇有吾，黒橋禎夫，“日本語未知語のテキストからの自動獲得”，電子情報通信学会技術研究報告.NLC, 言語理解とコミュニケーション，Vol. 111，No. 119，pp. 37-42，(2011)
[3] KyTea，http://www.phontron.com/kytea/index-ja.html
[4] 大西周，山田剛一，絹川博之，“文書索引生成における未知語の取り扱い方法の比較”，情報科学技術フォーラム講演論文集，2，pp.245-246，(2014)
[5] Yahoo! ブログ，http://blogs.yahoo.co.jp/
[6] R，http://www.r-project.org/