E-018

文書索引生成における未知語の取り扱い方法の比較
Comparison of Processing Unknown Words in Generation of Index Terms on Documents

大西 周† 山田 剛一† 絹川 博之†
Itaru Onishi Koichi Yamada Hiroshi Kinukawa

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，正しい解析を行えず未知語として出力され，索引付けに支障が出てしまう．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，辞書への登録が追いつかなくなると考える．

本研究では，汎用的な文書分別システムの構築を最終目的としており，未知語の間断のない出現による新語の辞書登録不全の問題に対応することを目的としている．今回は形態素解析ツールを用いた文書索引生成をするにあたり，未知語の取り扱い方法の比較と評価を行い，文書分別の際の索引生成に適する形態素解析ツールの選定を行う．

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)と Lucene-gosen[4] (Ver 4.1.0, ipadic 使用，以下 L-gosen と表記) について比較する．

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，**図 1** のように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．

† 東京電機大学大学院　未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University

また，**図 2** のように辞書未登録語がない状態でも，形態素の区切りの誤りによっては未知語と認識される．

鶏肉	名詞−一般	けいにく
さ	副詞−助詞類接続	さ
さ	副詞−助詞類接続	さ
み	動詞−自立	み
の	助詞−連体化	の
天ぷら	名詞−一般	てんぷら

図 1　語の誤分断により別のとして認識される例
（L-gosen を使用）

この	連体詞	この
前塾	名詞	UNK
の	助詞	の
先生	名詞	せんせい
に	助詞	に

図 2　区切り部を含む形態素の例
(KyTea を使用，「UNK」は未知語とみなしたことを示す)

3. 文書分別のための索引付与

3.1 文書分別のための索引とは

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．索引は文書の特徴語であり，索引の語は，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．よって，語として正しく認識されていれば，未知語でも索引として用いることができる．

3.2 付与するべき索引の形

文書の索引は，正しく認識された語である必要がある．

2.2 節の (1) にて述べたように，語の辞書未登録により本来の語が誤って分割される場合がある．分割された語の意味が本来の語の意味とかけ離れてしまう場合，語の意味の不整合により文書分別において誤った結果を生む原因となるため，索引に不適切な形であると言える．図 **1** は，「ささみ」という語が未知語であることにより「さ」と「さ」と「み」で別々の語とみなされている．このままの形では，元の「ささみ」という語の意味が失われてしまうため，索引として不相応であると言える．

2.2 節の (2) にて述べた場合に関しては，ツールによる処理結果が未知語でも，正しく認識された語であれば，そのまま索引として用いることができる．しかし，図 **2** のように区切られるべき部分が残ってしまっている場合は，元の語と意味が異なってしまうため，文書分別の結果に影響が出ると考えられる．

4. 形態素解析ツールの索引生成における適性比較

4.1 適性比較の判断基準

3.2 節より，文書分別をするにあたり，形態素解析ツールの語の区切り誤りが索引付与の際の大きな問題であると考える．区切りの誤りパターンを以下の２つに分類した．

(1) １つの語が複数の形態素に区切られ，かつ元の語の意味が保持できていない
(2) 区切られるべき箇所が区切られていない

これらの誤りの問題が少ないほど文書分別用の索引生成に適していると言える．

4.2 適性評価と実験データ

今回は形態素解析ツールである KyTea と L-gosen の適性比較を行う．L-gosen は MeCab[5] をベースとしており，最小コスト法を用いた解析をする．それに対し KyTea はポイントワイズでの単語分割を行うため語の区切り方が異なり，比較対象として相応しいと考えた．

解析対象の文書はヤフージャパン株式会社の運営する Yahoo! ブログ[6] に投稿された新着記事を無作為に収集したものとした．

文書を 200 件，2 種の形態素解析ツールにかけて解析処理を行う．解析結果の中から 4.1 節の誤りパターンを生じさせている形態素の数にて比較を行う．200 件文書を解析した結果，出力された形態素の数は KyTea が 75,051，L-gosen が 67,087 であった．

4.3 評価結果

表 1 に実験の結果を誤りパターン (1)，(2) それぞれについて示す．また比較の際は，誤って分割された本来の形態素（以下誤解析形態素と表記）の数と，誤って解析された結果生じた，誤った形態素（以下偽形態素と表記）の数を評価指標とする．

4.4 考察

本研究では，形態素解析の辞書未登録による語の区切りの誤りの数を索引生成における適性比較の評価指標とした．結果，誤解析形態素数は KyTea が L-gosen に比べ約３分の１ となり，偽形態素数に関しても KyTea は L-gosen に比べ４分の１程度に抑えられた．よって，KyTea の方が索引生成に適していると言える．また，区切りの誤りが生じる原因の未知語の内，KyTea では 97.8% が，L-gosen では 99.3%が名詞であり，誤りを生む原因の未知語は主に名詞であることも明らかになった．名詞は索引の候補として最も重要であるため，索引付与において未知語処理は重要である．

L-gosen に関しては，カタカナ名詞の解析誤りが KyTea に比べて多くあった．未知語の文字列の一部を辞書登録語と誤認識し，不要な分割をしていた．また，同様に人名の処理にも誤りが多く見受けられた．

KyTea に関しては，不要な分割をした語の数が L-gosen に比べて少なく，多くの人名やカタカナ名詞の未知語が語として正しく処理された．区切るべき部分が残っている形態素は L-gosen に比べ多く出現し，様々な品詞と名詞が連結され１つの形態素として誤解析されていた．しかし，この連結された形態素に関してはすべて**図 2** のようにツールが未知語として明示的に処理するため，後処理をしやすい．また，形態素総数を全体とした構成比としては少ないため，分別の際の結果への影響は小さいものであると考えられる．

5. おわりに

5.1 成果のまとめ

本研究では，形態素解析における語の区切り誤りが文書分別の性能に影響するため，索引付与に適する形態素解析ツールの調査を行った．KyTea と L-gosen の 2 種を比較したところ，形態素の区切り誤りは KyTea が L-gosen の約３分の１となり，KyTea が文書分別用の索引生成に適するという結果を得た．

5.2 今後の課題

本研究の結果をもとに文書分別のシステムを構築し，評価実験を行う．また，他の形態素解析ツールについても調査をし，文書分別用の索引生成における適性を比較する．

謝辞

本研究に際して使用させていただいた KyTea，Lucene-gosen，Yahoo!API 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] Lucene-gosen，https://code.google.com/p/lucene-gosen/
[5] MeCab，http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[6] Yahoo! ブログ，http://blogs.yahoo.co.jp/

表 1　形態素解析における誤解析形態素数と偽形態素数の比較

	KyTea			L-gosen		
	誤解析形態素数	偽形態素数	形態素総数	誤解析形態素数	偽形態素数	形態素総数
誤りパターン(1)	44	94	75,051	254	552	67,087
誤りパターン(2)	48	49		9	19	
計	92	143		263	571	