5P-9

文書分別のための未知語からのインデックスターム生成

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，インデックスターム（索引語，以下 IXT と略記）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，既存の形態素解析エンジンでは，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る．従来，未知語が出現した際は，新たな単語を辞書に登録することを前提としている．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 [2].

本研究では，汎用的な文書分別システムを制作すること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．具体的には文書の形態素解析を行った際に出現する未知語に対し，その未知語及び前後の形態素から IXT を構成する方法を提案するとともに，提案方式を実験評価する．

２．形態素解析ツールにおける未知語

形態素解析ツールによって，未知語を扱う手法は異なる．今回は IPA 辞書を内蔵した Lucene-gosen [3]（バージョン 4.1.0）が出力する未知語を対象として検討をする。

2.1 未知語とその発生

ある文書に対し形態素解析ツールを用い形態素解析する際，辞書に登録されている語では構成できない文字列が現れる場合がある．システムはその文字列を必要に応じ分割し，未知語として出力する．以降では，それらを未知語と呼ぶ．未知語は様々な新語や略語等の間断のない出現に対し，辞書登録が追随できないことにより発生する．

2.2 未知語の特徴

2.2.1 未知語の構成文字種

Lucene-gosen を用いた形態素解析の出力結果として現れる未知語は，登録されている語では構成できない文字列の中から，同一文字種の文字の並びが連結され，語として出力される．そのため，「あア」のように異なる文字種が混合してひとまとまりの未知語とされることはなく，「あ」と「ア」がそれぞれ別の未知語として出力される．

また，例外として，平仮名および漢字は，同一文字種の文字が連続したとしても連結されず，一つ一つの文字が未知語として出力される．

Index Term Generation from Unknown Words for Text Discrimination

†Itaru Onishi，††Koichi Yamada，†††Hiroshi Kinukawa

‡Graduate School of Science and Technology for Future Life, Tokyo Denki University

2.2.2 未知語の発生割合

ヤフージャパン株式会社の運営する Yahoo! ブログ[4]に投稿された記事をヤフージャパンが提供している RSS を用い収集した．その中から，本文の文字数が 200 字以上の文書 1,729,622件を解析した．出力結果のうち，7.62% が未知語であった．本研究では未知語を文字種別の 7つに分類した．その分類と出現割合を**表１**に示す.

表１．未知語の文字種分類とその出現割合

分類名	割合
半角記号	37.09%
半角英字	22.02%
半角数字	19.57%
片仮名	12.85%
平仮名	1.56%
CJK統合漢字	1.25%
その他	5.65%

３．未知語からのインデックスターム生成

本研究では，検索漏れの発生を防ぐため，必要な IXT をできる限り漏れのないように生成することを重視する．

形態素解析ツールが未知語として出力する文字列は，本来の語の区切とは必ずしも一致しない．本来の語が誤って分別され，その一部が未知語として表れている場合もある. その場合, 本来の語の区切りと一致させるため，未知語を前後の形態素を連結させることも必要である．本研究では，**表１**にて示した文字種分類ごとに，前後の形態素の品詞から連結を行うか否かを判定し，その結果の上で IXT とするか否かを判別する．

本来の語が誤って分割され，それら全てが別の辞書登録語とされる場合もあるが，本研究では対象外とする．

3.1 事前処理

IXT 生成のための形態素解析を行う前に, 以下を行う．

(1) 文書中の URL の除去
(2) 文書中にある半角カタカナの全角への変換
(3) 文章中の「(爆)」等，括弧内に漢字が 1 文字存在するものの除去

(1) と (3) は文書内に頻出し，未知語と解析されることが多いため行う．

除去を行うものに関しては，除去した箇所が語の境界であると解析されるよう，半角スペースを挿入する．

事前処理により，処理前に形態素解析を行った際に出力される未知語全体のうち，16.68% を IXT にならないとして除去することができる．

3.2 未知語と前後の形態素との連結による IXT 生成

以下に，連結条件と，連結した語句を IXT とすることの適否とを示す．適を「○」で，否を「×」にて示す．

(1) 未知語が名詞やサ変動詞語幹の一部であると周辺形態素の品詞から判定された場合：○
(2) 未知語が -（ハイフン），/（スラッシュ），.（ピリオド），:（コロン）のいずれかであり，前後に英字又は数字が存在する場合：○
(3) 未知語が数字であり，後に助数詞や名詞が存在する場合：○
(4) 未知語が平仮名であり，未知語と同一の文字が連続している場合：×

3.3 未知語単独による IXT 生成

単独にて扱う未知語の中から不必要なものを以下に示す．
(1) １字の英字
(2) １字，もしくは長音符の連続を含むカタカナ
(3) 直前の母音を伸ばす平仮名の小書き文字
(4) 半角記号の内，**3.2**節 (2) にて記したもの以外
(5) **表１**に示した分類のうち，その他に属するもの

上記以外に出現する，前後形態素との連結を行わない単独の未知語は全て IXT とする．

４．実験評価

4.1 実験データと実験方法

2.2.3項にて用いた文書データを実験データの抽出元とする．ランダムに文書を選び，**3.1**節に示した事前処理を行った後に形態素解析を行う．解析対象の文書数が少ない場合結果に偏りが生じるため，文書中に存在する未知語をランダムに選び，文書数を増やす．これを繰り返し，分類別に 500語の未知語を抽出し，IXT 生成を行う．

4.2 評価方法

評価の指標を以下に示す．正解 IXT 数に関しては処理された未知語の中から IXT にするべきであるか否かを個人の判断により正否を付け，以下の評価数値を出す．
(1) 精度＝[生成＆正解 IXT 数]／[生成 IXT 数]
(2) 再現率
　　＝[生成＆正解 IXT 数]／[正解 IXT 数]
(3) F値＝精度と再現率の調和平均
(4) 総合結果＝出現率を重みとし算出した値

4.3 実験結果

IXT 生成処理の性能を**表２**に示す．

4.4 考察

本研究では，必要な IXT の漏れをできる限り少なくし，再現率を高めることを目的としている．総合結果の再現率は良好な値と言えるが，その中で，半角記号の再現率が少し低い．**3.2**節 (2) に記したもの以外の記号を含む固有名詞を IXT とする処理の未実装によるものが多い．IXT として処理する記号を増加させる必要がある．また，半角英字に関しては，「T字路」などの１字の英字と漢字の複合語処理の未実装によるものが多い．１字の英字は，後ろに特定の漢字列が存在する場合，連結し IXT とする必要がある．これらにより再現率の向上が可能と考える．

精度面に関しては．平仮名において，「だじぇ」等の口語的な用いられ方の仮名小文字単独の未知語が多い．その処理方法の改善により，精度の向上が可能と考える．

５．おわりに

5.1 成果のまとめ

本研究では，文書を形態素解析し，解析結果として現れた未知語の処理を行った．連結の適否判別に基づき，未知語と前後形態素とを連結し，IXT を生成するシステムを提案した．実験の結果，IXT 生成における再現率は 94.7%，精度は 96.4%，F値は 95.4% を得た．

5.2 今後の課題

評価方法に関しては，評価データを増やし，正解の作成方法の見直すことにより，結果の信頼度を高める．

また，他の形態素解析ツールに関しても調査を行い，比較の上，未知語処理の手法の改良を行う。

謝辞

本研究に際して，使用させていただいた 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] Lucene-gosen，https://code.google.com/p/lucene-gosen/
[4] Yahoo! ブログ，http://blogs.yahoo.co.jp/

表２．未知語からの IXT 生成による処理結果

分類名	出現率	データ数	正解IXT数	生成IXT数	生成＆正解IXT数	精度	再現率	F値
半角記号	35.60%	500	64	68	62	96.9%	91.2%	94.0%
半角数字	22.05%	500	500	500	500	100.0%	100.0%	100.0%
半角英字	18.54%	500	408	388	376	96.9%	92.2%	94.5%
片仮名	16.32%	500	423	432	407	94.2%	96.2%	95.2%
平仮名	1.92%	500	182	321	173	53.9%	95.1%	68.8%
漢字	1.42%	500	411	418	407	97.4%	99.0%	98.2%
その他	4.15%	500	0	0	0	(*)100%	(*)100%	(*)100%
総合結果		500	287.0	289.0	277.5	96.5%	94.7%	95.4%

(＊)算出する際に分母が ’0’ となるため計算が不可能だが，100% と見なす．