社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DBS－141 (13)
2007－GN－ 62 (13)
2007／1／26

Webから新語を動的に獲得する形態素解析用辞書拡張方式

三枝 優一† 古井 陽之助†† 速水 治夫†

神奈川工科大学大学院 工学研究科 情報工学専攻†
神奈川工科大学 情報学部††

概要

辞書を用いる形態素解析においては，時代の流れと共に現われ変遷していく口語表現・省略表現・若者言葉等の新語を速やかに辞書に取り入れることで解析精度を高められると期待できる．そこで本研究では，Web上のblogを中心とした文書集合を字種別に切り分け新語候補とし，それらの出現頻度を既に辞書に登録されている語のそれと照合し評価することにより，新語を抽出する手法を提案する．実験では，カタカナのみ，あるいは漢字のみで構成される新語は，複合語を含め80%以上の精度で抽出することができた．また，出現頻度の低い新語についても一部抽出することができた．今後の課題としては，収集した新語の動的な品詞同定と，新語を辞書に取り入れることによる形態素解析精度向上の検証が挙げられる．

A dictionary extending method for morpheme analysis that acquires neologisms dynamically from Web

Yuichi SAEGUSA† Younosuke FURUI†† Haruo HAYAMI†

Graduate School of Engineering, Kanagawa Institute of Technology †
Information Faculty, Kanagawa Institute of Technology ††

Abstract

We believe that a dictionary promptly adopting neologisms produced by new generations will achieve higher precision in morpheme analysis. Therefore, we propose a new method that collects text data from the Web (mainly from blogs), separates the text into candidates of neologisms, and estimates the frequency of each candidate in comparison with that of the words in the dictionary, to determine neologisms. In our experiment, the precision reached 80 percents or higher, concerning neologisms only of Katakana, only of Kanji, and their compounds. In addition, this method also extracted some of neologisms of low frequency. The future work includes investigation of speech identification and experimental evaluation of morpheme analysis improved by the dictionary that adopts neologisms.

1. はじめに

自然言語処理の第一段階にあたる処理として形態素解析がある．形態素解析とは，自然言語で書かれた文を形態素(意味を持つ最小単位)に分割し品詞を見分ける処理である．

従来の形態素解析用辞書は，新聞記事に代表される文書集合を解析の元データ(コーパス)に用いて作られている．しかし，新聞記事のコーパスは用例

の種類や，量が限られ不十分である．[5] このため，新聞コーパスを用いて非定形的な文書集合を解析する際は，辞書に無い語による誤解析が引き起こされることが多い．

一方，インターネットの普及に伴い，個人によるWeb上での情報発信が一般化した．blogや電子掲示板などの普及がその後押しをすることにより，多くの人々の多種多様な意見が自由な記述で公開され，また誰もがそれを閲覧することができるようになった．こうした背景から，Web における文書集合は現在，自然言語処理において重要な解析対象の一つとなっている．

Web ページには，日常よく用いられる口語表現や省略表現，若者言葉など，従来の形態素解析用辞書[1]に含まれていない語(以下，**新語**と総称)が大量に含まれる．Web においては新語が急速に広まるため，そのような動きを適宜捕えることにより，効率良く新語を獲得することができる．

そこで本方式では，Web から新語を随時獲得することで，形態素解析の精度を向上させることを可能にする方式を提案する．

具体的に，本方式ではリンクを辿ることによってWeb ページから Web ページへと遷移しながらデータを収集する Web ロボット技術を用いた．これによりWeb 上の文書を周期的に収集し，字種別，形態素解析結果別に切り分け，新語を獲得する．

本研究では，Web ページから文書集合を収集し，その中から新語を獲得する実験システムを構築し，さらに評価実験を行った．本稿では，このシステムの概要および評価実験の結果を報告する．

以下，2 章で先行研究について紹介した後，3 章でシステムの構成について概説し，4 章で評価実験について述べる．5 章では実験結果を考察し，6 章で総括する．

2. 先行研究と本稿の位置づけ

Web 上から文書集合を収集し新語獲得を狙う研究は既にいくつか取り組まれている．

文献[2]では，あるドメインに属するWeb 文書集合から専門用語の獲得を行っている．実験の結果，専門用語と判断できる複合名詞や名詞句などを獲得できたという報告がなされている．本稿では，収集対象が専門用語に限らないことと，無作為に取り出した多量の Web ページを対象とするため実験結果の傾向が異なることが期待できる．

また，文献[7]では，「新聞記事テキストでは、書き手(話し手)の表現がそのままの形で再現されているわけではない。新聞社の側の責任で整理された言語で書かれていると考えるべきである。」と述べている．本方式では，書き手の表現がそのままの形で現れるWeb の文章集合を対象とするため，実験結果から新語が特定できればこれを裏付けることができる．

また，本研究では Web における新語の動きを適宜捕らえ，獲得した新語を解析に反映させる要件を満たすため，逐次データベース辞書に獲得した新語を登録できるよう文献[6]で示される解析システムを用いた．

3. 実験システム

3.1 システムの構成

本実験システムは，大きく下記の三つの機能に分けられる．

- Web ロボット部
  - ➢ Web を巡回し文書データを収集．
- 新語同定部
  - ➢ 収集された文字列から新語を抽出．
    - ✧ 文字種による新語同定
    - ✧ 解析を使用した新語同定
- DB 辞書を用いた形態素解析部
  - ➢ 新語をDB辞書に取込み解析に反映．

本実験システムの構成図を図 3-1 に示す．

[1] 本研究における従来の形態素解析用辞書は IPA 品詞体系辞書を想定した．

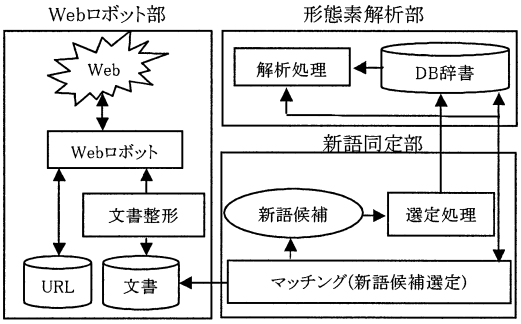

図 3-1. 実験システムの構成図

3.2 Web ロボット部

一般に Web ロボットとは，開始地点として任意の URLを設定すると，そのWebページを取得し，さらにそこからのリンクを辿ることによってWebを巡回しながら様々なWebページを収集する機能を持つプログラムである．

本実験システムの Web ロボット部は，Web を巡回しながら，Web ページに含まれるテキストデータを収集し，文書 DB に格納していく．取得した Web ページが HTML 形式であれば，アンカータグ(A タグ)として記述されたリンク先の URL を抽出し，未読 URL として重複なく URL DB に格納する．収集された URL は，既読 URL とし，以後これを未読 URL が無くなるまで繰り返す．また，本稿では主に日本語の新語を取得することを目的としているため，ひらがな又はカタカナを 1 文字も含まないWebページ，不要な拡張子へのリンクは読み飛ばした．

また，収集したテキストデータに対し，文字コードの変更，タグやスクリプトの削除，文書中に含まれる URL やメールアドレスの削除，不要な文字・改行・空白の削除などの文書整形処理を行い，以後の処理を容易にした．

3.3 新語同定部

新語同定部では，Web ロボットによって収集された文書から，新語を抽出する．本研究では，IPA 品詞体系辞書に掲載されていない言葉である新語を特定する手法として，以下の三ステップを経る．

1. 字種，形態素解析による切り分け
2. 新語信頼度の計測
3. 新語特定

3.3.1 字種，形態素解析による切り分け

字種による特定では，同じ文字種の形態素が連節するならば連結する規則を用いて，新語特定の第 1 段階とする．本方式で対象となる文字種は，カタカナ，ひらがな，漢字の 3 種類とする．アルファベット列は IPA 品詞体系辞書と比較できないため，ステップ 2 以降対象外とする．

また，精度を保つため，同一の Web ページに現れる文字列は重複を排除する．

次に字種切りした各文字列を IPA 品詞体系辞書に記載されている形態素とマッチングを行い，辞書に無い文字列を新語候補として選り分ける．

3.3.2 新語信頼度

前項では，字種により新語候補を絞り込んだ．この段階では，意味を成さない新語候補が含まれている．本方式の案出前は，出現頻度が高い新語候補ほど信頼度が高いと仮定し信頼度が高い新語候補を抽出したが．どの出現頻度までを信頼できる閾値として設定するか特定しづらいこと，出現頻度が低い新語候補を抽出することができないことから次に述べる手法を提案する．

提案手法は，辞書を参照して新語候補を選りだす前に，各新語候補に対し信頼度を計算し，絞り込みを行うものである．

1. 新語候補の出現率の計算

辞書を参照する前の文字列全てを対象に，取得した全 Web ページ数に対して，ある新語候補がどの程度の割合で出現したかを計算する(以下，出現率と呼ぶ)．

2. 区間信頼度の計算

計算した全ての新語候補の出現率を降順に並び替え，出現率が高いほど信頼度が高いと仮定する．

出現率が高い新語候補から適当な数ごとに区間にわけ，区間信頼度を計算する．辞書を参照した際に，一致した文字列は信頼度を 100%とし，辞書に無い文字列は信頼度を0%として区間信頼度を計算した．

区間信頼度 R はある区間 n に対して，文字列数 N_all と，ある区間内における辞書との一致文字列数 N_dic を用い，次式で与えられるものとする．

$$R_n = \frac{N\_dic_n}{N\_all_n} \qquad (1)$$

これを用い，区間信頼度 $R_1, R_2, R_3 \cdots R_n$ を全区間において計算する．

3.3.3 新語特定

前項で，計算した区間信頼度のデータから，実際に新語を取り出す処理を行う．

区間内文字列数による新語特定結果への影響を考慮し区間内文字列数を 50 と 100 の 2 パターンを実験的に用意した．各パターンともに，区間信頼度から対数近似曲線を計算し，それぞれの区間において対数近似値を超え且つ，全区間信頼度の平均値を上回る区間を選定した．

選定されたそれぞれの区間において，出現率が上位である文字列から，当該区間の信頼度の割合分だけ文字列を抽出する．抽出された文字列中に辞書に無い文字列が含まれていれば新語として獲得する．

獲得新語数 N_re は，ある区間 n において次式で与えられる．

$$N\_re_n = N\_all_n \times R_n \qquad (2)$$

3.3.4 漢字かな混じり文

多くの場合，漢字とひらがなは，字種による切り分けでは送り仮名などを不適当に切り分けてしまうことが挙げられる．そこで，漢字とひらがなの並びにおいて，漢字の直前で切り分ける手法を用いた．以後，上記手法を用い，さらに形態素解析にかけ新語部分を抽出した．

4. 評価実験結果

Web ロボット部では，blog を中心とし約 170 万サイトから URL を収集し，約 18 万サイトから文書を収集した．これをコーパスにし以降の実験を進めた．

新語同定部では，まず収集された文字列を字種別に切り分けた．その結果は表 4-1 に示す．

表 4-1. 字種別文字列数

	字種で切り分けた文字列数	割合(%)
カタカナ	322347	11.37
ひらがな	1540771	54.33
漢字	720699	25.41
英単語	251921	8.88
合計	2835738	

次に，辞書を参照し一致文字列数を計算した．その結果を表 4-2 に示す．

表 4-2 字種切りのみでの辞書一致数

	IPA 辞書との一致文字列数	字種別割合(%)
カタカナ	11518	3.57
ひらがな	10937	0.71
漢字	44714	6.20
合計	67169	

辞書を参照し一致しなかった文字列を新語候補とし，その数を計算した．その結果を図 4-3 に示す．

表 4-3 切り分けのみでの新語候補

	新語候補数	字種別割合(%)
カタカナ	310829	96.43
ひらがな	1529834	99.29
漢字	675985	93.80
合計	2516648	

区間信頼度を計算し，降順に並び替えた．グラフを図 4-4 に示す．

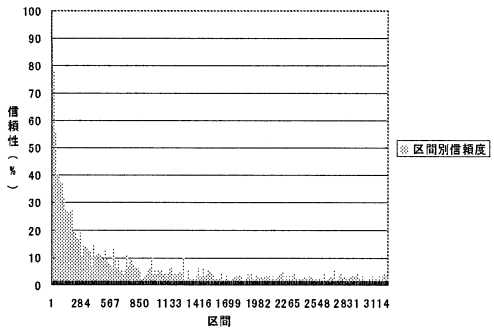

図 4-4 区間信頼度 （カタカナ例）

図 4-4 に，対数近似曲線と，平均信頼度直線を加えた．グラフを図 4-5 に示す．

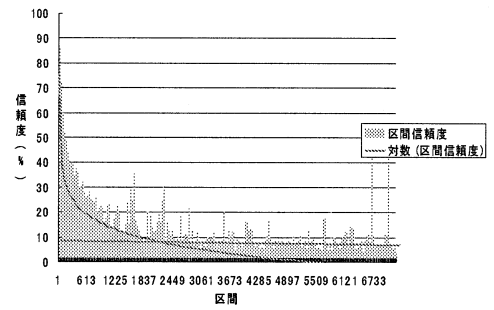

図 4-5. 区間信頼度と対数近似（漢字例）

実際に獲得した新語例を，表 4-6に示す．

表 4-6 新語獲得結果例

カタカナ	漢字
ブログ	過去問
アバター	湯壷
アイコン	女子高生
ワンコ	画嬢
コスメ	何気
カキコ	倖田來未
ヤバイ	爆睡
セレブ	特化
ディープ	食玩

また，漢字かな混じり文から獲得した新語を表 4-7 に示す．

表 4-7 新語獲得結果例

まっしぐら
ほんま
ぶろぐ
すんまへん
とりま
うぇぶ
島んちゅ
くだせぇ
どす

5. 考察

本研究では，Web 上の文書集合から新語を取得し，形態素解析用辞書に逐次追加することで解析の精度が向上するという仮定を基に実験システムを構築した．Web 上の文書集合に対する，解析の精度向上には，新語をいかに獲得できるかが結果に直結する．

本実験で獲得した新語数を表 5-1 に示す．

表 5-1 本実験で獲得した新語数

獲得した新語数		
カタカナ	ひらがな	漢字
5431	1503	17868

獲得された新語には以下の4種類がある．

1. 省略表現
2. 若者言葉
3. 複合語
4. 人名
5. 抽出誤り語

1～2 については，これらの新語を辞書に素早く登録し，解析結果の向上に利用する．動的に新語の獲得を行うことにより，Web を巡回しながら順次生み出される新語を獲得することが期待できる．

3 の複合語については，会社名や地名などの固有名詞や専門用語，また四字熟語などを獲得することができた．しかし，複合語は既に辞書に載っている言葉の組み合わせである場合がある．よって，その

表 5-2 本手法の字種切りによる新語獲得精度(複合語を含む)

字種		カタカナ		ひらがな		漢字	
区間内文字列数		50	100	50	100	50	100
区間数		6447	3223	30815	15407	14413	7206
平均区間信頼度(%)		1.79	3.57	0.35	0.70	3.04	6.15
対数近似	適合率(%)	98	100	55	72	87	82

まま新語として辞書に登録するのは辞書の信頼性を下げる恐れがある.

4 の人名については, 一般的な人名ほど苗字と姓にわけて既存の辞書と参照することできるので新語として登録できない. しかし, 芸名のような特殊な人名を獲得することができた.

5 の抽出誤り語については, ひらがな表現に多く見られた. 原因としては, 本方式ではひらがなのみの文字列に対して, 漢字かな混じり文からの新語獲得に伴い特定できる一部分しか, 対応していないことが挙げられる. また, 漢字かな混じり文では, "今日わ"や"だょぉ"のような新語を獲得したことから, 精度の問題点が残った.

また, 本実験の新語獲得精度として, 適合率を算出した. カタカナ, 漢字の新語は複合語を含み80%以上の適合率で獲得できた. しかし, ひらがなの適合率が低いことがわかる.

また, 区間内文字列数は大きい方が適合率が良い. 100 以上の区間内文字列数についても検討する必要がある.

6. まとめ

本研究では, カタカナや漢字のみで構成される新語は複合語を含み, 適合率80%以上の精度で新語を動的に獲得できた. また低頻度の新語も獲得できた. しかし, ひらがな文字については今後も検討が必要である.

また, 本研究では複合語をそのまま新語として扱った. 複合語を分かつ処理を加えて精度を計ることを今後の課題とする. さらに, 獲得した新語を辞書に登録するにあたり, 品詞情報などを推定し, 付加させることも同時に今後の課題とする.

・参考文献

[1] 太田晋,美馬秀樹 "用語抽出技術を利用したテキスト分類" 情報処理学会 情報学基礎研究会, 自然言語処理研究会, 情報処研報 2004-FI-76,2004-NL-163,Vol.2004,pp.61～66

[2] 山本英子,池野篤司,濱口佳孝,井佐原均 "検索支援に向けた Web 文書集合からの用語獲得" 情報処理学会, 自然言語処理研究会, 情報処研報 2004-NL-164 , Vol.2004 , pp171～176

[3] 濱口佳孝,池野篤司,井佐原均 "Web からの情報抽出・検索システムにおける全文検索" 情報処理学会 情報学基礎研究会, 自然言語処理研究会, 情報処研報 2004-FI-76,2004-NL-163, Vol.2004,pp9～14

[4] 情報処理学会編「データベース」オーム社 (2004)

[5] 関口洋一,山本和英 "Web コーパスの提案" 情報処理学会 情報学基礎研究会, 自然言語処理研究会, 情報処研報 2003-FI-72,2003-NL-157,Vol.2003,pp123～130

[6] 三枝優一,古井陽之助,納富一宏,速水治夫 "形態素解析におけるデータベース解析方式の提案" 情報処理学会 マルチメディア,分散,協調とモバイル シンポジウム論文集(II) IPSJ Symposium Series Vol.2006, No.6,pp573～576

[7] 後藤斉 "コーパスとしての新聞記事データ - 終助詞かしらをめぐって-" 東北大学言論学論集,Vol.5,pp.37～46,東北大学,1996