時系列データを用いたタイトルタグからの新語抽出法の提案

下山剛司* 秋岡明香† 村岡洋一‡
早稲田大学 電気通信大学/早稲田大学 早稲田大学

1 はじめに

世の中では，"Wii"といった商品名や"のだめカンタービレ"といったドラマ名，携帯電話を意味する"ケータイ"と言った略語など，新語と呼ばれる新しい言葉が絶えることなく出現している．このような新語の含まれる文章に対して形態素解析を行ったとしても，新語をうまく認識できないことが多く，新語を動的に取得し辞書に組み込む技術が研究されている [1]．また，新語の発生過程をたどれば，オピニオンリーダの発見ができる可能性もあり，新語を Web から動的に取得することは有用である．新語を動的に取得する技術としては，沖電気と NICT が共同で開発した，Web ページ 4 万ページを一日で処理し最新用語を継続的に獲得する技術がある [2]．

しかし，Lawrence らの報告 [3] [4] によると，1997 年に存在する全世界の web ページは，3.2 億ページと推測され，1998 年には 8 億ページになると推測されている．また，山名ら [5] らの報告によると，2006 年 2 月の時点で，全世界の web ページは 150 億ページ存在すると推測されている．つまり，Web 上の文書は爆発的に増加することが予想されており，このような膨大なデータ全てを対象に新語を抽出するのは非常に困難である．

そこで本研究では，HTML 文書本文の要約が書かれていると期待されるタイトルタグを対象に，独自の手法を用い高速な新語の抽出を試みた．

2 提案手法

本研究では，2007 年 9 月現在で，日本語によって記述された 58,252,033 件の HTML 文書内に含まれるタイトルタグから新語抽出を試みた．その際，2006 年 9 月，2007 年 3 月時点に収集された，日本語によって記述された HTML 文書群も使用し，2007 年 9 月のデータから新語抽出を行った．提案手法について以下で詳しく述べる．

まず，それぞれの年月の HTML 文書中のタイトルタグに対して形態素解析を行い，連続名詞，未知語を新語候補として抽出する．この一連の流れを新語候補フィルタリングと呼ぶことにし，ここで抽出されたものを新語候補群と呼ぶことにする．

次に，それぞれの年月における新語候補群を比較し，出現頻度の増加率と出現数が突出しているものを抽出する．ここで，抽出される新語の信頼度を，出現頻度の増加率に現在の出現頻度を重みづける式 (1) で定義する．この式中の tf_{pn} は過去のデータにおけるある単語の出現頻度を示し，tf_{cn} は新語抽出の対象としている年月におけるある単語の出現頻度を示す．式 (1) における l_n を新語信頼度と定義し，値が高いものほど新語の信頼性が高いものと仮定し，新語抽出を行う．この一連の流れを新語抽出フィルタリングと呼び，ここで抽出されたものを仮新語と呼ぶことにする．

最後に，仮新語間で重複して出現している言葉を排除する．これは，Web ページの増加に伴い，出現数も増加している新語とは考えにくいインパクトの低い言葉を排除するために行う．この一連の流れを，インパクトフィルタリングと呼び，ここで抽出されたものを提案手法で抽出できた新語と呼ぶことにする．

$$l_n = \begin{cases} \frac{tf_{cn}-tf_{pn}}{tf_{pn}} \log_{10} tf_{cn} & (tf_{cn} > tf_{pn}) \\ 0 & (tf_{cn} \leq tf_{pn}) \end{cases} \quad (1)$$

3 結果と考察

提案手法を用いた結果，表 1 に示す実験環境下で，収集済み HTML 文書約 6000 万件を対象に，約 2 日で新語を抽出することができた．2007 年 9 月における新語信頼度が上位 10 件の新語を表 2 に，新語抽出例を表 3 に示す．

提案手法を用いることによって"セカンドライフ"

*Takeshi Shimoyama, Waseda University
†Sayaka Akioka, The University of Electro-Communications/Waseda University
‡Yoichi Muraoka, Waseda University

表 1: 実験環境

OS	openSUSE10.3
CPU	Intel(R) Core(TM)2 Extreme CPU X9650 @ 3.00GHz
Memory	4GB

表 2: 新語信頼度が上位のキーワード 10 件

順位	キーワード
1	温泉旅行
2	セカンドライフ
3	金沢グルメナイト飲食店お酒情報満載
4	金沢グルメ
5	gifulog
6	岐阜様々情報お伝え
7	フォートラベル
8	ライブチャットチャットレディ
9	千代田区総合ホームページ
10	HomePLAZA

や"Wii"といったような新語を高速に獲得できた．しかし，"千代田区総合ホームページ"といったような新語とは考えてにくい単語群も抽出されるという課題が残った．これは，継続的に HTML 文書を増やすのではなく，あるタイミングにおいて大量に HTML 文書を更新，追加がされた場合に，提案手法では，誤って新語として抽出してしまうケースがあるからだ推測する．そのため，同ホスト内で HTML 文書の更新，追加数をチェックすることでこのようなケースで抽出される新語候補群を排除するような仕組みが必要であると考える．

表 3: 新語抽出例

順位	キーワード
2	セカンドライフ
22	ウィー
36	アレス
39	wii
50	アセットアライブ
277	ビリーズブートキャンプ
391	ドラゴンクエストモンスターズジョーカー
641	ワンセグチューナー

4 まとめと今後の課題

本稿では，日本語で記述された約 6000 万 URL の HTML 文書，3ヶ月分を対象に，タイトルタグから新語の抽出を行った．提案手法を用いれば，約 2 日で新語を抽出できることがわかったが，精度の面で課題が残った．

今後は，同ホスト内での HTML 文書の増減を考えた新語抽出法を提案，実験し，高速でかつ，より精度の高い新語抽出を行える手法を確立したい．

謝辞

本研究の一部は，文部科学省リーディングプロジェクト e-Society 「基盤ソフトウェアの総合開発」プロジェクトのサブプロジェクトである「インターネット上の知識集約を可能にするプラットフォーム構築技術」の支援により実施された．

参考文献

[1] 三枝　優一, 古井　陽之助, 速水　治夫:"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", 情報処理学会 Vol.2007, No.6 pp. 77-82 (2007)

[2] Web ページから新語を獲得する技術 ,http://www.oki.com/jp/Home/JIS/New/OKI-News/2005/07/z05039.html

[3] S.Lawrence, C.L.Giles:"Searching the World Wide Web", Science, Vol.280, No.5360, pp.98-100 (1998)

[4] S.Lawrence, C.L.Giles:"Accessibility of Information on the Web", Nature, Vol.400, pp.107-109 (1999)

[5] 加藤真, 山名早人:"Fact of the Web：30 億ページのウェブの解析",DEWS2006 3B-i6 (2006.3)