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

2003－FI－ 72　(17)
2003－NL－157　(17)
2003／9／30

Webコーパスの提案

関口 洋一 †　　山本 和英 †

{sekiguti,ykaz}@nlp.nagaokaut.ac.jp

Web をコーパスの情報源とした Web コーパスの構築手法を提案する．一般的に用いられている新聞コーパスの量やそれに伴う用例の少なさは否めない．そこで，我々は Web に着目した．Web を用いることで量的な問題を解決できるが，そのまま用いたのでは表現そのものや，文の構造に問題がある．そこでコーパスを質の面から検討を行う．質改善の手法として，HTML タグや日本語文章の書法を用いて改善を試みる外面的質の考慮を挙げる．さらに記号を多用した文や話しことばの崩れた文を削除し，文字種の割合を示す字面比を用いて文を削除する等の内面的質を考慮する手法を提案する．構築した Web コーパスに対して 2 種類の実験を行った．1 つめは，異なり単語数やシソーラスを用いて単語の特徴を観察した．2 つめは，有用性を調査するため，格フレームを用いて調査を行った．その結果，異なり単語数，格フレーム数ともに新聞や未処理の Web テキストを上回るコーパスを構築できた．

キーワード: Web コーパス, 新言文一致体，字面比，顔文字，格フレーム

Web Corpus Construction with Quality Improvement

Youichi SEKIGUCHI †　　Kazuhide YAMAMOTO †

{sekiguti,ykaz}@nlp.nagaokaut.ac.jp

We present a method for construction of a Web corpus. There is a quantity issue in a newspaper corpus as we use it as a text corpus for natural language processing. We use a collection of Web pages so that we can solve lack of resource amount. However, some of the Web texts have a low quality. We then propose some methods to reduce some of these texts out of the Web corpus. The methods include sentence determination using a part of HTML tags, and filtering out-of-range sentences by proportions of each character type. We have confirmed that our Web corpus outperformed a newspaper corpus, in terms of number of words and case frames. We also show that our Web corpus is also superior to unprocessed Web texts.

key words: Web Corpus, spoken style texts, character type proportion, smiley,case frame

1　はじめに

一般的に用いられている新聞コーパスは，その量が限られ不十分であり，用例の種類や数の不足などの問題がある．そこで，我々は Web に着目し，Web ページを用いてテキストコーパスを構築した．

Web は日々増加しているので，量もほぼ無限に存在すると言える．すなわち量的な問題は解決する．しかし，Web ページの中にはコーパスとしてふさわしくないものが存在する．例えば，製品紹介ページに含まれる固有名詞や値段，寸法などの情報がそれにあたる．また，日記などのサイトでは崩れた文面が多い．これらが混ざってしまうとコーパスの質を低下させる要因につながる．

そこで本稿では，Web コーパスを構築するにあたって，質を高めるための手法を提案する．提案する手法は，外面的質の向上および，内面的質の向上の 2 種類である．

†長岡技術科学大学 電気系
Department of Electrical Engineering,
Nagaoka University of Technology
http://nlp.nagaokaut.ac.jp/

外面的な質の向上とは，文の表層だけで処理できる作業のことを指す．具体的には，HTML タグを利用して Web の文を 1 行 1 文に整形する作業や，表現が完全に一致する文の削除，文字種の割合を考慮した規格外文の削除を行う．また，内面的質の向上とは，Web に見られる特徴的な文体の削除，Web 特有の表現を用いて行う文の加工/削除，文の雛型を発見し削除するという作業を指す．質改善が終わった Web コーパスと既存の新聞コーパスを表現の豊富さ，単語の網羅性という 2 つの観点から比較，検討を行い，Web コーパスの特徴について述べる．

以下，2 節で本研究の位置付けを行い，3 節で理想のコーパスについて言及する．さらに，4 節でコーパスの構築方法を説明し，5 節で評価実験について説明をする．

2　本研究の位置付け

コーパス構築の研究において，Web をコーパスとして扱っている研究はいくつか見られる．Jones らはタガログ語の Web 文書をシステムに与え，単語の 1-，2-gram を生成し，検索語とした．その検索語より得られた検索結果のうち言語フィルタを通過したものをコーパスとして登録するシステムを提案している [1]．しかし，コーパスの構築手順に着目している研究のため，構築するコーパスの質に焦点を当てている本研究とは異なる．

また，Oravsan らはコーパス構築処理に着目して，Web からコーパスを構築する際にサーバー・クライアント構造を用いた拡張性の高いコーパス構築支援ツールを提案している [2]．コーパスを構築するのは非常に大変な作業であるためこのような研究は有用であるが，コーパスの質を考慮していないため，本研究とは論点が異なっている．

本研究は，Jones らや Oravsan らの研究と異なり，コーパスを構築する手順や収集方法については言及はせず，Web ページを収集した後に「質」を高めるための方法を論ずる．

3　Web を良質のテキストコーパスに

後藤 [4] は，「新聞コーパスに含まれる新聞記事テキストデータを直ちに現代日本語の多様な使われ方を代表するサンプルコーパスと見なすことは難しい」と述べている．

テキストコーパスは，統計的手法による品詞の付与や共起関係抽出による意味的曖昧性の解消，文法規則や確率文法の確率値の学習，用例を用いた機械翻訳など，多岐に渡って用いられる．では，それらの研究で必要とされるコーパス，即ち良質のコーパスとは何か．統計的手法による品詞の付与であれば，周辺の語や品詞との共起によって尤度が推定されるため，ひとつの語に対して接続する周辺文脈がいくつも存在するコーパスである．また，共起関係抽出による意味的曖昧性の解消であれば，ひとつの単語が様々な使われ方をしていて，かつ，ある程度の頻度が存在するコーパスである．

即ち，「良質」のコーパスとは，人間が日常で使っている言語において，できるだけ広い範囲の単語を網羅し，さらにその語の用例を多く包含するものだと考える．しかし，いくら単語や用例が網羅されていても，本来使われることの無い文体や記号，特定の語などに偏った頻度情報を有しているコーパスは研究の対象とはなりにくい．そこで我々は，コーパスに含まれる全ての文が研究の対象となり得るコーパスを「良質」のコーパスと定義し，この構築を目標とする．

Web コーパスを構築するにあたり，具体的な目標を作成するため，松本ら [7] が提示するコーパスの特徴の一部を用いる．まず，抽出する文書の分野は限定せず，多様性を求める．また，文体はできるだけ日常で使われるものに限定する．すなわち記号で構成されていて書き手の感情を表す顔文字 ((^^) など) や，話しことばの崩れたような文体を避ける．さらに，ページの著者が意図している表現を崩さないために，改行位置の特定や文の結束性を考慮する．本来の共起情報が崩れてしまう可能性があるため，本文として表示していないコメント部分や画像の説明などは削除する．また，完全に一致する文は，単語や品詞情報のバランスをも崩す恐れがあるため削除する．本稿では，以上に述べた内面的および外面的な手法を用いて質の改善に取組み，「良質」のコーパスを構築する．

4　Web コーパス

4.1　コーパス構築の手順

実際に Web コーパスを構築する際の手順について述べる．本稿で作成したシステムの全体構成を図 1 に示す．以後，各モジュールについて説明を行う．

まず，基準となる任意の URL を与え，そのページに含まれるリンク情報を取り出し，URL データベースに保存する．次にデータベースから別の URL を取り出し，得られたページから新しいリンク情報を抽出する．この作業を繰り返し行い，データベースが適度な容量 (数 MB) になるまで行う．このとき，以下に定める制限を適用した．

- ドメインが .jp で終わるものを対象．
  - 日本語のコーパスを構築するため，.jp ドメイン以外の Web ページは除外した．

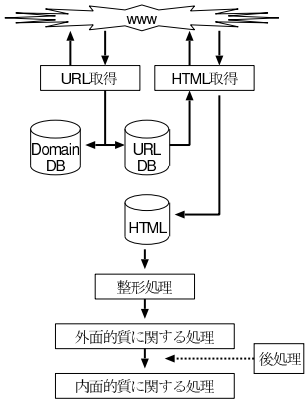

図 1: システム概要

- 拡張子は.html または.htm の文書を対象.
- 文字コードとして，euc-jp, iso-2022-jp, shift-jis, x-sjis が<meta>タグ内に表記されているページを対象.
- 同一ドメインからの取得は 100 ページ以内.
  - 企業などのページは非常に数が多い．異なった URL でも同じページを参照している場合がある．その様な同一ページを複数回取得することを回避するため，上限を定めた.

次に，構築された URL データベースより，取得対象となる URL を無作為[1]に取り出し，HTML ファイルを取得する．取得した HTML ファイルに対してタグ除去を行うが，その際「外面的質」に関する処理 (4.2 節) で用いるタグは残しておく．また，不必要な空白や改行の連続を削除する．これらの整形処理を施した後,「外面的質」に関する処理及び,「内面的質」に関する処理 (4.3 節) を行う．本稿では，外面的質の後処理を行うために以下の規則を適用した.

- 文末に句点を含む行に限定.
  - 単語の羅列などを含む行を取り除くため.
  - 文特定の処理 (4.2 節 (B)) 誤りを含む行を取り除くため.
- URL やメールアドレスを含む行を削除.
  - URL やメールアドレスは Web 特有のものであり，通常の書きことばや話しことばには使用されないため.
  - 文字種頻度の割合を狂わす危険性があるため.
- 1 行あたりの文字数を 150 文字以内に制限.
  - 1 行が長い文は，文の特定処理で誤った可能性が高い．そのため，単語の連続から成りかつ，句点を含む文などを削除するために長さを制限した.

[1]実際には，ハッシュを用いているため真意の無作為ではない.

4.2 外面的質に関する処理

(A) 完全一致文の削除

HTML ファイルには，本来閲覧ソフトに表示されない文や文字も含まれている．特に画像のサムネイル表示部 (注釈部分) などはその典型である.

サムネイル部には本文の説明と重複する場合があり，両方をコーパスとして登録することは，偏った頻度を与え質の低下を招く．したがって，同一ページ内に完全に一致する文が現れた場合，ひとつを残して，他を削除することにした.

(B) 文の特定

Web ページは HTML で表記されているため，ソースファイルを見ればその構造はつかみやすい．しかし，HTML は本来それが持つ意味通りに記述されていないことがしばしば見受けられる．特に <table>, <tr>, <td> タグは本来表を作ることに用いられるが，ページの体裁を整える場合に使用される場合も多い．この結果，これらのタグで構成されたページは文がセルにより切断されている場合が多く，そのまま利用したのでは本来の文の形が崩れてしまう.

そこで，本稿では HTML タグ自体を全て信用するわけではなく，一部分を用いて文を特定し，文末で改行することにした．文を特定するためのパターンは以下の通りである.

- 句点 (。)+
- 句点 (。)+</**>(全ての終了タグ)
- 句点 (。)+</p>
- 句点 (。)+</td>

- (行末文字)+
-

ここで行末文字とは，

)，)，>，＞，?，？，!，！，♪

の 5 種類 9 文字を対象にしている．

(C) 字面比の考慮

日本語で文章を書く際には，漢字・平仮名・カタカナ・英字・数字の 5 種類の文字を用いている．Web の文書においては記号が頻繁に使われる傾向がある．そこで，記号も文字として扱い 6 種類の文字種で字面比を考慮する必要がある．

それらの文字種の比率のことを「字面比」と呼び，例えば，字面比が良いとは，文字種の比率が一般の日本語のそれに近いということである．しかし，これを正確に測定することは難しい．

本稿では，毎日新聞<2>の 1 年分を基準にした上で，下記の字面比を超えたときは，その行を削除することとした．

数字>40%，英字>40%，
一般記号>30%，特殊記号>20%

ここで，一般記号とは文章によく見られる記号のことで以下のものを指す．

。.,、ー!??!

また，特殊記号は，以下に示すような通常表記に使われないものを指す．

☆♪■□○●△

平仮名やカタカナ，漢字の比率においては，文章を表現する上で自由度が高いため，本稿では対象外とした．この処理を行うことで，例 1 に示すような文が削除できる．

> ★★★★★腰痛こんにゃくゼリー。
> (^◇^) ノ」とのお答えでした
> ぼーーーーーーーっとしながら、
> 720 × 486/59.94i、720 × 480/59.94i
> をサポートしています。
> Anthropology resource son the
> Internet から。

例 1: 字面比による削除例

(D) 引用記号の対処

Web では掲示板などで他人の書いた文章を引用記号とともに参照することが良くある．よって，引用記号を外すと同一の文が同じページ内にあることが多い．そこで引用記号を削除した上で，完全一致文の削除工程 (本節 (A)) を再び行う．これにより引用記号を伴う文を削除することができる．

引用記号は複数あり，使用する記号も人によって異なる．本稿では良く見られる引用記号を処理の対象とし，以下に示す引用記号が文頭に付属しているものに適用した．

> > $ # ＞ # ＄

4.3 内面的質に関する処理

(A) 新言文一致体の削除

Web ページの表現は，「思い付くまま，感じるままに相手に話しかける」というものが多い．これを佐竹 [3] は，新言文一致体と呼んでいる．新言文一致体の例を例 2 に示す．この新言文一致体は，書き手の自然な表現と見ることもできる．しかし Web に存在する例 3 の様な極端なものは，コーパスの質の低下を招くと判断した．そこで，次の削除規則を作成し，例 3 に示す様な表現が削除されることを確認した．

- 「〜」が 3 つ以上連続で現れる文
- 「ー」が 3 つ以上連続で現れる文
- 「っ」が 2 つ以上連続で現れる文
- 「?」や「!」が行末で 3 つ以上連続する文

> おっと洗濯も簡単，すぐ乾いちゃうしね．
> それがめちゃ速いのです．
> 申し訳ないからホメルってワケでもない

例 2: 新言文一致体の例

> ん゛ あーーーーーーーーーーーーーー。
> 「もーーーーやだーーーーーーー!!」
> 映っててんよ☆★☆いやーーーーんモォ。
> ぴよぴよだけで反応してしまう〜〜〜〜〜〜
> どえええええっっ!?

例 3: Web 文体の削除例

(B) 雛型表現の削除

Web ページには独特の表現がある．例えば，HTML のフレーム機能が無いことを知らせる文面は「お使いのブラウザはフレームをサポートしていません」など (以下,「フレーム対応」表現) のようにほぼ固定されている．こういった特殊な雛型表現や単語が出現しているところから共起情報を得ようとする場合，雛型表現などはその精度を低下させる原因になりかねないと藤井らは指摘している [6]．

そこで本稿では，出現頻度の高い以下の雛型表現に対して削除を行った．削除は，各々の削除対象に含まれるキーワードやパターンを基に行った．

- 「フレーム対応」表現 (フレームなど 4 種)
- 都道府県名の連続 (<漢字>+県 など 4 種)
- 値段表現の連続 (<数字>+円 など 2 種)
- 日付表現の連続 (yyyy/mm/dd など 2 種)

(C) 顔文字や感情表現文字の削除

先述した新言文一致体とよく併用されるのが，顔文字や感情表現文字 ((笑) など) である．本稿では，文末における文字の 5-gram を収集し，顔文字となりやすい丸括弧を含むものを解析対象とした．そして，出現頻度の高かった上位 23 種の顔文字，52 種の感情表現文字を削除対象として処理を行った．

この結果顔文字を含んでいると判定された文は，その 1 文全てを削除した．これは，顔文字範囲の完全な判定が難しいためである．一方，感情表現文字を含んでいる文は，その該当する感情表現文字のみを削除した．

4.4 実装結果

以上の方法により Web コーパスを構築した．コーパスの一般性を検証するため，基準となる URL を変化させて 3 つのコーパスを構築した．それぞれのコーパスに対応する URL は以下の通りである．

Web コーパス A ：リンク集
`http://www.webring.ne.jp/`

Web コーパス B ：健康に関する情報サイト
`http://www.health-net.or.jp/`

Web コーパス C ：首相官邸
`http://www.kantei.go.jp/`

提案手法に伴うコーパスサイズの変化を表 1 に示す．ここで，処理前のコーパスサイズは，URL データベースを全て巡回した結果ではなく，途中でページの収集を打ち切った結果である．

表 1: 提案手法に伴うコーパスサイズの変化 [MB]

	Web A	Web B	Web C
処理前	3505	2441	1929
処理後	223	202	135

次に，外面的質および内面的質の質向上に関する手法でどの程度の文が削除対象になったかを表 2 に示す．

表 2: Web コーパス A における各処理の効果

削除の要素	削除できた文の数	割合 [%]
完全一致文	51691	19.7
同一ページ	14878	5.7
字面比	4937	1.9
感情表現文字	3003	1.1
フレーム対応表現	2582	0.9
新言文一致体	2214	0.8
顔文字	1736	0.7

5 評価実験

構築したコーパスの特性と有用性を確認するために実験を行った．実験の際，形態素解析器として茶筌<1>を利用した．本節では，Web コーパスとして表 1 の Web A について調査を行った結果を示す．

5.1 単語を用いた Web コーパスの特性評価

Web コーパスにおいて，異なり単語数を調査した．その結果を表 3 に示す (表 3 の Web A_0 は，5.3 節で述べる)．Web コーパスは，新聞と同等のサイズ (21MB) に調整した．その際，無作為に行を選んでいる．

表 3 において，名詞以外は Web コーパスが新聞の単語数を超えていて，有利であることが判る．しかし，Web の未知語数がおよそ倍になっている．この未知語の中には表記揺れと見受けられるものが観察された．これを鑑みると実質的には名詞についても Web コーパスの方が多いと考えられる．また，多様な表現の収集としても Web コーパスの方が優位である．

Web コーパスは新聞に比べ多くの単語を含んでいるが，茶筌の辞書に比べるとまだ少ない．そこで，コーパスのサイズと単語数の関係を調査し

表 3: 同一規模 (21MB) の Web と新聞の比較

	毎日新聞	Web A	Web A_0	茶筌辞書
名詞数	55120	52512	44559	208232
動詞数	5124	6530	4303	14367
形容詞数	646	809	549	1624
未知語数	14788	34391	53523	—
全単語数	77789	96833	104867	234034

た．その結果を図 2 に示す．収集した Web コーパスの全て (220MB) を用いたとき，これは茶筌辞書の全てを超えている (25 万語)．図 2 より，動詞数や形容詞数が茶筌辞書のおよそ 3 分の 2，名詞数が 2 分の 1 であることが分かる．これは Web 文書において良く使われる単語 (HTML や WWW など) が茶筌辞書に登録されていないため，未知語として判定されているからである．したがって，名詞数も動詞数や形容詞数と同様，3 分の 2 かそれ以上が包含されている可能性は十分にある．

次に，得られた単語に偏りがあるかを調査するため，茶筌の品詞毎に 5 単語を抽出し，新聞と Web とで比較を行った．この結果を表 4 に示す．新聞コーパスにおいて，名詞 (サ変接続) では，発表，代表，選挙，会議といった政治，事件を容易に想像させる単語が得られた．一方 Web コーパスでは，利用，使用といった一般的な語が得られている．また，新聞コーパスにおいて，名詞 (一般) には，東京，大阪といった都市名が並ぶのも特徴的である．その他の品詞の名詞 (非自立) や動詞，形容詞においては，Web コーパス，新聞コーパスともに上位に現れる単語にそれほど変化は見られないことから，偏りはないと言える．

更に，構築した Web コーパスが，どのような単語を含んでいるかシソーラス (角川類語新辞典) <3>を用いて調査した．この調査結果を図 3 に示す．図 3 は多義性を考慮していない計量のため 100%を超えることがある．新聞コーパスと同規模の Web コーパスでは，シソーラス上の全単語に対する被

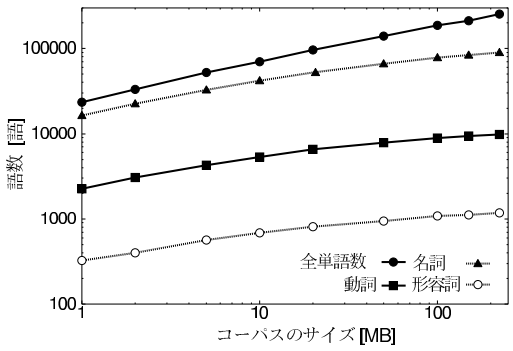

図 2: コーパスサイズと単語数の関係

表 4: 茶筌の品詞別に見た高出現単語 (上位 1〜5 位)

	Web A	毎日新聞
名詞 (サ変接続)	利用 規定 研究 使用 サービス	発表 代表 選挙 会議 調査
名詞 (一般)	こと の 者 もの 情報	こと 者 東京 の 大阪
名詞 (非自立)	的 必要 可能 重要 自然	的 明らか 可能 特別 必要
動詞 (但し，する，いる， なる，あるを除く)	できる くださる いう 思う 行う	行う 開く いう くる 受ける
形容詞	ない 強い 多い 高い 良い	ない いい 高い 多い いい

覆率が 47.6%となっている．一方，新聞コーパス一年分では 41.5%でその差は 6% (約 3 千 5 百語) であった．また，Web コーパス 220MB で同様に調査したところ，その被覆率は 62.1%となり高い被覆率を得た．図 3 において，大分類毎の差異はほとんど見られない．また，Web コーパス A(220MB)，Web コーパス A(21MB)，新聞コーパス (21MB) の順位も崩れず，同容量において Web コーパスが新聞コーパスよりも効率良く単語の収集ができている．図 3 と表 4 より，Web コーパスでは単語の分布，頻度共に偏りが見られない．一方，新聞コーパスでは語の分布の様子は Web と同様であるが特定分野の語の出現傾向が高いため，頻度分布に偏りがある可能性がある．

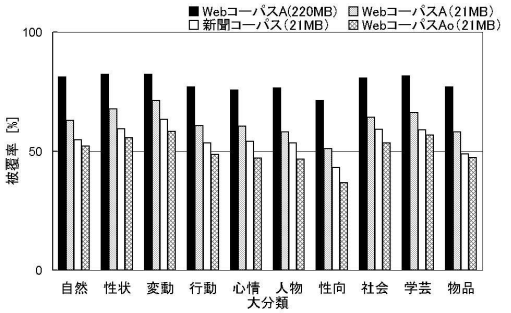

図 3: 角川類語新辞典の大分類による単語分布

表 5: 格フレームの異なり数とその例

	新聞	Web A
格フレーム数	145337	170274
頻度 上位 10 件	明らかにする	ことができる
	ことになる	ことになる
	ことを決める	ものとする
	容疑で逮捕する	ことがある
	明らかになる	必要がある
	罪に問う	ようになる
	方針を固める	ようにする
	会を開く	日から施行する
	性がある	場合がある
	ことが分かる	目的とする

5.2 Web コーパスの有用性

構築した Web コーパスの有用性を確認する．表現の豊富さを計量するためには動詞の使われ方を観察すれば良いと考え，格フレームを調査した．河原らは，「用言とその直前の格要素の組を単位として考えると，用言の用法はほとんど一意に決定される」と述べている [5]．本稿でもその性質より格フレームを簡便に構築し，それを用いて表現の豊富さを評価する．

格フレームの異なり数

表 5 に，Web コーパスと新聞コーパスより構築した格フレームの異なり数およびその上位 10 例を示す．同一サイズでは，Web コーパスの方がおよそ 2 万 5 千件多く収集できた．格フレームの用例としては，新聞コーパスで事件の展開などを示す「明らかになる」や「容疑で逮捕する」という用例が上位に現れるのに対し，Web コーパスではそのような特定の分野の用例は見受けられず，より一般的と言える．

頻度上位単語の格フレーム

コーパスに頻出する単語 (動詞) と格フレームの用例数にどのような関係があるかを調査した．Web コーパスおよび新聞コーパスにおいて，主な高頻度単語，5 単語を対象に収集できた格フレームの数とその例 (頻度 10 以上) を表 6 に示す．表中において 10 以上の頻度が無い場合は，‘—’ で示し，格フレームが存在しない場合は，‘(0)’ と示した．動詞「する」においては，名詞 (サ変接続)+する の形もあるが，ここでは格助詞+するの形のみを収集した．表 6 より，両方のコーパスで頻度が高い単語であっても新聞コーパスには存在しない格フレームがある．しかし，それらが Web コーパスで存在することも確認できた．

頻度 1 の格フレーム

一方で，多様な格フレームがどの程度収集できたかを確認するため，それぞれのコーパスにおいて頻度が 1 の格フレームを調査した．また，その格フレームが他方でどの程度の頻度を持っているのかを調べた．新聞コーパスに含まれる頻度 1 の格フレームは 108135 件，Web コーパスは 137926 件存在した．その中から無作為に抽出した 5 例を以下に示す．ここで，[] の中の数字はその例が他方のコーパスで何件の用例があるかを示す．

新聞コーパスの例
- 事件にぶつかる [0]
- 妻と会う [0]
- 人柄を想像する [1]
- 公開となる [2]
- 面倒を見る [9]

Web コーパスの例
- 指標を含む [0]
- ネジを外す [0]
- 十年間を見る [0]
- 制作に専念する [0]
- 建物ごとに行う [0]

特定分野の格フレーム

新聞コーパスおよび Web コーパスにおいて，どちらとも直接関係の無い話題「料理」で格フレームの有無を比較した．ここで用いる格フレームは，料理雑誌<4>の作り方から抽出したものである．

雑誌より得られた 18 件の格フレームのうち，コーパスに含まれていた格フレームのみを表 7 に示した．Web コーパスでは，18 件のうちの 7 件の格フレームを確認できた．一方，新聞コーパスではいずれも検出できなかった．上記に示した格フレーム以外で，Web コーパス，新聞コーパス共に検出できなかったものには以下のようなものがある (数字は料理の手順を示している)．

(3) をのせる，コンテナに残す，皿をとる，むらが防げる，ごま塩をふる，ふたをずらす，皿に取り分ける，マッシュを作る，電子レンジで加熱する，(2) を加える

Web コーパスにおいては，表現を少し変化させると検出できるものもあった (表現を変えて検出できたものを△で示した)．例えば，「ラップをかける」という格フレームは無かったものの，同意の「ラップをする」という表現は存在した．また，料理に関係する名詞から格フレームを作成し，Web コーパスから得られた格フレームの例を以下に示す．

醤油	を	使う，かける，加える ...
	に	漬け込む
塩	を	振る，まぶす，すり込む...
	に	漬け込む
フライパン	で	焼く，炒める

5.3 提案手法の有効性

本研究で構築した Web コーパスの特性が Web そのものの特性か，あるいは我々の提案手法による貢献があったのかを検討する．

Web コーパス A と同等の情報源から，単に HTML タグを除いて同一サイズのコーパスを構築した (Web A_0)．このコーパス A_0 の異なり単

表 6: 主な高頻度動詞

	Web A					新聞				
	ガ格	ト格	ニ格	ヲ格	総数	ガ格	ト格	ニ格	ヲ格	総数
する	気，感じ	もの，目的	よう，こと	仕事，話	4198	気，音	中心，柱	明らか，こと	行為，けが	2611
いる	人，者	にし，はし	そば，近所	ー	480	人，君	ー	近く，一緒	ー	390
なる	ー	必要，こと	こと，よう	ー	4212	ー	人，見通し	こと，明らか	(0)	3449
ある	こと，必要	ー	傾向，中	ー	3341	性，こと	ー	背景，市	ー	2140
行う	会，等	ー	日，前	等，活動	1426	試合，式	ー	日，一 斉	試合，戦	998

表 7: 料理の格フレームの実例

	皿に盛る	ラップをかける	などをのせる	室温に戻す	薄切りにする	ふたをする	水をきる
Web A	○ (4)	△ (6)	○ (1)	○ (3)	○ (4)	○ (4)	○ (2)
Web A_0	×	×	○ (2)	×	×	×	×
毎日新聞	×	×	×	×	×	×	×

語数および単語の被覆率調査の結果を表 3 に示す．表中の名詞数，動詞数，形容詞数は提案手法を適用することで増加傾向にある．また，Web コーパス A_0 に対して Web コーパス A は，全単語数が 8 千語しか減少していないのに対して，未知語は 2 万語減少している．これは提案手法が不必要な単語を削除し，良質な単語の収集に成功していると言える．また，図 3 より提案手法の処理を行った後の方が被覆率が上昇していることが確認できる．さらに，表 7 に示すように，格フレームの増加も観察できた．これらより，Web ページをそのままコーパスとして用いることは，Web が持つ言語情報を 100%生かしていないと言える．したがって，Web コーパスを構築する上で，提案手法の重要性が示された．

6 今後の課題

実装した手法の他に，対処しなければならない問題がある．(1) 現在，HTML より本文を抽出する際には文の結束性を無視している．しかし，これを無視すると共起情報に影響を与える恐れがあるため，処理を追加する必要がある．(2) 多くの分野に渡ってページを収集するためには，収集したページの URL やタイトルからページをクラスタリングする必要がある．(3) 改行箇所同定をより正確なものにするため，記号による箇条書きを認識することが必要である．これは，タグを使用せずに箇条書きをしている場合の認知である．以上の問題を考慮することで，さらに良質なコーパスを構築できると考えている．

7 おわりに

インターネット上の Web ページを用いた良質な Web コーパスの構築手法を提案した．新聞コーパス一年分と同一規模にして比較をした場合，異なり単語数 9 万 7 千語 (20%増)，異なり格フレーム数 17 万件 (15%増) を含む Web コーパスを構築できた．また，提案した質改善の手法により，Web ページをそのままコーパスとして構築するよりも良質の Web コーパスが構築できた．

使用した言語資源とツール

<1> 形態素解析器「茶筌」Ver.2.3.0． 奈良先端科学技術大学院大学 松本研究室．
http://chasen.aist-nara.ac.jp/

<2> 毎日新聞 全文記事データベース 2000 年版． 毎日新聞社．

<3> 大野晋，浜西正人． 角川類語新辞典． 角川書店．1981．

<4> レタスクラブ． 2003 年 8 月 25 日号, SS コミュニケーションズ．

参考文献

[1] Rosie Jones and Rayid Ghani. Automatically building a corpus for minority language from the Web. In ACL Proceedings of the Student Research Workshop 2000, pp. 29–36, 2000.

[2] Constantin Orăsan and Ramesh Krishnamurthy. An open architecture for the construction and administration of corpora. In Proceedings of LREC 2000: Second International Conference on Language Resources and Evaluation, Vol.2, pp. 793–799, 2000.

[3] 佐竹秀雄. 新言文一致体の計量的分析. 武庫川女子大学言語文化研究所年報, Vol. 3, pp. 1–14, 武庫川女子大学, 1991.

[4] 後藤斉. コーパスとしての新聞記事テキストデータ—終助詞「かしら」をめぐって—. 東北大学言語学論集, Vol. 5, pp. 37–46, 東北大学, 1996.

[5] 河原大輔, 黒橋禎夫. 用言と直前の格要素の組を単位とする格フレームの自動構築. 自然言語処理, Vol. 9, No. 1, pp. 1–16, 2002.

[6] 藤井洋一, 鈴木克志, 今村誠, 高山泰博. 共起情報を利用した文書の自動分類. 情報処理学会研究報告, 97-NL-118(16), pp. 97–104, 1997.

[7] 松本裕治, 徳永健伸. 言語コーパスの動向と将来展望. 「言語データ共有計画」シンポジウム (LRSI シンポジウム) 資料集, pp. 1–16, 1994.