ライティング教材作成を目指した日本語学術文長単位解析の試行

堀 一成 †, 坂尻 彰宏 †, 石島 悌 ‡,
大阪大学 全学教育推進機構 †, 大阪府立産業技術総合研究所 製品信頼性科 ‡
{hori, sakajiri}@.celas.osaka-u.ac.jp, ishijima@tri-osaka.jp

1 はじめに

大学学部初年次生向け日本語アカデミック・ライティング指導教材を作成する際の基礎データとするため、学術文・技術文の長単位による形態素解析を行い、用いられている一般動詞と普通名詞の頻度情報を得た。学術文の代表として、大阪大学に提出された博士学位論文の要旨文を、技術文の代表として、大阪府立産業技術総合研究所の技術報告概要文を、解析の対象として選定した。いずれも規模は小さく、試行を始めた段階に過ぎないが、長単位による形態素解析をすることで、学術文・技術文の表現特徴をよりよくマイニングすることができるであろうとの予測に至ったので、本稿で報告するものである。

2 研究の背景

大学学部初年次生を対象とした日本語アカデミック・ライティング指導教材は多数発行されている。筆者らも独自の教材を作成し、学部初年次生全員に配布するとともに、自由に利用できるようデータ公開している [5],[4]。しかし、その指導書などでは、文を書く際に使用する用語や言い回しの事例紹介がなされる例が多いが、その用例・文例を採用した根拠が明示されていることはまれである。一般に日本語教材の説明は著者の内省によっており、その根拠となる情報を示されることは少ない。（二通　他著「レポート・論文表現ハンドブック」:東京大学出版会 [7] は少ない例外であるが、これも心理・経済分野の特定著者の論文が論拠である。）

2013 年に、我々は、国立国語研究所開発の現代日本語書き言葉均衡コーパス (BCCWJ) を対象とした解析を行い、一部をライティング指導時の教材として提供する試みを行った [2],[3]。これは、特定の著者や学会に偏らないデータが得られ、その 成果をライティング指導に活用することで、より広範囲に応用できるライティング技能を受講者に身につけさせることができると考えたからである。BCCWJ の高校教科書とラベル付けされたデータを対象とし、長単位で区切りされた一般動詞・普通名詞の頻度情報を抽出した。国立国語研究所の発表した基本語彙調査の結果を参考に、一般文でも利用される頻度が高いと判断される語を、頻度上位のリストから除き、日本語アカデミック・ライティングで用いることを推奨する用語集として受講者に提供した。また作業を MySQL 上での SQL プログ ラムで行い、一部の自動化を実現した。実際のセミナー授業での活用の様子なども併せて報告した。

しかし、この時点での用語集の内容は、高校の教科書データが基盤であることが反映されており、大学で求められるアカデミックな語彙を十分提示できていない難点があった。そこで、一般の日本語学術文を対象として特徴語抽出を行えばより有用な情報が得られると予想されるが、BCCWJ のように形態素解析済み長単位情報が付与されている学術文データは少なく、平文テキストを効率よく処理する適切な手法がない状態であった。

2014 年に平文テキストデータを入力とし、長単位解析可能な形態素解析ソフトウェア Comainu と、関連した形態素辞書である Unidic2 が公開され、長単位基準の多様な言語資源に対する特徴語抽出が可能になった。

3 長単位解析ソフトウェア Comainu

小澤らが開発している形態素解析ソフトウェア Comainu[6] は、その解析単位に長単位を選ぶことができることが、大きな特徴である。冨士池ら [1] によると、長単位とは、文節の内部を自立語部分と付属語部分に分解することで認定される区切りである。長単位は資料の特徴語を取り出せることが利点であるとしている。対して、短単位は基準がわかりやすく、ゆれが少ないが、合成語を構成要素に分割してしまう問題点が

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

あるとしている。

これまで広く使われてきた JUMAN, ChaSen, MeCab といった日本語形態素解析ソフトウェアの出力は、基本的に短単位に分解された情報であるといえ、平文テキストを長単位で解析するソフトウェアの不足から、長単位の応用事例の報告が少なかったといえる。

4 頻度リストの作成方法

本稿は、日本語アカデミック・ライティング教材として日本語学術文データを長単位解析することの有効性を見極めるため、試行として小規模の処理を行った事例を報告するものである。

以下に長単位の動詞・名詞の頻度リストを作成した手順を説明する。作業は CPU Intel Corei5 560M/2.66GHz, メモリ 6GB 搭載のノート PC 上で行った。OS は Ubuntu 14.04 LTS である。

1. **解析対象平文データの準備**

   学術文として、大阪大学リポジトリ OUKA 上で公表されている博士学位論文概要を対象とすることとした。分野は、言語学・医学・法学・生物学・教育学のものを一件ずつえらび、概要文の箇所のみを UTF-8 テキストデータとして集約した。総字数は約 6,000 字である。技術文として、大阪府立産業技術総合研究所の技術報告および技術論文概要 (Web で公開されている。

   参照 URL http://tri-osaka.jp/c/syoho/) の平成 24 年度～平成 27 年度分のデータを UTF-8 テキストデータとして集約した。総字数は約 15,000 字である。

2. **Comainu を用いた解析作業**

   用意した平文テキストデータを Comainu Ver.0.71 で、平文から長単位の解析結果が得られるようオプション設定して処理する。実行時のコマンド設定例を付録 1 に示す。

3. **シェルスクリプトによる頻度情報抽出作業**

   得られたタブ区切り短単位・長単位解析結果から頻度情報を得る作業を行う。シェルスクリプト例を付録 2 に示す。

このようにして得られた学術文と技術文に含まれる一般動詞と普通名詞の頻度を、長単位と短単位の解析結果にわけ対照表とした。表 1 と表 2 は、いずれも頻度上位 50 位までの動詞データである。

5 得られた頻度リストに対する考察

得られた頻度リストのうち、一般動詞データに対して考察する。

普通名詞頻度データは、この規模の解析では各文の専門分野に関する用語が抽出できているだけで、汎用の学術文作成の参考になる情報をまだ含んでいないと判断したため、今回は提示していない。

長単位と短単位の処理結果の違いをみると、印象的説明になるが、2 例ともに、長単位がより学術的な硬い表現がマイニングできているといえる。とくに「抽象名詞＋する」動詞がきちんと拾えている。このことから、普段このような表現を書きなれない学部初年次生が日本語アカデミック・ライティング時に使用する語彙を検討する場合、選定の基準に関する有用な示唆を、長単位解析結果が与え得ることを表していると解釈できる。

また、これも印象的記述にすぎないが、博士学位論文データと大阪府産技研データの長単位リストを比較すると、大阪府産技研データの方に、よりやさしい（義務教育中で習得するような）表現が多いようである。これは、大阪府産技研の研究者は一般企業の技術者を、博士学位論文執筆者は大学研究指導者を、読者対象として想定していることの違いが表れているのではないかと考えている。

6 今後の展開

本報告は、学術文の長単位解析データを有効活用するための手法確認という位置づけである。今回の手順をきっかけに、さらに大規模・有用な結果がえられる手法開発へと進みたいと考えている。

◎ 解析対象コーパスデータの大規模化

近年多数の大学が整備を進めているリポジトリに掲載されている論文データを広く対象にする。今回対象としたデータは論文概要数例と小規模であったが、今後本文データや大阪大学以外のリポジトリも対象とすることで大規模データ化を図る。国立情報学研究所の CiNii 論文情報なども対象にすべきと考えている。技術文解析対象範囲についても同様である。

◎ 特徴的な語・表現の抽出方法の改良

今回、特徴語の抽出方法は、長単位解析して頻度情報を得るだけという、簡易な手法であった。今後適切なデータ集団の差異抽出手法を検討し、より良い抽出結果を得たいと考えている。単なる語彙情報のみに注目

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

するのではなく、連語 (コロケーション) の情報の提供がより有用であろうと予想している。

◎ 資料インストラクション手法の改善

受講生に資料の有効活用法を説明する方法も改善が必要である。作成した資料のみを渡すだけでは、有効活用は期待できない。頻度リストや関連 Web ツールを使用して、より良い文を選定する具体的な方法を、文章作成指導手順に組み込み提示したいと考えている。

既存の教材を、研究成果を含めたものに改善していきたい。

7 おわりに

以上のように、大学学部初年次生向け日本語アカデミック・ライティング指導教材を作成する際の基礎データとするため、学術文・技術文の長単位による形態素解析を行い、用いられている一般動詞と普通名詞の頻度情報を得た。学術文の代表として、大阪大学に提出された博士学位論文の要旨文と、技術文の代表として、大阪府立産業技術総合研究所の技術報告概要文を、解析の対象として選定した。現段階は試行状態であるが、長単位による形態素解析をすることで、学術文・技術文の表現特徴をよりよくマイニングすることができるであろうとの予測に至った。

謝 辞

本研究は、学術研究助成基金助成金挑戦的萌芽研究課題番号: 25540163 「XML コーパスからの抽出データに基づく日本語学術ライティング教材作成法の研究」（研究代表者：堀一成）による補助を受け推進したものである。

本研究は、小澤俊介氏を代表とする Comainu 開発グループの成果に依存したものである。有用なソフトウェアの開発と公開に対して深く謝意を表したい。

参考文献

[1] 冨士池優美, 小椋秀樹, 小木曽智信, 小磯花絵, 内元清貴, 相馬さつき, 中村壮範. 「現代日本語書き言葉均衡コーパス」の長単位認定基準について. 言語処理学会 第 14 回年次大会発表論文集, pp. 931–934, 2008.

[2] 堀一成, 坂尻彰宏, 石島悌. BCCWJ 教科書データより抽出した頻度情報に基づく日本語ライティング指導教材の作成. 第 4 回コーパス日本語学ワークショップ　予稿集, pp. 45–52, 2013.

[3] 堀一成, 坂尻彰宏, 石島悌. 現代日本語書き言葉均衡コーパスより抽出した頻度情報に基づく日本語学術ライティング指導教材の作成. 電子情報通信学会　技術研究報告　第 3 回テキストマイニングシンポジウム IEICE Technical Report NLC2013-14, pp. 1–6, 2013.

[4] 堀一成, 坂尻彰宏. 大阪大学におけるアカデミック・ライティング教育の実践と教材作成. 大阪大学高等教育研究 Vol.3, pp. 27–32, 2015.

[5] 堀一成, 坂尻彰宏. 阪大生のための アカデミック・ライティング入門第 2 版. 大阪大学 全学教育推進機構, 2015. http://hdl.handle.net/11094/27153 から自由に PDF ファイルをダウンロードできる.

[6] 小澤俊介, 内元清貴, 伝康晴. BCCWJ に基づく長単位解析ツール Comainu. 言語処理学会 第 20 回年次大会発表論文集, pp. 582–351, 2014.

[7] 二通信子, 大島弥生, 佐藤勢紀子, 因京子, 山本富美子. 留学生と日本人学生のためのレポート・論文表現ハンドブック. 東京大学出版会, 2009.

付録 1　Comainu 実行コマンド例

DThesis.txt が解析対象平文ファイルとする
実際は 1 行で書いて実行する

```
$ ./comainu.pl plain2longout
--input DThesis.txt --output-dir out
```

付録 2　頻度情報計算シェルスクリプト例

```
#! /bin/sh
# programmed by Dai Ishijima 2016
cat DThesis.txt.lout | nkf -e |
perl -e '
while (<>) {
@x = split(/\t/);
if ($x[8] =~ /^動詞-/) { # 長単位品詞
# 長単位語彙素読み出力
printf "%s\t%s\n", $x[12], $x[8];
}
}
' | sort | uniq -c | sort -nr
```


Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 1　大阪大学博士論文概要 5 例 (計約 6000 字) から抽出した動詞頻度表 (頻度上位 50 語まで)(堀一成 作成)

長単位動詞	長単位頻度	短単位動詞	短単位頻度
用いる	12	為る	140
行う	11	於く	31
有る	9	有る	31
する	8	居る	28
成る	7	因る	18
存在する	7	言う	14
割る	7	用いる	12
見る	7	つく	11
考える	7	行う	11
示す	6	成る	8
異なる	6	見る	7
活性化する	6	割る	7
燐酸化する	6	考える	7
着目する	5	異なる	6
調べる	5	示す	6
発現する	4	調べる	5
減少する	4	出来る	3
議論する	4	描く	3
示唆する	4	呼ぶ	3
呼ぶ	3	行く	3
踏まえる	3	踏まえる	3
構成する	3	経る	2
描く	3	対する	2
不活性化する	3	応ずる	2
終息する	3	働く	2
確認する	2	置く	2
提示する	2	介する	2
介する	2	受ける	2
応ずる	2	試みる	2
働く	2	得る	2
基づく	2	持つ	2
置く	2	基づく	2
観察する	2	生ずる	2
試みる	2	引く	1
発生する	2	比べる	1
持つ	2	向ける	1
機能する	2	結び付ける	1
寄与する	2	分かる	1
設定する	2	引き寄せる	1
考察する	2	当たる	1
象徴する	2	縮める	1
経る	2	好む	1
期待する	2	育つ	1
貢献する	2	付き合う	1
採用する	2	招く	1
受ける	2	歩く	1
得る	1	起こる	1
知る	1	来る	1
補記する	1	経つ	1
育つ	1	当てる	1

表 2　大阪府産技研技術報告概要文集 (計約 15,000 字) から抽出した動詞頻度表 (頻度上位 50 語まで) (堀一成 作成)

長単位動詞	長単位頻度	短単位動詞	短単位頻度
用いる	35	為る	349
行う	35	有る	81
する	21	居る	67
有る	20	因る	60
得る	19	つく	47
成る	16	行う	35
使用する	12	用いる	35
報告する	11	出来る	34
検討する	10	於く	30
分かる	10	得る	19
示す	10	成る	18
利用する	8	対する	12
有する	8	分かる	11
含む	8	示す	10
解説する	7	来る	10
形成する	7	含む	8
発生する	7	考える	8
試みる	6	有する	8
測定する	6	関する	7
述べる	6	述べる	6
注目する	6	言う	6
考える	6	試みる	6
紹介する	6	図る	6
優れる	5	優れる	5
作製する	5	比べる	5
期待する	5	従う	5
比べる	5	基づく	4
提案する	4	目指す	4
基づく	4	及ぼす	4
掛かる	4	見出す	4
及ぼす	4	引っ張る	4
図る	4	加える	4
加える	4	渡る	4
適用する	4	行く	4
評価する	4	求める	4
活用する	4	掛かる	4
求める	4	関わる	4
見出だす	4	生ずる	4
目指す	4	伴う	3
比較する	3	知る	3
発揮する	3	異なる	3
導入する	3	呼ぶ	3
決定する	3	高まる	3
伴う	3	抜ける	3
開発する	3	際する	2
放散する	3	施す	2
推測する	3	至る	2
呼ぶ	3	持つ	2
生成する	3	望む	2
抜ける	3	取る	2

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.