

論文作成支援のための学生論文における不適切な表現の分析

尾崎 遼^{*1} 村田 真樹^{*2} 都藤 俊輔^{*2} 三浦 智^{*2} 徳久 雅人^{*2}

^{*1} 鳥取大学 工学部 知能情報工学科

^{*2} 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻

^{1, 2}{s092019,murata,s082034,s072052,tokuhisa}@ike.tottori-u.ac.jp

1 はじめに

学部4年生など、研究を始めたばかりの論文を書き慣れていない者は、誤字や脱字、説明不足な表現を含め、論文として不適切な表現を用いがちになる。しかし学生自身でその不適切な文や表現に気づくことは困難である。そこで不適切な表現を検出する支援技術があれば便利である。そのためには学生の不適切な表現の事例の収集と分析を行うことが要である。そこで本研究では、論文として不適切な表現の収集と分析を行う。

本研究では、論文を執筆した学生の指導担当教員により修正が行われた学生論文を修正後論文、教員による修正が行われる前の論文を修正前論文とし、この二つで差分を取り、得られた差分をもとに不適切な表現の分析を行う。ここで得られた差分の分類分けを行い、どのような誤り修正があるかも分析する。さらに、頻度に基づく分析も行い、学生に用いられやすい不適切な表現の傾向や偏りを分析する。

本研究の主要な点を以下に整理する。

- 差分抽出の技術により修正前と後の学生論文から多くの不適切な表現を抽出した。その不適切な表現を分類することで、どのような種類の誤りが学生論文に存在するかを明らかにした(2.4節の表1)。
- 差分に含まれる高頻度の2単語連続を分析することで、種々の特徴のある傾向を発見した。例えば、「ている」という表現が修正前の表現に多く存在していた。この表現は積極性に欠けた他人事のような印象を与える表現であり修正されたものと思われる。

2 差分の抽出と分類分け

ここでは、指導担当教員による修正が行われる前の学生論文と、修正が行われた後の学生論文で差分をとる。差分を得るための手法として mdiff コマンド [1] を用いた。さらに抽出によって得られた差分箇所の分類分けを行う。分類を行うことで、どのような不適切な表現があるかを知ることができる。

2.1 データ

指導担当教員による修正の行われた5人分の学生論文で差分の抽出を行い、そこからさらに適切に差分が抽出

されているかを確認するために、差分箇所を含む一文を出力して人手で確認を行った。なお本研究で用いる学生論文データは2011年度の言語処理学会年次大会論文である。

2.2 手順

分類分けまでの手順を以下に示す。

1. 修正前の学生論文と修正後の学生論文に対して mdiff コマンドを用いて差分を抽出する。
2. 1で得られた差分箇所を含む一文を原文から抽出する。(例1に抽出した例を掲載)
3. 差分箇所の前後の共通部分(例1の共通部分(前)(後))の文字数を求める。
4. 前後の共通部分の文字数の小さい方の値を調べ、この値の降順にソートする。(共通部分が短いものは断片的に一致しただけであり適切な差分でない場合が多い。そのため前後の共通部分がある程度の長さをもっているものを有用な差分と考える。)
5. ソートの上位のものから順に人手で分類分けを行う。

例1

差分部分 が(を)
 共通部分(前): どちら
 共通部分(後): 空白に入れるべきかを推定する
 共通部分(前)の文字数: 3文字
 共通部分(後)の文字数: 14文字
 修正前文: どちら が 空白に入れるべきかを推定する
 修正後文: どちら を 空白に入れるべきかを推定する

2.3 分類分けの結果

前節の手順に基づき、差分抽出を行った。抽出した差分の上位約650個を人手で確認し、分析に有用な差分を258個獲得した。この258個の差分の分類分けを行った。分類分けに利用する分類の設定には古本ら [2] の「誤りおよび不適切表現の分類」を参考した。

以下に設定した分類を示す。分類ごとに例文を掲載す

る。例文の見方はアンダーラインを引いている部分が修正前の表現であり、括弧の中の表現が修正が行われた後の表現である。は空表現である。

1. 表記の修正

(a) 表記の統一 (漢字・カナ・ひらがな)

例文：『余分な漢字表現を含む言い回しは、冗長で分かり(わかり)にくい』

解説：同一論文の中でひらがなの「わかる」を用いているので、ひらがなで統一している。

(b) 専門用語の統一

例文：『教師あり機械学習 手法で(に) は性能の優れたサポートベクトルマシンを利用する』

解説：「機械学習」か「機械学習手法」どちらかの表現で統一させている。

2. 語彙・表現の修正

(a) 冗長性 (文を短く書き換えても大きな意味変化をもたらさないような修正)

例文：『機械学習 を行った場合(では) あまりよい結果は得られなかった』

解説：大きな意味の変化を起こさずに、より短い文へと修正している。

例文：『要約前の文章から得られる情報を用いて文の順序推定を行う 手法(の) が主な手法である』

解説：同じ単語や文が二回以上用いられて冗長なため修正している。(例文の場合「手法」が二回用いられているため。)

(b) 情報補完・詳細化 (読者に意味が伝わりづらい表現に言葉を補いわかりやすくなるようにする)

例文：『また、_(副助詞「は」と格助詞「が」に関わる) データの分析を行うことにより、日本語学習者にとって有用な情報を獲得する』

解説：どのようなデータを分析したのかを明確にするために情報を補完した。

例文：『適合率では 優るものの(ベースラインより高かったが、) F 値ではベースラインより低かった』

解説：何に優るのかが書かれておらず、わかりにくい表現になっている。

(c) 大雑把・論文として安全な表現へ

例文：『対象語列の出現頻度 と 照合(を利用) して誤り表現の検出を行う』

解説：英語で use の意味をもつ「利用」という

語を用いて違和感のない表現に修正している。

例文：『素性を拡充することでさらに性能向上が期待できる(を目指したいと考えている)』

解説：論文として指摘を受けにくいような安全な表現へと修正している。

(d) 適切な単語・表現へ (適切な単語や論文の内容に沿った語への書き換えなど)

例文：『素性を拡充することで より良い精度(さらに性能) 向上を目指したいと考えている』

解説：向上という語に係る語として、良い精度向上は日本語としておかしいため修正している。

3. 文法による修正

(a) 助詞・接続詞の修正

例文：『どちら が(を) 空白に入れるべきかを推定する』

解説：「入れる」に係る語として「が」は不適切なため修正している。

(b) 係る語との対応

例文：『機械学習を用い_(た) 格助詞「が」、副助詞「は」の分類を初めて行った』

解説：「用い」の係り先として『行った』ではなく「分類」に係ってほしいので「用いた」に修正している。

(c) 時制の修正

例文：『ヒューリスティックルールに加え教師あり機械学習法を利用することで性能の改善が可能であることが わかる(わかった)』

解説：実験などを行った際の結果なので過去形で表している。

4. 文体の修正

(a) 口語の修正

例文：『結果を さらによく(改善) する方法として次の方法が考えられる』

解説：口語を論文らしい表現に修正している。

(b) 硬い表現の軟化

例文：『近年、パソコンやインターネットの普及により、計算機を使って文字を入力する機会が 増し(増え) ている』

解説：硬い印象を与える語を柔らかい表現へと修正している。

2.4 分類の結果と考察

人手で分類した不て節な表現の個数を集計した結果を表 1 に示す。

表 1: 分類の結果

修正項目	人物	A	B	C	D	E	合計
(表記の修正)							
用語の統一		1					1
その他の表記の統一			1		1	2	4
小計		1	1	0	1	2	5
(語彙・表現の修正)							
冗長性		4	5		3	20	32
情報補完・詳細化		9	13	7	13	34	76
大雑把・安全な表現へ		1	5		2		8
適切な表現へ		9	16	5	8	44	82
小計		23	39	12	26	98	198
(文法による修正)							
助詞・接続詞		6	6	3	3	20	38
係ることの対応				2			2
時制		2	1		1	3	7
小計		8	7	5	4	23	47
(文体の修正)							
口語		1			1	3	5
硬い表現の軟化					2	1	3
小計		1	0	0	3	4	8
合計		33	47	17	34	127	258

表 1 より、大きな項目の分類で見ると『語彙・表現の修正』が圧倒的に多く、次いで『文法による修正』が多いことがわかった。さらに細かい分類項目で見ると『適切な表現への修正』、『情報補完・詳細化の修正』、『助詞・接続詞の修正』の修正箇所が多くみられた。これらの原因として、学生は論文を書き慣れていないため、読み手に伝わりにくい内容の欠落した文章を書きがちであるということ、助詞や接続詞の誤用、知識不足で言葉をあまり知らないため誤った単語を用いるということが想定される。

3 頻度による分析

差分箇所の頻度を分析することによって不適切な表現の偏りや傾向を調べる。

3.1 差分の頻度分析

3.1.1 頻度の集計

3 章で抽出した差分の頻度の集計を行った。頻度 3 以上の差分を表 1 に示す。表中の は空表現を意味し、修正前の文に矢印の後の文を追加した箇所に相当する。例えば表 2 の番号 1 の修正前と後の表現は以下のとおりである。

修正前：要約前文章から得られる情報

修正後：要約前 の 文章から得られる情報

表 2: 頻度 3 以上の差分

番号	修正前	修正後	頻度
1		の	11
2	簡潔な	冗長でない	7
3		は	5
4	単語	自立語	5
5	前後	文の順序を	4
6	した	する	4
7	のっている	につく	3
8	もの	表現	3

3.1.2 考察

表 2 から、表現の挿入、特に助詞の挿入が多くなされていることがわかった。しかし抽出した差分のままでは頻度の集計を行っても、あまり特徴的な偏りや傾向は見られず、9 割以上が頻度 1 となる抽出結果であった。そのため次節では、抽出した差分箇所を単語単位に分解してから頻度分析を行う。

3.2 差分に含まれる 2 単語連続の頻度分析

抽出した差分をそのままの形で頻度を数えると、頻度 3 以上のものは 8 箇所しか見つからなかった。よって、2 節で抽出した差分の表現から、2 単語連続を取り出し頻度の集計を行い、不適切な表現や修正後表現の出現傾向を調べる。ここで取り出す単語を 2 単語連続としたのは、1 単語だと短く分析に用いるには情報不足であり、3 単語連続だと出現頻度が小さく頻度分析がしづらいためである。ここで取り出す 2 単語連続は 5 人中 2 人以上の論文に出現している表現に限定して行う。

3.2.1 手順

2 単語連続の頻度分析の手順を以下に示す。

1. 差分抽出によって得られた差分表現を利用する。
2. 1 の差分表現の修正前表現と修正後表現に対して形態素解析 (ChaSen[3]) を行い、それらを単語単位に分解する。
3. 形態素解析を行った結果から 2 単語連続を取り出し、その頻度を数える。
4. 修正前表現で得られた 2 単語連続の頻度を a とし、修正後表現で得られた 2 単語連続の頻度を b とする。 $a/(a+b)$ という式を利用して 0~1 までの数値で得る。
5. 上記の式で得られた値が 1 に近いものが、より修正後表現で利用される可能性が高いと考えられる。逆に 0 に近いものは修正前表現で利用される可能性が高いと考えられる。

3.2.2 結果

分析の結果、数値が0または1に近く、なおかつ特徴的な傾向が見受けられたものを表3、表4、表5に示す。

表3: 修正後に数が減った2単語連続

2単語	修正前頻度	修正後頻度	数値
ている	38	7	0.84
用いて	21	5	0.81
とし	16	5	0.76
を行う	10	3	0.77

表4: 修正後に数が増えた2単語連続

2単語	修正前頻度	修正後頻度	数値
ように	4	11	0.27
出現する	4	9	0.31
である	14	22	0.39

表5: 修正前か修正後にしか出現しなかった2単語連続

2単語	修正前頻度	修正後頻度	数値
され	16	0	1.00
を行っ	8	0	1.00
を利用	0	9	1.00
それを	0	8	1.00

3.2.3 考察

表3より「～ている」という表現が修正後には大幅に数を減らしていることがわかった。この原因としては、「～ている」という表現は執筆者が論文を書くにあたって積極性に欠けた他人事のような文を書いている印象を読者に与えてしまうため、修正がなされたと考えられる。実際に抽出した文からは「用いている」といった表現を「用いた」などに書き換えられている場合が多く見受けられた。2番目の「とし」が修正されているものとして、「結果としては」という表現が「結果は」のように書き換えられている場合が多く見受けられた。これは冗長性による書き換えに該当し、表1でも冗長性が多く修正されていることがわかるため、これはあまり論文に用いるにはふさわしくない表現である可能性がある。「～を行う」も同様に冗長性の問題から「～する」に修正されている場合が多く存在した。

表4から読み取れることとして「ように」、「出現する」、「である」の3つ全てが、文の情報を明確にするために書き足された表現ということである。実際に「ように」が用いられた例として「読みやすく修正した」という文を「読みやすくなるように修正した」というように書き換えが行われている。他に2例も同様に明確化のための書き換えで用いられていた。

表5の考察として、「され」という表現は「する」「した」「でき」という表現に修正されている場合が多く見受けられた。これは「～ている」の考察と同様に、受け身文ということで執筆者自身に積極性が欠ける表現な

めに修正が多くなされたと考えられる。「を行っ」は表3の「を行う」と同様に冗長性の問題から修正が行われていた。「を利用」という表現は、「使う」や「用いる」といった表現を書き換えている場合が多く見受けられた。これは論文修正者の好みの問題と考えられるが、論文として安全な表現への修正とも考えられる。「それを」の修正は、文として意味は通じるが、指示詞は読み手に不親切なため、文の情報の詳細化、明確化の点から修正がなされていると考えられる。

4 関連研究

古本ら[2]は工学を専門とする日本人学生が書いた文章に見られる基礎的問題点として、学生の書く文章に現れる誤用を分析している。本研究での分類分けでは古本らが分析で用いた不適切表現の分類を参考にした。

村田ら[4]は差分を用いた言い換えパターンを抽出する技術を利用して、英語運用における個人的な誤りパターンを抽出するシステムを作成した。英文校閲前のもとの英文校閲後のもので差分を取り、この差分を誤りパターンとして抽出し、頻度を計算し、結果を考察した。この研究は差分を用いた分析手段として参考にした。

阿辺川ら[5]は下訳から修正訳に至る過程でどのような要因で修正操作が施されるかの解明を試みている。

5 おわりに

本研究では差分を用いた学生論文の不適切な表現の分析を行った。その結果『語彙・表現の修正』『文法による修正』が多く行われていることがわかった。また、差分に含まれる高頻度の2単語連続を分析することで、特徴のある傾向を発見した。例えば、「ている」という表現が積極性に欠け良くない表現の場合があることを発見した。今後は本研究で得られた知見を活かして、学生論文の作成支援に役立つ技術の開発をしたいと考えている。

6 謝辞

本研究は科学研究費(23500178)の助成を受けたものである。

参考文献

- [1] 村田真樹. “di を用いた言語処理-便利な差分検出ツール mdi の利用-”, 自然言語処理(言語処理学会誌), 9巻, 2号, pp.91-110, 2002.
- [2] 古本裕子, 苗田敏美, 八重澤美知子, 川西琢也. “工学を専門とする日本人学生が書いた文章に見られる基礎的な問題点”, 専門日本語教育研究, 第7号, pp.47-52, 2005.
- [3] ChaSen <http://chasen.naist.jp/hiki/ChaSen/>
- [4] 村田真樹, 井佐原均. “自動言い換え技術を利用した三つの英語学習支援システム”, 情報科学技術レターズ, 3巻, pp.85-88, 2004.
- [5] 阿辺川武, 影浦峯. “下訳と修正訳を用いた訳文修正パターンの発見”, 言語処理学会年次大会発表論文集, 13巻, pp.919-922, 2007.