論文作成支援のための学生論文における誤り表現の分析

尾崎 遼 [*1]　村田 真樹 [*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}@ike.tottori-u.ac.jp

1　はじめに

文章を読む際に、語句や単語の説明が不足していると、文の意味を理解することが難しい．特に若手の研究者など, 論文を書き慣れていない者は、誤字や脱字を含め、論文として不適切な表現を用いがちになる．そこで本研究では差分を用いた手法に着目し、論文として不適切になりがちな表現の分析を行い、論文の文章作成支援、解決を目指す．まず学生論文の修正前、修正後の文章で差分をとり、どのように文章が修正されているかを分析する．分析の結果から学生が用いやすい誤り表現の傾向や偏りを調べ、パターンによる修正などが可能でないかを検討する．さらに余力があれば文章の自動修正法の検討も行う．

2　研究の進め方

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、得られた修正差分をもとに誤り表現の分析を行う．また、得られた修正差分の分類分けを行い、どのような誤り修正が多いのかを分析する．

3　差分の抽出と分析

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 [**?**] を用いた．

3.1　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、そこからさらに有意な差分であるかどうかを判断するために差分箇所を含む一文を出力し、人手で考察した．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

3.2　結果

差分抽出を行い、合計で約 650 個の修正差分を獲得した．さらに人手の考察の結果、有意と考えられる修正差分を 258 個獲得した．例 1 に抽出した差分の一つを載せる．

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
前 3 文字
後 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

4　分類分け

抽出によって得られた有意な差分箇所の分類を行う．分類を行うことで、用いられやすい誤り表現の傾向や偏りなどを知ることができると考えた．なお分類分けには古本ら [**?**] の「誤りおよび不適切表現の分類」を参考に行った．

4.1　分類の手順

1. 差分箇所を含む文を句点までで区切る．これを得られた差分箇所全てで行う．
2. 差分箇所の前後にある、共通部分 (例 1 の共通部分 (前)(後)) の文字数を測る．
3. 前後の共通部分の文字数の小さい方の値を調べ、この値が大きい順にソートする．(共通部分が短いものは適切な差分が取れていない、もしくは断片的に一致して差分が抽出されている場合が多かった．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が有意な差分と考えた．)
4. ソートの上位のものから順に人手で分類分けを行い表を作成した．

4.2　分類の例

各分類項目について、実際に抽出した差分箇所を例文として掲載する．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, 括弧の中に入れられている文が修正が行われた後の文となっている. Φは修正後の文を挿入する．

1. 表記の修正
   (a) 表記の統一 (漢字・カナ・ひらがな)
      - 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
      - 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タ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. 助詞の修正もあり

2. 語彙・表現の修正
   (a) 冗長性
      - 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

      - 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.（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
      - また，Φ(→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

      - 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
      - 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

      - 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)
      - 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：向上と言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
      - 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『を』は不適切なため修正している（係る語との対応にも含まれる）

   (b) 係る語との対応
      - 機械学習を用い Φ (→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
      - 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 (文体の修正は一部『語彙・表現の修正』の適切な表現にも含まれる)
   (a) 口語の修正
      - 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
      - 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

4.3 結果と考察

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
係るごとの対応			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 の分類結果から、大分類の項目で見ると『語彙・表現の修正』が圧倒的に多く、次いで『文法による修正』が多くみられた．さらに細かい分類項目で見てみると『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考えられることとして、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．これには若者が日常的に適切な日本語を使っていないということも原因となっているのではないかと考えられる．

5 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誤り表現の偏りがないかを調べる．

5.1 抽出した差分の原型での頻度分析

3 章で抽出した差分を、修正のパターンごとに頻度のカウントを行った．カウントの結果、頻度 2 以上のものを表にまとめた．表中の『Φ→』は挿入を意味し、修正前の文に矢印の後の文を追加した箇所である．例として表 2 の番号 1 を挙げて解説を以下に示す．

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

5.1.1 結果

表 2: 頻度 2 以上の修正差分

番号	修正前		修正後	頻度
1	Φ	→	の	11
2	簡潔な	→	冗長でない	7
3	Φ	→	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	のとっている	→	につく	3
8	もの	→	表現	3
9	Φ	→	法	2
10	している	→	する	2
11	Φ	→	で	2
12	付く	→	存在する	2
13	付く	→	出現する	2
14	や	→	と	2
15	に	→	で	2
16	Φ	→	すべての	2
17	Φ	→	本研究の	2

5.1.2 考察

結果の表から、文の挿入、特に助詞の挿入が多く出現していることが読み取れる．しかし抽出した差分を抜き出したままの形で頻度のカウントを行っても、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．そのため次項では、抽出した差分箇所を単語単位に分解を行ってから頻度分析を行った．

5.2 2 単語での頻度分析

抽出した差分箇所の文をそのままの形で頻度を数えると、頻度 2 以上の箇所は表 2 の 17 パターンしか見つからなかった．なので 3 章で抽出した有意な修正差分を、2 単語ごとに取り出し頻度のカウントを行い、単語の出現の偏りも調べた．ここで取り出す 2 単語は 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った．以下に 2 単語連続の例を載せる．

例：性能を向上させる

性能 を

を 向上

向上 さ

さ せる

というように 2 単語を獲得して頻度をカウントした.

5.2.1 手順

1. 差分抽出によって得られた有意な修正表現を集める.
2. 抽出した箇所に形態素解析 (ChaSen) を行い, 単語単位に分解する.
3. 形態素解析を行ったものを 2 単語連続で取り出していき, 頻度をカウントする.
4. 修正前差分で得られた 2 単語連続の各頻度を a とし, 修正後差分で得られた 2 単語連続の各頻度を b とする. そこに a/(a+b) という式を利用して偏りを求める. そして式の値の結果を 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の必要な表現である可能性が高いと考えられる.

5.2.2 結果

表 3: 修正後に数が減った 2 単語

2 単語	修正前頻度		修正後頻度	偏り
ている	38	→	7	0.844
が存在	9	→	2	0.819
用いて	21	→	5	0.808
とし	18	→	6	0.750
を行う	12	→	4	0.750

表 4: 修正後に数が増えた 2 単語

2 単語	修正前頻度		修正後頻度	数値
な文	4	→	12	0.750
ように	4	→	11	0.733
存在する	4	→	9	0.692
出現する	4	→	9	0.692
である	14	→	22	0.611

表 5: 修正前か修正後にしか出現しなかった偏り 1 の 2 単語

2 単語	修正前頻度		修正後頻度	数値
され	16	→	0	1.0
を行っ	8	→	0	1.0
よって	8	→	0	1.0
しない	7	→	0	1.0
を利用	0	→	9	1.0
基づく	0	→	8	1.0
それを	0	→	8	1.0
わかった	0	→	7	1.0

5.2.3 考察

6 関連研究

村田ら [?] 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. 英文校閲前のものと英文校閲後のものを UNIX の Diff コマンドで差分を取り, この差分を誤りパターンとする. その抽出された誤りパターンの頻度を計算し, 結果を考察する. この研究は差分を用いた分析の手順として参考にした.

古本ら [?] 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の表を参考にした.

7 おわりに

本研究では差分を用いた学生論文の誤り表現の分析を行った．その結果『語彙・表現の修正』『助詞・接続詞の修正』が多く行われていることがわかった．この修正箇所に焦点を絞りさらに分析を行い、修正のパターンを見つけることができれば学生論文の作成支援に大きな貢献をもたらすことができると考えられる.

参考文献

[1] 村田真樹. “diff を用いた言語処理-便利な差分検出ツール mdiff 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] 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 P.85-88,2004 8 月

[4] 村田真樹, 井佐原均.“尺度に基づく変形の利用” 自然言語処理 11 巻 5 号 P.113-133,2004 10 月

[5] 村田真樹, 金丸敏幸, 井佐原均.“複数の辞書の定義文の参照に基づく同義表現の自動獲得” 自然言語処理 11 巻 5 号 P.135-149,2004 10 月

[6] 阿辺川武, 影浦峡.“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 P.919-922,2007 3 月