## 形態素解析を利用した文章校正手法の提案

坂本 俊介<sup>†</sup> 須藤 崇志<sup>†</sup> 丸山 広<sup>†</sup> 中村 太一<sup>†</sup>

技術文章力を高めるには,個別に添削指導を受けることが有効であるが,時間的な制約により実現が難しい.この問題に対処するため,書き手自身が短時間で何度も文章の客観的な見直しを繰り返し行い,さらに指摘の重要性や修正例を提示する校正支援手法により,学生の文章作成能力の向上に寄与できることを示した.また,学生の実験レポート 2,285 編から,140,214 件の誤りを指摘した結果,学生に多い誤りは句読点および体言止めで 79.4%を占めた.同時に,技術文章作成のノウハウを課題やレポートを通じて習得させることの重要性を改めて確認できた.

# A Method for correcting technical writing by Japanese language morphological analysis

Shunsuke Sakamoto<sup>†</sup> and Takashi Sudou<sup>†</sup> and Hiroshi Maruyama<sup>†</sup> and Taichi Nakamura<sup>†</sup>

There has been discussion about improving technical writing skills. Individual tutoring which a mentor provides a trainee to correct a technical writing might be the most effective way to enhance the skills. However it is difficult to repeat the tutoring many times because of time consuming job. We propose a method that an author oneself can emend a manuscript repeatedly. The method is based on a recommendation system which implements Japanese grammar and technical writing grammar with a morphological analysis. This paper describes the experimental results obtained from 2,285 reports written by students. The proposing method detected 140,214 technical writing errors in their reports. 79.4% of 140,214 errors are a punctuation mark and the indication of the termination with a substantive and speak that it is effective to teach these errors.

#### 1. はじめに

研究者や技術者には論理的で分かりやすい文章を書く能力が求められる。文章作成能力を向上させるには、文章を繰り返し見直し、修正し、文章の修正方法のノウハウを積み重ねる必要がある。執筆者は自ら、何回も繰り返し見直すことができるが、客観的に自分が作成した文章を見直すことが難しい。第三者に見直しを依頼すると客観的な指摘が得られるが、見直しを繰り返し依頼することは困難である。また、作文技術を向上させるには、個別の添削指導が効果的であるが、添削指導する教育者の負担が大きい。

この問題に対処するため、本学では、文章表現規則に則り、客観的に文章を見直し、修正すべき部分を文章作成者に通知する推敲支援手法を提案し、実装し、学生の実験レポート作成支援に供している[1][2][3]. 具体的には、修飾語の並べ方の規則、読点の打ち方の規則、およびそれら二つの規則の競合回避規則を持つ推敲支援手法を開発した. 現在は、上記の二つの規則を検出アルゴリズム化し、一つの検出アルゴリズムにどの程度の機能を持たせるのが妥当か、検出アルゴリズム同士が競合しないかを分析し、個々の検出アルゴリズムをどの程度まで細分化するか、適切な検出アルゴリズムの優先順位や例外条件は何かを検証している。

文章を精査する作業は校正,推敲,および筋立ての三つで構成される.図1に三つの作業の位置付けを示す.縦軸は指摘の検出対象を,横軸は指摘箇所の検出方法を示す.四角に囲まれた文字は,指摘を行う規則である.

校正は、一つの文から単語の単位で文法に違反する箇所を修正する行為である。本研究では、形態素解析で抽出される単語の品詞や活用形を文法と照合し、誤りを検出することで、校正作業を支援する。推敲は、一つの文を対象に、読みやすさを向上させる。従来研究では、係り受け解析を用いて読みやすさを評価した。筋立ては、文全体の論理の展開と章立て構成を検証する。この検証は、文章の意味解析により行うが、正確に文章の意味を解析するには、解析対象の文章が少なくとも文法や技術文章の規則に則り、校正され、推敲されていることが必要である。文法が誤っていると、係り受け解析を正しく行うことができず、推敲を要する箇所を検出するアルゴリズムが誤った判断をしてしまい、効果的な推敲支援を行うことができない。

多くの学生は、大学のレポートで初めて技術文章を作成するので、作成された文章には文法に関する多くの誤りがある。そのため、推敲に先立ち校正の文章作成支援を提供することで、推敲の検出アルゴリズムが正しく判断できるようにする。また、指摘された内容を、文法や検出手法の専門知識を必要とせずに理解し、学生が校正できる工夫が必要である。

<sup>†</sup> 東京工科大学大学院

Tokyo University of Technology Graduate School

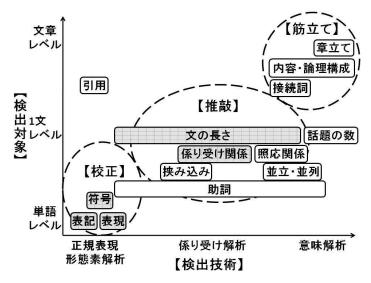


図1 文章改善方法の分類

## 2. 関連研究

## 2.1 10 種類の校正ルールと係り受け関係に基づいた木構造を用いた推敲支援

技術文章作成能力の向上を目的とした、ルールベースによる校正支援と、係り受け構造により執筆者に文意を確認させる推敲支援の取り組みがある[4].この校正支援のためのルールは以下の10個である.

- ① 句読点の誤り, 漢字含有率, 長すぎる文
- ② 不適切な表現
- ③ 全角の英数字と、半角のカタカナ
- ④ 文末の口語文章体
- ⑤ 当て字
- ⑥ 常用漢字以外の漢字の使用
- ⑦ 不適切な一人称
- ⑧ 形式名詞のひらがな表記化
- ⑨ 接続詞のひらがな表記化
- ① 副詞の表現

推敲支援のための文意の確認は、複数の解釈が成り立つ曖昧な文章を視覚的に発見するために、係り受け関係を木構造で表示することで実現している.

この方法は、利用者が文の構造を意識しながら、修正する工夫があるが、視覚的に表示された係り受け解析結果の意味を理解し、文章の訂正を行うには、形態素解析と係り受け解析を利用者が理解することが求められる。しかし、自然言語処理を専門に学んでいない学生が、木構造で表示された文構造や、その深さを基に提示される文章構造を把握することは難しい、文章作成の訓練に先立ち、自然言語処理を学生が習得することは現実的ではない。我々の提案手法は、利用者には指摘をした根拠と修正例と修正方法を提示する。利用者は自然言語の事前知識を必要とせず、指摘内容を読むだけで修正方法を理解できる。

#### 2.2 字面解析による活用語の活用形を推定する方法を利用した推敲支援

字面の並びを解析して修正すべき箇所を指摘することで推敲を支援する研究がある[5]. 用言を語幹と活用語尾に分け、それぞれで品詞を推定している. そして、語幹と活用語尾で推定された品詞が一致した場合に該当の品詞や活用形で用言が存在すると判断している. 語幹と活用語尾を分離するため、用言をそれぞれ2つの部分に分けて語幹表(漢字表と平仮名表)と活用語尾表に登録している. この活用形チェック法を推敲支援に適用すると、逆茂木型の文の検索と動詞の表記揺れの検出、並列構造の推定ができる.

この研究では、情報検索の分野で使用されている再現率と適合率を抽出精度の指標に使用する.再現率を式(1)で、適合率を式(2)で定義している.

再現率 = 
$$\frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{文章中の抽出すべき対象の数}} \times 100 (%)$$
 (1)

適合率 = 
$$\frac{\text{候補中に含まれる抽出すべき対象の数}}{\text{抽出法で得られる候補の数}} \times 100 (%)$$
 (2)

文章から不適切な表現の箇所を抽出する際の問題は、不都合な表現を指摘できないことと、指摘する必要がない表現まで検出してしまうことである。前者の誤りは、再現率を低下させ、後者の誤りは適合率を低下させる。この研究では、文章中に問題となりそうな箇所があればそれを漏れなく指摘することを基本方針とし、再現率を100%の状態において適合率を上げることを目的とする。この手法は、指摘された個所に問題があって修正するべきか、あるいは誤検出で修正する必要がないかを利用者が判断する。しかし、技術文章を初めて作成する学生は、指摘された箇所から修正すべき箇所と修正する必要がない箇所を判断することが困難である。そこで、教育に利用する場合は、必ず誤りである個所だけを検出する適合率重視で指摘する必要がある。

## 3. 提案手法

#### 3.1 文章一般のルールと技術文章のルールの違い

本研究は、技術文章を対象に技術文章の書き方の規則を適用して不都合な個所を検出する方法を確立することを目的とする。技術文章の書き方の規則と日本語の一般的文章の規則を同時に適用すると不都合なことが起きる場合がある。例えば、日本語の表現の一つの体言止めは、日本語の表現として誤りではないが、技術文章では用いないので、技術文章の規則を援用し修正を促す必要がある。換言すると、適用規則の衝突回避が重要な課題である。そこで、指摘すべき項目を文章一般のルールと技術文章のルールの二つに分類し、技術文章のルールに規定がある部分とない部分を分けて考えることにした。指摘すべきかチェックするときに、技術文章のルールに規定がない場合は、文章一般のルールを適用する。また、技術文章のルールに規定がある場合は、技術文章のルールを適用し、両方に規定がある場合は、技術文章のルールを優先することとした。文章一般のルールは、助詞の誤り・不適切な接続詞・形式名詞・呼応の副詞・話し言葉・ら抜き言葉の6種類 20 項目である。文章一般のルール6 種類を表1に、そこに含まれる項目の例を表 2 に示す。

表1 文章一般のルール

指摘 No.	指摘の種類	指摘理由(項目数)
1	助詞の誤り	助詞の用法が誤っている箇所を指摘(5)
2	不適切な接続詞	接続詞以外の語で文同士を接続している箇所(1)
3	形式名詞	形式名詞の用法が誤っている箇所を指摘(1)
4	呼応の副詞	呼応の副詞の決まった表現が対応していない箇所(1)
5	話し言葉	文章に用いる表現は、話し言葉より書き言葉が適切(8)
6	ら抜き言葉	ら抜き言葉は,動詞の活用を無視した表現で不適切(4)

表 2 文章一般のルールの指摘項目の例

	• • • • •	1 101
指摘 No.	指摘の種類	指摘項目の例
1	助詞の誤り	1文中に助詞「たり」を1回だけ使用
2	不適切な接続詞	接続詞の「なので」を,文頭で使用
3	形式名詞	1 文中に形式名詞「ため」を複数回使用
4	呼応の副詞	「全然」と「ない・ぬ・ん」がセットでない (「全然」 を「全く」に修正するようにも指摘)
5	話し言葉	「だけど」「けど」「でも」「みたいな」
6	ら抜き言葉	「見れる」「来れる」

技術文章のルールは、句読点・体言止め・文末表現・文の長さ・難解な表現・ぼかし表現・不適切な漢字変換・内容の不足の8種類85項目である。技術文章のルール8種類を表3に、そこに含まれる項目の例を表4に示す。なお、本稿で適用した指摘ルールは、理科系の作文技術という文献や、山之内総合研究所のテクニカルライティングセミナーを参考にして策定した[6][7]。

表 3 技術文章のルール

指摘 No.	指摘の種類	指摘理由(項目数)
7	句読点	句読点は「、」「。」ではなく,「,」「.」を使用(2)
8	体言止め	心情的表現を含む体言止めは技術文章には不適切(1)
9	文末表現	技術文章の文末に「です」「ます」は不適切(3)
10	文の長さ	長すぎる文は文意が分かりにくく,誤解を招く(2)
11	難解な表現	理解の妨げになるので、より平易な表現で記述(2)
12	ぼかし表現	正確さが要求される技術文章は数値の範囲を明確に(9)
13	不適切な漢字	一般的に用いない漢字は,文章の理解を阻害する(65)
13	変換	
14	内容の不足	で指し示す内容が明記されていないと理解を阻害(1)

表 4 技術文章のルールの指摘項目の例

指摘 No.	指摘の種類	指摘項目の例
7	句読点	[,] [,]
8	体言止め	(名詞)+(句点 or 全角ピリオド)
9	文末表現	「です」「ます」
10	文の長さ	「1 文が 200 字以上」「1 文が 120 字以上」
11	難解な表現	「不可欠」
12	ぼかし表現	「ほぼ」「ほど」「約」「程」「ぐらい」「など」「等」
13	不適切な漢字 変換	「乃至」「或いは」「即ち」「出来る」「殆ど」「予め」
14	内容の不足	(名詞) +「に関しては」

#### 3.2 校正による文章作成能力効果向上の研究

指摘した点の重要度と修正例を利用者に提示することで理解を容易にする効果があることが確認されている[8][9]. そこで、校正すべき個所とその重要度を付記し、具体的な修正方法を提示することにした. 指摘の重要度は、表 5 に示すように Error と Waning の二つである. Error は、文章一般のルールは日本語文法の違反、技術文章の

#### 情報処理学会研究報告 IPSJ SIG Technical Report

ルールは論文原稿の様式の違反であり、必ず修正すべき指摘である。Waning は、日本語文法や論文原稿の様式に反してはいないが、修正することで内容が明確で分かりやすくなる指摘である。なお、文の長さの指摘のみ、一文に含む文字の数で Error とWarning を設定した。

表 5 15 16 16 20 20 20 20 20 20 20 20 20 20 20 20 20					
指摘の分類	指摘の種類	指摘の重要度			
	助詞の誤り	Error			
	不適切な接続詞	Error			
文章一般のルール	形式名詞	Error			
文卓一叔のルール	呼応の副詞	Error			
	話し言葉	Warning			
	ら抜き言葉	Warning			
	句読点	Error			
	体言止め	Error			
	文末表現	Error			
	文の長さ(200 字以上)	Error			
技術文章のルール	文の長さ(120字以上)	Warning			
	難解な表現	Warning			
	ぼかし表現	Warning			
	不適切な漢字変換	Warning			
	内容の不足	Warning			

表 5 指摘の分類

## 4. 運用結果

文章作成能力向上に効果が期待できる指摘を把握するため、指摘した項目の数を調査した. 提案する手法を実装した校正支援システムを、本学の学部2年生に学生実験レポート作成のために使用してもらった. 2,285 編の実験レポート作成の利用ログから、140,214件の指摘箇所が検出された.

140,214 件のうち文章一般のルールに則った指摘箇所の数を表 6 に示す. ら抜き言葉の指摘が最も多く,逆に呼応の副詞はほとんど指摘されなかった. 技術文章のルールに則る指摘箇所の数を表 7 に示す. 句読点の指摘が最も多く,難解な表現がされている箇所が最も少なかった. 句読点の指摘は,技術文章をはじめとする横書き文章の句点は全角ピリオドで,読点は全角カンマで表わすという規則に反している誤りである. 項目ごとの指摘数の上位 10 件を表 8 に示す. 句読点の誤りと体言止め,より詳細を記述した方が良いとされる「など」の使用の 4 項目の指摘が 10,000 件を超え,技術文章のルールの指摘が多いことが分かる.

表 6 文章一般のルールに基づいた指摘箇所の数

指摘 No.	指摘の種類	指摘箇所の数
6	ら抜き言葉	1,778
3	形式名詞	775
1	助詞の誤り	435
5	話し言葉	180
2	不適切な接続詞	81
4	呼応の副詞	5

表 7 技術文章のルールに基づいた指摘箇所の数

	200 200 200 200 200 200 200 200 200 200				
指摘 No.	指摘の種類	指摘箇所の数			
7	句読点	100,179			
12	ぼかし表現	14,697			
8	体言止め	11,212			
10	文の長さ	3,612			
13	不適切な漢字変換	3,491			
9	文末表現	2,574			
14	内容の不足	848			
11	難解な表現	337			

表 8 指摘数の上位 10 項目

順位	指摘項目	指摘箇所の数
1	読点に「、」を使用	59,994
2	句点に「。」を使用	40,185
3	体言止め	11,212
4	ぼかし表現の「など」	11,183
5	文の長さが 120 字以上	3,488
6	ぼかし表現の「等」	2,914
7	文末が「です」	1,754
8	ら抜き言葉の「見れる」	1,669
9	内容の不足の「に関しては」	848
10	形式名詞の「ため」の複数使用	775

#### 情報処理学会研究報告 IPSJ SIG Technical Report

次に、文章一般のルールと技術文章のルールに則る指摘箇所の数を表9に示す.表9から技術文章のルールの指摘箇所の数が文章一般のルールの40倍であることが分かる.

<b>≠</b> : ∩	/ <b>人</b> 米石 ~ "	1 n +E		コレナシ	$\mathbf{H} *\mathcal{H} \sigma$	くしん 事六
表 9	分類ご	とり担	11間 固	別作	出数の	ノレレ甲文

指摘の分類	指摘箇所の数
文章一般のルール	3,264
技術文章のルール	136,950
合計	140,214

本提案手法が文章作成能力向上に与える効果を確かめるため、同一人物が文章を作成する度に、校正支援システムを使い、その時の指摘数の変化を調べた。研究室の学部4年生が毎週作成する週間報告書の作成時にシステムを利用してもらい、利用ログを分析した。同一人物が週間報告書を毎週作成するので、継続利用の効果の検証に適していると判断した。利用期間は2008年12月中で、計4回である。利用者は校正作業にあたり、複数回システムを利用するが、2回目以降の利用ログは、一度校正支援システムが指摘した箇所を修正した後の利用ログであるのでので、本提案手法を継続利用することで向上する文章作成能力の調査には適さない。そのため、各週間報告書の初回利用分のログを収集した。また、利用を依頼した人の2008年4月から11月までの8か月分の過去の週間報告書に提案手法を適用し指摘数を調べた。

校正支援システムを 9 か月分の週間報告書 546 編に適用し、週間報告書の単語数の違いを正規化するため、9 か月分 546 編の週間報告書を形態素解析し、形態素数に対する、校正支援システムが指摘した数の比を求め比較することとした。この比を指摘数の割合 Rc とし、式(3)で表す。

指摘数の割合 
$$Rc = \frac{\text{指摘数}}{\text{形態素数}}$$
 (3)

図3に、評価対象者14人の4月から12月までの指摘数の割合Rcの平均値を各提出日ごとに表す、縦軸は、指摘数の割合の平均値で、横軸は週間報告書の提出回である、4月の初めに提出する週間報告書を第1回とし、12月の利用依頼をした週報提出回は、第36回から第39回である。

#### 週間報告書利用期間 第36回~第39回

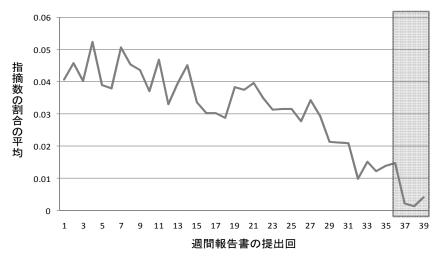


図3 週間報告書の提出回ごとの指摘数の割合の平均の推移

指摘数の割合の平均は、増減を繰り返し、全体的に右下がりに減少している.特に、第36回の週報で一度システムを利用した後の第37回から第39回は、指摘数の割合の平均が0.01を下回り、大きく減少していることが分かる.

## 5. 考察

表 8 の項目ごとの指摘箇所の数の 8 位と 10 位以外が全て技術文章のルールの指摘であり、表 9 の分類ごとの指摘箇所の数では技術文章のルールの指摘の数が文章一般のルールの 40 倍であることから、修正すべき箇所のほとんどが技術文章のルールによって検出された部分であることが分かる。大学教育における技術文章作成の指導の重要性が示された。今後、更に技術文章の規則を追加し、文章一般規則との衝突問題に対して整合を図る必要がある。

また、本システムを継続して利用すると、週間報告書での利用を依頼した 12 月の第1週分である第36回目以降、指摘数の割合Rcが大きく減少したことが確認できた. 第36回から第39回の指摘数の割合Rcと第36回との指摘数の割合Rcの比較は、表10に示すとおりである。本システム利用前と同様に一様に減少しているわけではない

が,第36回以前に1度も無かった0.01を下回る値が得られた.このことから,校正支援システムからの指摘を受けることで,問題のある表現を学習していることが分かる.以上より,提案手法が学習支援に寄与していることが確認できた.

表 10 週間報告書の第 36 回から第 39 回の指摘数の平均の値と第 36 回との比較

週報提出回	指摘箇所の数の平均	初回(第36回)との比較
第 36 回	0.014798	ĺ
第 37 回	0.002279	15.4%
第 38 回	0.001316	8.9%
第 39 回	0.004101	27.7%

また、システムの利用者にインタビューを行った結果、文章作成時には句読点を「、」や「。」で記述し、最後に Word やテキストエディタの置換機能を利用して一括変換することで体裁を合わせている利用者がいることが確認された。表 6 で示した項目ごとの指摘箇所検出数の 1 位と 2 位が読点と句点であることを合わせて考えると、同様の操作を行っている利用者が多いことが推測される。本稿は、文章作成の効率を論じているわけではないので、置換機能の是非については触れない。

しかし、1つ1つの文で意識して読点にカンマを、句点にピリオドを使っている学生に比べ、従来と同じように「、」「。」で文章を作成して、最後に一括置換する学生は、ルールの定着が薄いようである。新しい漢字を覚える際に、目で見て答えを覚えるよりも、手を使って何度も紙に練習した方が定着することは一般的に知られているが、同様のことがコンピュータ上の文章作成にもいえるようだ。技術文章の書き方を習得し、定着するまでは置換機能は利用しない方が学習効果は高いと思われる。

## 6. おわりに

本研究では、提案する手法を実装した校正支援システムにより書き手自身が短時間で何度も文章を客観的な見直しを繰り返し行い、さらに指摘の重要性や修正例を提示することで、学生の文章作成能力の向上に寄与できることを示した。

また、技術文章のルールに則った指摘が文章一般のルールより学生の文章作成支援に効果を上げているという分析結果から、レポート作成方法を論じた文献に記述されている文章作成のノウハウを、課題やレポートを通じて習得させることができる点の重要性を改めて認識できた。同時に、学生が指摘を受けた内容を分析することで、教員が学生の苦手なポイントを把握することができ、教育する側の支援にもなると考える。

本稿では、提案手法の評価は週間報告書により行ったが、実験のレポートを作成している学生に対しての評価も行う予定である。現在、学生が実験中に提出する課題に記された文章を用いて、校正支援システムを利用した学生と、利用していない学生の指摘数を比較することを考えている。また、今回得られた指摘項目ごとの指摘箇所の検出数から、指摘の分類ごとに設定した Error と Warning の二つの重要度の設定を、項目ごとに細分化して設定しなおすことを検討している。

今後は、利用者ごとに、過去に指摘した指摘項目とその項目を指摘した回数を記録して指摘項目ごとの累計指摘数をカウントするユーザプロファイルを考えている。現状では、誰が操作してもシステムに同じ文章を通せば、同じ指摘結果が表示される。それに対して、利用者が過去に入力した文章の指摘状況に応じて、その利用者が常にされている指摘を文章への指摘とは別に表示するようにすれば、個々の癖に対する注意の喚起が可能で、より高い学習効果を期待できると考える。

**謝辞** 本研究は、文部科学省の平成 19 年度私立大学学術研究高度化推進事業オープン・リサーチ・センターのタンジブル・ソフトウェア教育の研究の助成による.

## 参考文献

- 1) 大長達也, 丸山広, 中村太一: "係り受け構造を用いた読みやすい文章への推敲支援", 情報 処理学会第 69 回全国大会, 6Q-6 pp.2-433 2-434 (2007)
- 2) 須藤崇志,丸山広,鈴木健之,中村太一:"読点の打ち方と語の並べ方の競合回避による日本語文章の推敲支援手法",情報処理学会第70回大会,6ZG-3 pp.4-733 4-734 (2008)
- 3) 須藤 崇志, 丸山 広, 中村 太一: "文を分かりにくくする要因の分析と改善支援手法の提案", 電子情報通信学会技術研究報告. KBSE,108(65), pp.41-46 (2008)
- 4) 大野博之,稲積宏誠:"技術文書作成能力の育成を目指した教育支援ツールの開発",電子情報通信学会第18回データ工学ワークショップDEWS2007, D9-4 (2007)
- 5) 菅沼明, 牛島和夫: "字面解析を応用した日本語文章推敲支援ツールの開発", 情報処理学会研究報告, Vol.97, No.14, pp.1-8 (1997).
- 6) 木下是雄: "理科系の作文技術", 中公新書, pp.244, 1981.
- 7) "山之内総合研究所のサイト わかりやすいマニュアル作成のための実践テクニカルライティングセミナー -", http://www.yamanouchi-yri.com/.
- 8) 矢野米雄,緒方広明,榊原理恵,脇田里子:"日本語作文教育のためのネットワーク型添削支援システム CoCoA の構築",教育システム情報学会誌 VOL.14 No.3 特集号:臨時増刊 (1997).
- 9) 片山章郎: "新入生の文章力に対する一考察", 日本教育情報学会, 年会論文集 No.17, pp. 184-187 (2001).