# 推敲課題発見のためのテキスト分析 - 推敲支援システムの構築を目指して -

乾 裕子　　岡田直之
{ h_inui, okada }@pluto.ai.kyutech.ac.jp
九州工業大学大学院情報工学研究科

## 1. はじめに

### 1.1 背景

報告書やマニュアルなどのテキストの文書化にともない、さまざまな文章作成支援システムが開発されている。形態素解析技術やパタンマッチングを利用した同音異義語の使い分けや送り仮名の使い方など校正支援機能は実用化のレベルにある[9]。構文レベルの支援としては、長文の分割処理[10,14]や名詞句の並列構造に関する研究[12]が進められているが、意味処理については不十分である。そこで、本研究では文章の意味内容を考慮した推敲支援システム構築に向けて研究を進めている[13]。具体的な課題としては、1) 推敲課題の整理、2) 推敲課題の検出および修正規則の作成、3) 規則の評価、4) 規則の実装などが挙げられる。本稿では、とくに推敲課題の整理について文生成過程で生じるさまざまな不具合がどのように表現されているか、また不具合の要因を調べ、問題の体系を明らかにする。

### 1.2 関連研究

推敲のモデルについては、認知科学の立場からFlowerらの提案がよく知られている[3]。Flowerらは、プランニング部、アイデア・概念などから文字列への変換部、推敲部の三つのモジュールを含む作文過程を中心に、話題・読み手プランについての知識などを管理する書き手の長期記憶や、話題・読み手によって変化する修辞学的課題を管理するモジュールとの相互的処理によって統合的に文生成される作文モデルを提唱している。しかし、モデル全体の妥当性は明らかにされていない。とくに推敲過程の実装という立場から見ると妥当性評価の不足だけでなく、プロセスの各モジュールの入出力と処理範囲が明らかでない、処理プロセスと知識との対応関係が具体的でない、処理レベルや推敲すべき問題の選択条件が明らかでない等さまざまな問題がある。すなわち、語彙的な処理から文脈、段落間処理までを行う包括的な推敲支援システムを構築する場合のシステムのコンテンツとなるべき文章の不具合（以下、推敲課題と呼ぶ）の整理、各課題の検出・修正方法などは明らかになっていない。また、Faigleyらは推敲の方法を意味変化の有無という観点で区別しているが[1,2]、変化対象となっている意味や意味の単位については事例に一貫性がなく、単位についても具体的な言及がない。

一方、文章構成論では、いわゆる"悪文"を対象にその修正方法が種々議論されている[5,6,7,11]。しかし、これらの分類には観点の異なる不具合が含まれており、実装に必要な一貫した観点から整理された推敲課題を挙げるには至っていない。また、作文教育・教育心理学の分野では教育支援が目的であるため、推敲能力に関する個人差の研究や、生徒・学生の書いた文章に関する特徴や傾向の調査分析などを中心とした研究が多い[8]。

## 2. 文生成モデルと推敲過程

ここでは、本研究における推敲モデルについて述べる。推敲過程は文章生成モデルの中の各モジュール内およびモジュール間のフィードバックとみなす。したがって、文章生成モデルの各モジュール間の機能および入出力を明確にする必要がある。

図1に示すように、文生成モデルへの入力となるのはイメージ、音、記号などの外界からの情報である。これらの情報が解釈過程（ステップ1）を経て論理式あるいは意味ネットワークなどに記号化される。記号化された情報から格フレーム・談話構造など言語構造にするのが変換過程（ステップ2）であり、言語構造を実際の単語や文、文章として出力するのが表出過程である（ステップ3）。ステップ1に関しては、[4]など研究を進めてきた。本研究では特にステップ2の言語構造への変換過程における意味処理について推敲の観点から検証する。文生成の過程において、推敲のフィードバックは各過程の中でも、また前段階・前々段階に向かっても起こりうる。各過程で、どのような推敲課題が生じるのか、さまざまな推敲課題においてどのようなフィードバック（=検出）をすればよいか、検出条件は何か、検出した推敲課題の修正方法はどのようなものか、これらの問題を解決することで、各過程の処理役割の具体化と推敲知識の作成を進める。

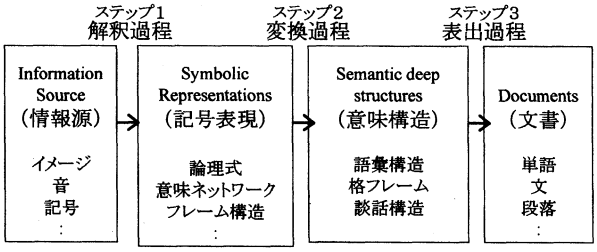


図1 文生成モデル

## 3. 分析

文生成モデル（図1）における各段階の不具合がどのように推敲課題として現れるかを網羅的に把握するためには、ある程度まとまった分量のデータを対象に分析する必要がある。そこで本研究では、推敲課題の対象である誤りやわかりにくい表現を分析するために、専門家による悪文の推敲例を多数収集する。以下、データの作成方法について説明する。

### 3.1 分析データの作成

推敲支援システムにおいて、テキスト中の誤りだけを検出するのは不十分である。読み手にとってわかりにくい表現や構文をわかりやすくすることが重要である。しかし、実際には、文書の目的や読み手の知識によってわかりにくさは変動する。そこで、われわれは、推敲対象を調査報告書、業務報告書、企画書、案内書など実務文書に絞り文書を収集した。また、テキストの一部を図2に示す。対象としたテキストから連続する10文程度の文章を取り出した195文が推敲データの元となるテキストである。

《 original sentence 》
#1 都市内の大規模開発による周辺交通への影響価と交通計画の策定のために建設省が作成した交通計画検討マニュアルの改訂を目的とする調査である。……..

**図2 交通計画検討調査のテキスト例**

《 original sentence 》
#1 都市内の大規模開発による周辺交通への影響評価と交通計画の策定のために建設省が作成した交通計画検討マニュアルの改訂を目的とする調査である。……..
《 corrected sentence-1 》
本調査は(593)、都市内の….を目的としている。
( *主題がない* )

《corrected sentence-2》
本調査の目的は(626)、次の二つである(627)。一つは、…..すること(628)。もう一つは、….. することである。
( *結論を先に述べるべき。文が長いので分割する* )
:

**図3 異なる担当者による推敲テキストの例**

次に推敲方法は、校正担当者、言語学および言語処理の研究者15人を対象に一人あたり3-4種類の文章に対して下記の作業を行った;

1) 理解しにくい部分あるいは不適切な表現に対する指摘
2) 上記の箇所に対する修正
3) 指摘と修正に対する理由

ステップ1での形成内容に関わるような不具合については、厳密には筆者の意図を確認しないと修正ができない。特に、2)については、図1に従うとステップ1からステップ3までの、三段階で問題が発生する。しかし、その場合にも、できる限り前後の文脈などの情報から、各担当者の判断で意味あるいは形成内容を損なわない修正を指示した。また、修正の個人差をできるだけ少なくするため、元の文書の意味を大幅に変更する修正、段落を修正を行わないように指示した。

この結果、得られた推敲データは述べ 1218 文となり、一種類の文章あたり平均 6 例、修正対象として指摘された箇所は約 1400 個所にのぼった。修正の具体例を図3に示す。図3は同じ文に対して異なる専門家による異なる修正結果の例である。修正推敲テキストにおけるゴシック体は修正箇所を示し、通し番号がふられている。また、斜字体で書かれているものは修正理由である。なお、修正理由については記述されていないものもある。

### 3.2 分析方法

本章では、推敲データの分析に関し、その方法、作業および結果を示す。図3において、通し番号をふった修正箇所は文生成モデル(図1)における各過程での不具合に起因する。本稿では、いずれの過程にどのような不具合が生じるかを分類・分析し、特に変換過程での不具合に注目して生成文章を読みにくくする要因を分析する。

分析は以下の方法で行った。自由記述の推敲理由に書かれた「意味がわかりにくい」「理解しにくい」「○○を明確に」「冗長」「くどい」など顕著な意見から;

・「わかりにくさ」の要因を意味的なものと構文的なものに分ける
・意味が「明確でないもの」と「冗長・くどいもの」を分ける
・システム化する際の処理レベルを明らかにする

以上の点からの分析が必要なことが明らかになり、下記の分析手順を作成した。

**分析手順**

1) 指摘箇所が変換過程での不具合であるかをチェック
2) 指摘箇所とその修正例の間の;
  2a) 意味の過不足の分析
    2a-1) 意味情報の不足
    2a-2) 意味情報の過多
    2a-3) 不適切な意味情報
  2b)指摘箇所の意味構造レベルの分析
    2b-1) 談話構造における課題
    2b-2) 格構造における課題
    2b-3) 語彙構造に関する課題
3) 過不足、不適切の対象となっている意味内容の考察

以下、上記の手順にしたがった分析例を示す。

### 3.3 分析例

分析方法にしたがって分類の事例を示す。事例は、原文Sと、修正文Tであらわす。

#### 3.3.1 情報の不足

図3に示した＜主題化 topicalize＞の不足の例は、この項目に分類される。また、同様に談話構造上の不足としては、＜詳細化 elaborative＞(#1)、＜比較 comparetive＞(#2)もある。また談話構造だけでなく、語語句の不足＜詳細化 elaborative＞(#3)や、＜行為の相手 dative＞(#3)のように格構造上にも見られる。しかし、格要素自体が欠落する推敲課題は今回のデータには見られなかった。

#1 S: 本研究では、...非集計モデル適用の必要性を考察し、マニラ首都圏をケーススタディとして、...
  T: 本研究では、...非集計モデル適用の必要性を考察し、具体的にはマニラ首都圏をケーススタディとして

表1 情報の過不足に着目した推敲課題の分類

| 情報の過不足 | 言語構造 格構造 | 語句 | 談話 | 総計 |
|---|---|---|---|---|
| 不足 | 6 | 12 | 50 | 68 |
| 過多 | 19 | 28 | 42 | 89 |
| 誤り | 25 | 1 | 21 | 47 |
| 不適切 | 1 | 5 | 5 | 11 |
| 総計 | 51 | 46 | 118 | 215 |

#2 S: A市の成長は、...他都市との立地環境面での比較優位をもたらしたという側面もある。

T1: 立地環境面において他都市に比べ、比較的優位をもたらしたということも挙げられる。

T2: 他都市に比べ、立地環境面で優位にあることによる

T3: 立地環境面で他都市より比較優位をもたらしたという側面がある

#3 S:「マルチメディア社会到来の背景と道路行政の対応」に関する講演会開催について (講演者依頼の検討案の題目)

T: ...講演依頼について

#4 S: ...今後、A市は...高度な都市機能の集積が期待されるなかで、...イメージを損なわないような都市づくりのあり方が求められている。

T: 今後、A市には...が期待されるなかで、...が求められている。

### 3.3.2 情報の過多

この項目の推敲課題としてもっとも多いのは、同じ語の重複、あるいは同義語・類義語の重複である。また、重複でなくても過多とみなされるのは文の意味に寄与しない表現、あるいは実務文書にとっては不要な主観的表現(一部の副詞、係助詞、接続助詞(#5)など)である。情報の不足同様、過多にも語句<同義語 sem repetition>の重複(#4, #5)、格構造では<対象格 objective>の重複(#6)、談話構造では不要な<逆接 adversative>(#5)のように、すべての段階に推敲課題が見られる。

#4 S: ...△△駅周辺整備上の観点からも有意義である。

T: 整備上も／整備の観点からも

#5 S: 図1はその権利変換の仕組みをごく簡単に模式化し図解したものであるが、権利変換には原則型...がある。

T: ...模式化したものであるが図解したものであるが

T: 図1はその権利変換の仕組みをごく簡単に模式化し図解したものである。権利変換には原則型...がある。

#6 S: ...都市交通計画手法は、パーソントリップ調査をベースとした4段階推定法による集計型のモデル分析によって計画が立案されている。

T1: ...都市交通計画手法は、...集計型のモデル分析である。

T2: ...都市交通計画手法は、...集計型のモデル分析によって立案されている。

T3: ...都市交通計画手法は、...集計型のモデル分析によって計画されている。

表2 情報の過不足・不適切の対象である意味内容(semantic contents)

| 情報の過不足 | SC(semantic contents) | 言語構造 格構造 | 語句 | 談話 | 総計 |
|---|---|---|---|---|---|
| 不足 | elaborative<*-*> | | 7 | 6 | 13 |
| | cause-effect | 1 | | 9 | 10 |
| | contrastive<*-*> | | 4 | 5 | 9 |
| | topicalize<*> | | | 8 | 8 |
| | means-aim | 1 | | 6 | 7 |
| | and<*-*> | | | 7 | 7 |
| | ref-anaph | | | 5 | 5 |
| | parallel<*-*> | 1 | 1 | 1 | 3 |
| | dependent-governer | 2 | | | 2 |
| | condition-apodosis | | | 2 | 2 |
| | aim-contents | | | 1 | 1 |
| | or<*-*> | 1 | | | 1 |
| 不足 計 | | 6 | 12 | 50 | 68 |
| 過多 | repetition<*> | 11 | 15 | 5 | 31 |
| | repetition<ref-anaph> | | | 11 | 11 |
| | topicalize<*> | 1 | | 9 | 10 |
| | repetition<dep> | | | 7 | 7 |
| | redandancy<adversative> | | | 6 | 6 |
| | repetition<objective> | 5 | | | 5 |
| | redandancy<*> | 1 | 3 | | 4 |
| | repetition<adversative> | | | 3 | 3 |
| | emphasis<*> | | 3 | | 3 |
| | emphasis<extent> | 1 | 2 | | 3 |
| | repetition<extent> | | 3 | | 3 |
| | redandancy<extent> | | 2 | | 2 |
| | redandancy<illustrative> | | | 1 | 1 |
| 過多 計 | | 19 | 28 | 42 | 89 |
| 誤り | dependent-governer | 23 | | 1 | 24 |
| | emphasis<topicalize> | 1 | | 8 | 9 |
| | contrastive<*-*> | | | 8 | 8 |
| | redandancy<*> | | | 2 | 2 |
| | cause-effect | 1 | | | 1 |
| | means-aim | | | 1 | 1 |
| | ref-anaph | | | 1 | 1 |
| | lexical error<*> | | 1 | | 1 |
| 誤り 計 | | 25 | 1 | 21 | 47 |
| 不適切 | condition-apodosis | | | 4 | 4 |
| | simple←difficult | | 4 | | 4 |
| | dependent-governer | 1 | | | 1 |
| | and<*-*> | | | 1 | 1 |
| | positive←negative | | 1 | | 1 |
| 不適切 計 | | 1 | 5 | 5 | 11 |
| 総計 | | 51 | 46 | 118 | 215 |

### 3.3.3 情報の誤り・不適切

不足あるいは過多に属さない推敲課題を不適切・誤りとした。意味のわかりにくい語句の<簡単化 simplify>(#7)、不適切な<対象格 objective>(#8)、不適切な<包含的 inclusive>関係の表出(#9)など、各構造に推敲課題が見られる。

#7 S: ...市街地の土地の合理的で健全な高度利用と

T: ...市街地の土地を合理的かつ健全に効率よく利用することであり

#8 S: ...「マルチメディア社会」については近年急速に脚光を浴びはじめた概念であり、

T: ...「マルチメディア社会」は近年急速に脚光を浴びはじめた概念であり、

#9 S: 大規模店や地下街の開設が相次ぐ東口周辺地区では、商業活動に大きな変化があり、中でも北口周辺の商業は低滞しつつある……

T: …商業活動に大きな変化があるのに対し、北口周辺の商業は低滞しつつある

## 4. 考察

### 4.1 結果の考察

前章の手順により分類した結果について示す(表1)。

全体的な傾向としては、情報の過不足、誤り・不適切は談話レベルに生じやすい、すなわち、書き手にとって談話レベルにおける形成内容から意味構造への変換が難しいことがわかる。この傾向は、情報の不足について、より顕著である。文章の書き手は、一文ごとの文を完成させることに意識を傾けがちである。したがって、予測と異なり、述語に示される行為に対する主体や対象の格の欠落は少なかった。

さらに、分析過程において<>で表した過不足・誤りの対象となっている意味内容<semantic contents>を加えてまとめた結果を表2に示す。例えば、不足している情報について、<cause>, <aim>, <addition>などが上がっているのを見ると、従来も指摘されているが、報告書や技術文などの文章では論理的を明示した文章展開が重要視されるということが直観だけでなく帰納的に調査分析から得られた。

### 4.2 分析の評価 ー森岡の分類との比較ー

森岡の分類では、作文における誤りの類型を思考上の誤りと文法上の誤りに分け、思考の誤りを以下の14種類に分類している[6]。

断片文(不完全な文)/混合文(文をいくつもつなげたもの)/だらだら文( 分離すべきもの,従属関係にすべきもの,むだな語句を省くべきもの)/重複/論理的な区切り/短すぎる文/思想の飛躍/あいまいな代名詞の指示/宙に浮いた修飾語/位置の不適当な修飾語/対をなす思想のととのえ方/語の脱落/よじれ/強調

われわれの分類と比較すると、この分類観点は一貫性に乏しいことがわかる。たとえば、宙に浮いた修飾語、語の脱落は格構造上の、また対をなす思想のととのえ方は格構造・談話構造上の不足と整理できる。重複は各構造上の過多、強調はどのような強調かによって過多の場合、不適切・誤りの場合いずれも考えられる。よじれとは、格要素あるいは修飾要素と、それらが係る述部との意味的共起制約の違反である。また、だらだら文にまとめて分類されている誤りは、[6]の事例を見ると、むだな語句を含むものが情報の過多、分離すべきもの・従属関係にすべきものが談話構造上の情報の不足であることがわかった。

## 5. おわりに

本稿では、情報の過不足・不適切といった意味上の不具合が文のどのような構造で生じているかという二つの観点で推敲課題を分類することにより、不具合の原因となっている意味内容(SC)を明らかにすることができた。これは、言い換えると推敲支援システムで検出・修正すべき問題の体系を明らかにしたということになる。この分類体系は、従来さまざまな観点から分析されてきた、文をわかりにくくする要因を一貫した観点によって整理した結果である。

さらに、semantic contentsを取り出したことにより、談話関係における因果関係の重要性や具体性・論点の明確化、対比的な関係を明確に表現することの難しさなど帰納的に推敲課題を取り出すことができた。

今後の課題は、分類された推敲課題を各semantic contentsごとの検出・修正規則を記述し、規則の評価と実装を行うことである。

**参考文献**:


[1] Lester Faigley and Stephen P. Witte 、 Analyzing Revision、 College Composition and Communication、 32、 pp.400-414、 1981.

[2] Lester Faigley and Stephen P. Witte. "Measuring the effects of revisions on text structure", New Directions in Composition Research, pp.95-108、 1984.

[3] Linda Flower, John R.Hayes, Linda Carey, Karen Schriver, and James Stratman, "Detection, Diagnosis and the Strategies of Revision", College Composition and Communication, vol.37, No.1, 1986.

[4] Okada, N. "Story Generation Based on Dynamics of the Mind," Computational Intelligence, 8(1), 1992.

[5] 岩淵悦太郎『悪文』日本評論社、 1960

[6] 森岡健二『文章構成法』至文堂、 1963

[7] 樺島忠夫『文章構成法』講談社現代新書、 1980

[8] E・D・ガニエ著 赤堀・岸監訳『学習指導と認知心理学』パーソナルメディア, 1989

[9] 高橋・牛島:計算機マニュアルの分かりやすさの定量的評価方法、情報処理学会論文誌 vol.32.No.4、1991

[10] 武石・林:接続構造解析に基づく日本語複文の分割、情報処理学会論文誌 vol.33.No.5、 1992

[11] 中村明『悪文』、 ちくま新書、 1995

[12] 菅沼・山村・牛島:日本語文における名詞句の並列構造の推定およびその推敲支援への適用、 情報処理学会論文誌 vol.38.No.7、 1997

[13] 乾、岡田:「わかりにくい」表現の検出規則作成-推敲支援システムの実装をめざして-、言語処理学会第6回年次大会発表論文集、pp179-182, 2000.

[14] 野上、藤田、乾:文分割による連体修飾節の言い換え、言語処理学会第6回年次大会発表論文集、pp215-218、2000.