冗長な文章の人手による分析

村田 真樹 [*1]　徳久 雅人 [*1]　馬 青 [*2]
[*1] 鳥取大学大学院 工学研究科 情報エレクトロニクス専攻
[*2] 龍谷大学 理工学部 数理情報学科
[*1]{murata,tokuhisa}@ike.tottori-u.ac.jp
[*2] qma@math.ryukoku.ac.jp

1　はじめに

文の改善の研究としては「誤字の修正・適切な語の選択」[1, 2, 3] と「語順の修正・語と語の係り受けの誤りおよび複雑性の修正」[1, 4, 5, 6] と「冗長な表現の改善」が考えられる．このうち「誤字の修正・適切な語の選択」と「語順の修正・語と語の係り受けの誤りおよび複雑性の修正」の研究に関しては既に先行研究が多数ある．しかし「冗長な表現の改善」に関係する研究はあまりないため本研究で扱う．

例文として「まず初めにマシンの点検を行う．」という文を考えてみよう．文中の「まず」と「初め」の 2 つの単語は同じ意味を含んでおり冗長である．また「点検を行う」については意味の薄い「行う」を省くことができる．このように文内に同じ意味の単語が複数回出現する文や，余分な漢字表現を含む言い回しは，冗長でわかりにくい．上述した例文は冗長箇所を削除・修正することで「まずマシンを点検する。」という簡潔な文に修正できる．本研究では，上記のような文を冗長な文とし，冗長な文に関する研究を行う．本研究は文書作成者の推敲を手助けするシステムの構築や，限られた字数で文字入力をする際に，綺麗に収めるのにも役立つ．

筆者らは既に冗長な文に関する研究を進めている．文献 [7, 8] では 1 文における冗長な文の分析・検出・修正を行った．文献 [9] では複数文にまたがる冗長な文章の分析・検出を行った．文献 [9] での冗長な文章の分析は，作例データで分析を行っていた．これに対して，本稿では，冗長な文章の分析を，実際の文章において行う．

冗長な文章の分析を実際の文章で行うために，本稿では，序盤の取り組みとして，既存の手法や簡単に行える手法を用いて冗長な文章を自動で収集する．文献 [9] の研究で，冗長な文章の検出に冗長度（同じ語が多く出現すると高くなる指標）を利用することが良いことがわかった．このため，冗長度の高い文章を収集することで，冗長な実例の文章を収集しやすい．また，1 文が長い文，1 段落に多くの文を含む文章も冗長な文章になりやすい．これらの情報を考慮して，本研究では，冗長度，文の長さ，段落内の文数の情報を用いて，効率よく冗長な文章を収集する．そして，収集した冗長な実例の文章を人手で分析し，冗長な文章の特徴を調査する．

本研究の主な主張点と特徴を以下に整理する．

- 冗長度，文の長さ，段落内の文数の情報を用いることで，効率よく冗長な実例の文章を収集した．
- 冗長な実例の文章を人手で分析し，その特徴を調査した．具体的に，冗長な実例の文章の改善の種類として，「文章の分割」「内容のある表現の削除」「文章の順序変更」「同じ表現の繰り返しの削除」「装飾・表現の変更」「複数の文を 1 文にまとめる」「説明の追加 」があることがわかった．

2　冗長な文章の収集

冗長な文章の収集には，以下の特徴を利用する．

- 文章中に同じような語を多く含む文は冗長である．
- 長い 1 文は冗長である．
- 多くの文を含む段落は冗長である．

実験に用いるデータは以下の 3 種類を用いる．

- KNB コーパス (Kyoto University and NTT Blog コーパス)
- 「Yahoo!知恵袋データ」の回答データ
- ウィキペディア

それぞれについて以下の 4 種類のデータを作成する．

- 1 段落で文数大
- 1 文で文字数大
- 1 段落で冗長度大
- 1 文で冗長度大

ただし，京大ブログコーパスは段落の情報がないため，4 種類のうち，段落が関係する「1 段落で文数大」「1 段落で冗長度大」のデータは作成しない．このため合計で 10 種類のデータを作る．

上述における冗長度は以下の式で計算する．

$$\text{冗長度} = \frac{N}{V} \tag{1}$$

ただし，V は対象データ内での単語の異なり数であり，N は対象データ内での延べ単語数である．

表 1: 冗長な文章のデータ数

	データ 1	データ 2	データ 3	データ 4	データ 5	データ 6	データ 7	データ 8	データ 9	データ 10	合計
文書の種類	KNB コーパス		知恵袋回答データ				ウィキペディア				
1 文か 1 段落か	1 文	1 文	1 段落	1 文	1 段落	1 文	1 段落	1 文	1 段落	1 文	
1 段落で文数大			文数大				文数大				
1 文で文字数大	文字数大			文字数大				文字数大			
冗長度大か否か		冗長度大			冗長度大	冗長度大			冗長度大	冗長度大	
収集した冗長な文章の数	174	100	162	53	41	81	134	50	25	85	905
冗長な文章を人手で修正して作成した冗長でない文章の数	204	117	193	69	48	96	148	59	29	93	1056
1 つの冗長な文章から 2 つの冗長でない文章を作成した回数	30	17	31	16	7	15	14	9	4	8	151

上記の方法で収集したデータを人手で冗長かいなかを判定し，冗長な場合は，冗長でない文章になるように修正を行う．記号を多く含んでいるなど不適切な文章は除外する．修正や分析が比較的困難な文章も除外する．1 つの冗長な文章から複数の冗長でない文章を作成できるときは 2 つまで作る．3 個以上は作らない．

実際に上記の手法により冗長な文章を収集した．その結果を表 1 に示す．10 種類のデータは便宜上，表のようにデータ 1 からデータ 10 と呼ぶ．

表の「冗長でない文章の数」は，冗長な文章を修正して得た冗長でない文章の数である．

データ 1 は 1 文が約 200〜400 字で，データ 2 は冗長度が約 1.5〜1.8 で，データ 3 は 1 段落 9 文で，データ 4 は 1 文が約 100〜200 字で，データ 5 は冗長度が約 2.5〜3.3 で，データ 6 は冗長度が約 1.8〜3.2 で，データ 7 は 1 段落が約 15〜40 文で，データ 8 は 1 文が約 100〜250 字で，データ 9 は冗長度が約 2.8 で，データ 10 は冗長度が約 2.5 であった．

冗長度，1 文での文字数，1 段落での文数を利用することで，多くの冗長な文章を簡便に収集できた．

3 冗長な文章の人手による分析

3.1 改善例における修正の種類の数

2 節で収集した冗長な文章を人手で分析する．データ 1 からデータ 10 の 10 種類のデータについて，それぞれから 20 個の冗長な文章の改善例を取り出し，その改善例でなされた修正の種類の数を調査した．その結果を表 2 に示す．

「行う」「思う」などを削除したり末尾の表現を修正して文章を簡素にする修正はほとんどすべての改善例においてなされていた．このため，この種類の修正は調査していない．

表 2 では，「文章の分割」「内容のある表現の削除」「文章の順序変更」「同じ表現の繰り返しの削除」「装飾・表現の変更」「複数の文を 1 文にまとめる」「説明の追加」を大項目とし，大項目の出現頻度順に示している．また，表では大項目の下にも下位の項目を設けている場合もある．

表 2 の修正の種類の項目は，一つの改善例に複数付与する場合がある．

3.2 改善例における修正の種類ごとの例

改善例における主要な修正の種類ごとに例文を示す．

- 1 文を複数の文に分割する
  - (修正前)
    これだけだとオーソドックスな京都観光からあまりに外れてしまい、ややもすると「こいつは５００円の図書券欲しさにこんなしょうもないエントリーを投稿したのか」という誇りをも受けかねないと思うので、最後にこれから卒業までに一度は行っておきたい京都の観光名所を挙げて終えることにします。
  - (修正後)
    これだけだとオーソドックスな京都観光から外れてしまう。「こいつは５００円の図書券欲しさにこんなしょうもないエントリーを投稿したのか」という誇りを受けかねないので、卒業までに一度は行っておきたい京都の観光名所を挙げることにします。
- 箇条書にする
  - (修正前)
    建造物は、中世以前のものも残るが、フランス第三共和政期のパリ改造やベル・エポックの建造物、あるいはフランス革命 200 周年期のグラン・プロジェの建造物など、各時代の世界の最先端のものが多い。
  - (修正後)
    建造物は、中世以前のものも残るが、以下のように各時代の世界の最先端のものが多い。
    - フランス第三共和政期のパリ改造やベル・エポックの建造物
    - フランス革命 200 周年期のグラン・プロジェの建造物
- 1 段落を複数の段落に分割
  - (修正前)
    私は受話器を取り、大学事務局へ電話を掛けてみた。するとオペレーターは、「カリフォルニア大学です。」と答えた。私は、「バークレー校かね？」と聞くと、彼女は「いいえ、違います。」と答えた。「では、どこへ繋がっているのかね？」と私が問うと、彼女は「UCLA です。」と答えた。すかさず「じゃあ、何故

表 2: 冗長な文章の改善例における修正の種類の数

	データ 1	データ 2	データ 3	データ 4	データ 5	データ 6	データ 7	データ 8	データ 9	データ 10	合計
文書の種類	KNB コーパス		知恵袋回答データ				ウィキペディア				
1 文か 1 段落か	1 文	1 文	1 段落	1 文	1 段落	1 文	1 段落	1 文	1 段落	1 文	
1 段落で文数大			文数大				文数大				
1 文で文字数大	文字数大			文字数大				文字数大			
冗長度大か否か		冗長度大			冗長度大	冗長度大			冗長度大	冗長度大	
文章の分割	14	7	3	11	4	8	14	13	10	15	99
1 文を複数の文に分割する	14	7	0	10	0	5	0	7	2	7	52
箇条書にする	0	0	3	3	4	4	2	6	4	9	35
1 段落を複数の段落に分割	0	0	0	0	0	0	13	0	5	1	19
内容のある表現の削除	4	2	10	1	8	2	11	2	10	2	52
補足表現の削除	1	1	10	0	8	2	11	1	9	2	45
修飾語の削除	3	0	0	1	0	0	0	0	1	0	5
接続詞の削除	0	1	0	0	0	0	0	1	2	0	4
条件節の削除	0	1	0	0	0	0	0	0	0	0	1
文章の順序変更	4	5	1	7	3	1	1	5	4	9	40
結論を前に補足を後ろに	1	2	0	3	3	2	1	0	2	6	20
主語と述語を近づける	0	1	0	0	0	0	0	3	1	3	8
理由を後ろに	3	0	0	3	0	0	0	0	0	0	6
理由を前に	0	0	0	1	0	0	0	0	0	0	1
主題を前方に	0	1	0	0	0	0	0	0	0	0	1
順位の順に	0	1	0	0	0	0	0	0	0	0	1
時系列の順に	0	0	0	0	0	0	1	0	0	0	1
同じ表現の繰り返しの削除	0	6	2	1	2	4	3	2	0	3	23
装飾・表現の変更	0	1	1	1	2	6	0	3	1	0	15
引用符・括弧の利用	0	1	0	0	2	0	0	2	0	0	5
読点の挿入，場所の変更	0	0	0	0	0	2	0	1	1	0	4
読点の削除	0	0	0	0	0	3	0	0	0	0	3
平仮名を漢字に	0	0	1	1	0	0	0	0	0	0	2
漢字を平仮名に	0	0	0	0	0	1	0	0	0	0	1
ある表現を片仮名に	0	0	0	0	0	1	0	0	0	0	1
複数の文を 1 文にまとめる	0	0	1	1	0	0	0	0	0	0	2
説明の追加	0	0	0	0	1	0	0	0	0	0	1

最初からそう言わないのかね？」と聞くと、「そう言う様に言い付けられているもので・・・。」と答えた。そこで私は翌朝、大学事務局へ赴き、「今日正午よりカリフォルニア大学ではなく、UCLA と電話で受け答えるように言いつけてくれ。」とメモを認めた。すると彼ら（事務局）は「バークレーは気に入らないでしょうね。」と言った。私は「まあ、様子を見ようじゃないか。私達にだって、バークレーの事務局の許可なしに出来ることだってあるさ。」と答えた。

– (修正後)

大学事務局へ電話を掛けると、オペレーターが「カリフォルニア大学です。」と答えた。「バークレー校か」と尋ねると否定するので、どこに繋がっているかを問うと「UCLA」と答えた。最初から UCLA と言わないのは、そう言い付けられているとのことだった。

そこで翌朝、私は大学事務局へ赴き、「今日正午よりカリフォルニア大学ではなく、UCLA と電話で受け答えるように言いつけてくれ。」と要求した。事務局は「バークレーは気に入らないでしょうね。」と言ったが、私はこう答えた。「様子を見よう。私達にもバークレーの事務局の許可なしに出来ることがある。」と。

- 補足表現の削除 (下の例では括弧内を補足表現として削除)
  – (修正前)

柱の中に巣をつくるタイプの蟻もいます。(うちの玄関脇の柱にも一家族住んでいます。……外だからいいようなものの……) 古い家に多いようですが、新築の家にも出ることがあるようです。自治会で床下の噴霧消毒をやっていませんか？やっているようでしたら、参加しましょう。それから、本職の方に電話して聞いてみたほうがいいと思います。業者は複数、見積もりをとりましょう。一番安いところが一番良い業者とは限りません。電話で相談したときの印象なんかにも気をつけて選んでくださいね。こちらから電話したときの契約はクーリングオフの対象外ですから。

– (修正後)

柱の中に巣をつくる蟻もいます。古い家に多いようですが、新築の家にも出ることがあります。自治会で床下の噴霧消毒をやっていたら、参加しましょう。それから、本職の方に電話で聞いてみましょう。業者は複数、見積もりをとりましょう。一番安いところが一番良い業者とは限りませんので、電話の印象にも気をつけて選んでください。こちらから電話したときの契約はクーリングオフの対象外ですから。

- 修飾語の削除
  – (修正前)

Ｊａｖａ対応機種では 世界標準のＪａｖａ規格のＭＩＤＰを採用しているし、日本におけるおそらく最強のスマートフォンといわれるW-ＺＥＲＯ３

ファミリーに採用されているＷｉｎｄｏｗｓ　Ｍｏｂｉｌｅに対しては強力な開発環境や豊富なアプリケーションが用意されている。
  - (修正後)
    Ｊａｖａ対応機種ではＭＩＤＰを採用しているし、Ｗｉｎｄｏｗｓ　Ｍｏｂｉｌｅに対しては強力な開発環境や豊富なアプリケーションが用意されている。
- 結論を前に補足を後ろに
  - (修正前)
    脂肪肝・・・高脂血症・・・「私は薬に殺される」という本を読んで勉強中。なんでも治療薬がものすごく危険らしい。いわゆる副作用が・・・だ。医者から出された中性脂肪とコレステロールの薬のせいで、二度と治らぬ体にされ、俺は今、死にかかっている。元気で働くために、家族で幸せになるために、そして長生きするために飲んだ薬のせいで――。あなたの飲んでいる薬は大丈夫か？
  - (修正後)
    脂肪肝、高脂血症の治療薬がものすごく危険らしい。あなたの薬は大丈夫か？医者から出された中性脂肪とコレステロールの薬のせいで、二度と治らぬ体にされ、俺は今、死にかかっている。元気で働くため、家族で幸せになるため、長生きするために飲んだ薬のせいで。
- 主語と述語を近づける
  - (修正前)
    今後、そういった高次心理過程も、心理学における行動・認知レベルの研究に加えて、生物学における分子レベルの、細胞レベルの、皮質のグローバルなレベルでの研究を進めることにより、両分野のあいだで統合的に説明できるようになるかもしれない
  - (修正後)
    心理学における行動・認知レベルの研究に加えて、生物学における分子、細胞、皮質のグローバルなレベルでの研究を進めることで、今後、そういった高次心理過程も、両分野で統合的に説明できるだろう。
- 理由を後ろに
  - (修正前)
    これだけだとオーソドックスな京都観光からあまりに外れてしまい、ややもすると「こいつは５００円の図書券欲しさにこんなしょうもないエントリーを投稿したのか」という誇りをも受けかねないと思うので、最後にこれから卒業までに一度は行っておきたい京都の観光名所を挙げて終えることにします。
  - (修正後)
    最後に、卒業までに一度は行くべき京都の観光名所を挙げて終えることにします。これだけだとオーソドックスな京都観光から外れすぎて、５００円の図書券欲しさにこんなしょうもないエントリーを投稿したのかと非難されかねませんので。
- 同じ表現の繰り返しの削除
  - (修正前)
    スポーツをしている人や散歩している人，昼寝をする人など，日常の京都人の姿を拝見できました．
  - (修正後)
    スポーツや散歩，昼寝をする人など，日常の京都人の姿を拝見できました．
- 引用符・括弧の利用 (下の例では判決結果を引用符でくくっている)
  - (修正前)
    パラマウント社は「シェーン」についてのみ知的財産高等裁判所に控訴したが、同裁判所は 2007 年 3 月 29 日、著作権は 2003 年 12 月 31 日をもって消滅したとする一審判決を支持し、パラマウント社の控訴を棄却する判決を言い渡した（知的財産高等裁判所判決平成 19 年 3 月 29 日）
  - (修正後)
    パラマウント社は「シェーン」についてのみ知的財産高等裁判所に控訴した。しかし、同裁判所は 2007 年 3 月 29 日、「著作権は 2003 年 12 月 31 日をもって消滅した」とする一審判決を支持し、パラマウント社の控訴を棄却する判決を言い渡した（知的財産高等裁判所判決平成 19 年 3 月 29 日）

4　おわりに

本研究では，冗長度，文の長さ，段落内の文数の情報を用いて，効率よく冗長な文章を収集した．そして，収集した冗長な実例の文章を人手で分析し，冗長な文章の特徴を調査した．冗長な実例の文章の改善の種類として，「文章の分割」「内容のある表現の削除」「文章の順序変更」「同じ表現の繰り返しの削除」「装飾・表現の変更」「複数の文を 1 文にまとめる」「説明の追加 」があることがわかった．また本稿では主要な改善例の種類について，例文を記載した．

謝辞

本研究は科研費 (23500178) の助成を受けたものである．

参考文献

[1] 菅沼明, 牛島和夫. テキスト処理による推敲支援情報の抽出. 人工知能学会誌, Vol. 23, No. 1, pp. 25–32, 2008.
[2] Masaki Murata and Hitoshi Isahara. Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples based on positive examples. *IEICE Transactions on Information and Systems*, Vol. E85–D, No. 9, pp. 1416–1424, 2002.
[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, Vol. 3, pp. 85–88, 2004.
[4] 内元清貴, 村田真樹, 馬青, 関根聡, 井佐原均. コーパスからの語順の学習. 言語処理学会誌, Vol. 7, No. 4, pp. 163–180, 2000.
[5] 村田真樹, 内元清貴, 馬青, 井佐原均. 日本語文と英語文における統語構造認識とマジカルナンバー７±２. 言語処理学会誌, Vol. 6, No. 7, pp. 61–71, 1999.
[6] 乾裕子, 岡田直之. 長い文は常にわかりにくいか? : わかりにくさの要因とその依存関係. 情報処理学会 自然言語処理研究会 2000-NL-135, pp. 63–70, 2000.
[7] 都藤俊輔, 村田真樹, 徳久雅人, 馬青. 冗長な文の機械的分析と機械的検出. 言語処理学会第 18 回年次大会発表論文集, pp. 1114–1117, 2012.
[8] 都藤俊輔, 村田真樹, 徳久雅人, 馬青. パターンと機械学習による冗長な文の修正と修正のヒント出力. 言語処理学会第 19 回年次大会発表論文集, pp. 588–591, 2013.
[9] 都藤俊輔, 村田真樹, 徳久雅人, 馬青. 機械学習と冗長度を用いた冗長な文章の検出. 言語処理学会第 20 回年次大会発表論文集, pp. 939–942, 2014.