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

都藤 俊輔 [*1]　村田 真樹 [*2]　徳久 雅人 [*2]　馬 青 [*3]
[*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
[*3] 龍谷大学 理工学部 数理情報学科
[*1,*2]{s082034,murata,tokuhisa}@ike.tottori-u.ac.jp
[*3] qma@math.ryukoku.ac.jp

1　はじめに

文の生成や推敲 [1] において，注意すべきことの一つに文の冗長性の問題がある．冗長な文は読みづらく，読みやすくなるように修正する方が良いと考える．

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，冗長な文の自動修正を試みる．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に役立つ．

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] 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」を扱う研究はほとんどないため本研究で扱うこととした．

「冗長な表現の改善」についてはわれわれの先行研究 [6] があり，そこでは以下の知見を得ている．

1. 冗長な文の分析により，「可能」「という」「すること」などの表現が入った文は冗長である可能性が高い．
2. 機械学習 (サポートベクトルマシン) を利用した冗長な文の検出では，特定の表現を含む文の集合ごとに機械学習する手法 [7](特定の表現の種類の数だけ機械学習が必要) を利用した結果，0.7 から 0.8 という比較的高い F 値で冗長な文を検出できた．

先行研究 [6] では冗長な文の分析・検出を行っているが修正に関しては行われていないため，本研究では修正に関わる研究を行う．具体的には，パターンと機械学習により冗長な文の自動修正を試みる (2 節)．また，冗長な文の修正に役立つヒントを出力する方式について検討する (3 節)．

2　パターンと機械学習による冗長な文の修正

先行研究 [6] において「可能」「という」「すること」の表現が入った文は冗長である可能性が高いとの知見が得られている．このため，本節では，これらの表現を含む文について冗長な文の修正を行う．

冗長な文の検出はすでに先行研究 [6] でなされているため，本節の実験では冗長な文であることがわかっている文を入力として，その文を冗長でない文に修正することを試みる．

冗長な文の修正には，パターンを利用する方法 (2.2 節) と機械学習を利用する方法 (2.3 節) を試す．

2.1　データ

ウィキペディア[*1]において，「可能」を含む文を収集する．一文内に複数回「可能」が出現する文は本研究では用いない．収集した文の集合から文中の「可能」が別の冗長でない表現に言い換えができる文を 100 文取り出す．取り出した 100 文を人手で修正し，取り出した 100 文 (冗長な文) とその修正文を対としたものを作成し実験に用いるデータとする．上記の修正は，「可能」が存在していたことにより冗長となっていた個所のみに対して行う．例えば，「無排土での施工が可能であり、経済的である。」の文からは次のような対を獲得する．

冗長な文　無排土での施工が可能であり、経済的である。
修正文　無排土での施工ができ、経済的である。

「という」「すること」についても同様にして，上述のような文対をそれぞれ 100 文対ずつ獲得し，合計 300 文対を獲得する．それぞれの表現ごとに，100 文対を 2 分割し，50 文対を学習データ，残りの 50 文対をテストデータとする．

2.2　手法 1:パターンを用いた冗長な文の修正

手法 1 ではパターンを用いて冗長な文を修正する．

「可能」を含んだ冗長な文についてのパターンを用いた修正は次の手順で行う．

[*1] Wikipedia:http://ja.wikipedia.org/wiki/

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

1.「可能」について 2.1 節で作成した学習データに含まれる冗長な文とその修正文をそれぞれ形態素解析 ChaSen[*2]に適用し単語単位に分割をする．
2. 単語単位に分割した冗長な文とその修正文を文対ごとに差分を取る．差分検出には diff コマンド [8] を使用する．
3. 獲得した差分部分を，修正に利用するパターンとする．例えば，「無排土での施工が可能であり、経済的である。」の文からは次のようなパターンを獲得する．

   冗長な文 無排土での施工が 可能であり、経済的である。
   修正文 無排土での施工が でき、経済的である。
   パターン 可能であり → でき

   ここで「可能であり」が差分における修正前の表現であり，「でき」が修正後の表現である．このパターンは「可能であり」を「でき」に修正するパターンとなる．
   一文中に差分箇所が複数存在する文については，複雑であり処理するのが困難であるため，それらの文は「その他」という分類に分類できればよいとする．この種の文からは，「可能」を含む差分箇所の修正前の表現に対して「その他」という分類先に分類するパターンを作成する．
4. テストデータの 50 文にパターンを適用し文を修正する．

上記の 4 のパターンの適用の際，適用可能なパターンが複数存在することがある．修正に利用するパターンを一つ選ぶために，以下の二種類の方法を設けた．

PT1 最長一致 (パターンの修正前の表現が最も長いもの) するパターンにより修正する．最長一致するパターンが複数存在する場合，パターンを構成する差分表現の学習データにおける出現頻度を求め頻度が高かったものにより修正する．

PT2 パターンを構成する差分表現の学習データにおける出現頻度を求め頻度が高かったものにより修正する．

「という」「すること」についても同様にしてパターンに基づく冗長な文の修正をする．

2.3 手法 2:機械学習を用いた冗長な文の修正

手法 2 では教師あり機械学習 [**?**] を用いて冗長な文の修正を行う．機械学習法には，最大エントロピー法 (以下 MEM) を用いる．[*3]

「可能」を含んだ冗長な文についての機械学習を用いた冗長な文の修正は次の手順で行う．

1.「可能」について 2.1 節で作成した学習データに含まれる冗長な文とその修正文をそれぞれ形態素解析 ChaSen に適用し単語単位に分割をする．
2. 単語単位に分割した冗長な文とその修正文を文対ごとに差分を取る．差分検出には diff コマンド [8] を使用する．
3. 獲得した差分部分に基づき機械学習の分類先を設定する．分類先の設定方法として以下の 2 種類を用いる．

   ML1 差分部分の修正前表現 X と修正後表現 Y をあわせた「X → Y」を分類先とする
   ML2 差分部分の修正後表現を分類先とする

   分類先の例を表 1 に示す．
   一文中に差分箇所が複数存在する文の場合は，方法 ML1 では，「可能」を含む差分部分の修正前表現と「その他」の組を分類先とし，方法 ML2 では，「その他」を分類先とする．
4. 上述の分類先を付与した学習データを用いて機械学習を行う．テストデータを機械学習により分類し，分類先をもとめる．機械学習の際には，表 2 に示す素性を用いる．
   表 2 中の対象表現 X とは，修正対象となる表現のことである．「可能」を含んだ冗長な文における対象表現 X は「可能」となる．
   「という」「すること」についても「可能」と同様の処理を行う．
   方法 ML2 では修正後表現のみを分類先とするため，修正前表現を認識できず，元の文のどこの個所を修正するとよいかを把握できない欠点があり，自動修正を行うことができない．方法 ML2 を実際に文書修正に使う際は，ユーザに修正後表現を見せるのみで修正前表現は対象表現などを手掛かりにユーザに人手で判断してもらうこととなる．一方，方法 ML1 では修正前表現と修正後表現の組を分類先とするため，修正前表現を認識でき，ML2 のような欠点はなく，文中の冗長な表現を自動で修正できる．

2.4 ベースライン手法

比較のため，ベースライン手法として以下の BL1 と BL2 を用いる．

BL1 手法 2 の方法 ML1 に基づく分類先のうち最も高い頻度で学習データに出現したものを常に分類先とする

BL2 方法 ML2 に基づく分類先のうち最も高い頻度で学習データに出現したものを常に分類先とする

2.5 結果

パターンを用いる手法 (PT1 と PT2)，機械学習を用いる手法 (ML1 と ML2)，ベースライン手法 (BL1 と BL2) による冗長な文の修正結果における正解率を表 3

[*2] ChaSen:http://ChaSen-legacy.sourceforge.jp/
[*3] 本研究では，最大エントロピー法の方がサポートベクターマシンよりも高い性能を出すことを実験で確認している．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 分類先の設定方法の例

方法 ML1	方法 ML2	文
可能である→できる	できる	例えば、固体の融解や固化のプロセスをメッシュフリー法で解析することが 可能である。
不可能でない→できる	できる	単に隠れた変数理論が 不可能でない ことを示そうとしただけだった。
不可能である→できない	できない	都市の数が無限であれば、全ての都市についてチェックするのは原理的に 不可能である。
可能であり→でき	でき	無泥水・無排土での施工が 可能であり、経済的である。

表 2: 機械学習で用いる素性

素性番号	素性
1	文中に存在する対象表現 X の前 2 単語
2	文中に存在する対象表現 X の後ろ 2 単語
3	文中に存在する対象表現 X の前 2 単語の品詞
4	文中に存在する対象表現 X の後ろ 2 単語の品詞

と表 4 に示す．表 3 は修正前と修正後の両方の表現を推定できた場合に正解とする場合の評価結果である．表 4 は修正後の表現さえ推定できただけで正解とする場合の評価結果である．

BL1 で用いられた修正前と修正後の表現対を以下に示す．

可能　　可能である→できる
という　という→（Φ）
すること　することが→その他

「という」の表現対における修正後表現の「Φ」は修正前の表現を削除することを意味する．

表 3 と表 4 のどの場合でもパターンを用いる手法と機械学習を用いる手法のいずれかがベースライン手法よりも同等以上の正解率であった．表 3 のように，修正前表現を含めた推定では，パターンを用いる手法と機械学習を用いる手法のいずれかで 6 割以上の正解率を得た．表 4 のように，修正後表現のみの推定では，パターンを用いる手法と機械学習を用いる手法のいずれかで 7 割以上の正解率を得た．修正後表現だけがわかる場合でも文書の修正作業を行う作業者にとって有用な場合があるので，修正後表現のみの推定で 7 割以上の正解率を得ることができたことは有益な結果である．以上により，「可能」「という」「すること」については，パターンを用いる手法と機械学習を用いる手法がある程度冗長な表現の修正に役立つことがわかった．

しかし「すること」については「その他」という分類が多数を占めており (全体の 4 割強)，実際の修正先の表現までを推定していない場合が多い．これの対処は今後の課題とする．

表 3: 修正前表現と修正後表現の推定の正解率

表現	PT1	PT2	ML1	BL1
可能	60 %	48 %	42 %	27 %
という	54 %	68 %	62 %	65 %
すること	78 %	46 %	64 %	46 %

表 4: 修正後表現の推定の正解率

表現	PT1	PT2	ML2	BL2
可能	70 %	66 %	64 %	66 %
という	58 %	86 %	80 %	76 %
すること	80 %	46 %	72 %	46 %

パターンを用いる手法と機械学習を用いる手法で誤りが生じた原因としては，修正先の表現が非常に多く，パターンと機械学習ではそれらを網羅したり，適切に選択することが難しいということが挙げられる．

3　冗長な文の修正のヒント出力

3.1　研究の動機

2 節の実験では「可能」「という」「すること」について 6 割程度の性能で冗長な文を修正でき，7 割程度の性能で修正先の表現を推定できた．しかし，実際の文書の推敲での冗長な文の修正ではもっと確実な手法で行う必要がある場合も考えられる．

そこで，修正を行うのではなく，図 1 のように修正箇所の検出を自動で行い (先行研究 [6] の利用が可能)，さらに検出した冗長箇所の修正候補を頻度の高い順に並べ，ユーザーに提示するという方式を検討する．この方式では，冗長な個所とその修正候補が提示されるため，文書作成者の修正作業の負担を軽減できると考える．

3 節では，ユーザに提示するデータの構築を試みる．

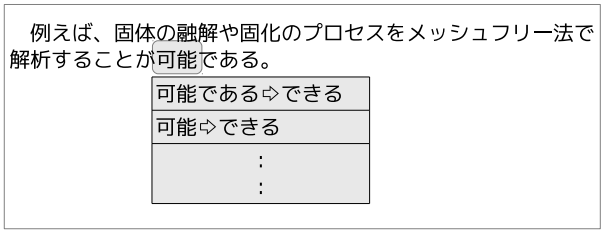

図 1: 修正のヒント出力の様子

3.2　研究の進め方

2 節の実験データとは別に冗長な文とその文を修正した文を準備する．2 節の実験データでは，「可能」「という」「すること」のみについてデータを作成したが，本節ではそれらの表現に限定せずにデータを作成する．作成したデータに基づき，修正先の候補が複数存在する冗長な文の修正例を収集する．収集したものを頻度により分析し，図 1 のアプリケーションの構築に役立つデータを構築する．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表 5: 差分部分の抽出結果

頻度	前方一致部分	差分部分	後方一致部分
11	地震の強い揺れ	によって(→で)	引き起こされる災害。
6	女性だけの劇団	である(→の)	ため、男性役も女性が演じる。
5	阿武隈山系北部の山あい	にある(→の)	美しい村を訪れた。
4	お酒に漬け込む	ことで(→と)	、生のときのクセが消えておいしく味わえるそうです。

表 6: 修正候補

修正表現	修正候補上位
によって	によって→で 場合によっては→Φ 施すことによって→施し
である	である→の である→Φ である→だ
にある	にある→の 立場にある→Φ にあることを→だと

3.3 使用データ

人手で冗長な文とそれを修正した文の対を 1,300 文対作成した．このデータを分析に利用する．

3.4 分析手法

データの分析は以下の手順で行う．

1. 使用データをそれぞれ形態素解析 ChaSen に適用し単語単位に分割をする．
2. 単語単位に分割したデータを文対ごとに差分を取り，差分箇所の表現の頻度をもとめる．差分検出には diff コマンドを使用する．一文中に差分箇所が複数存在する文については，それらの差分箇所を包含する最小の範囲を一つの差分箇所として扱って処理する．
3. 頻度の高い差分の表現についてその他の修正先の候補を収集する．

3.5 結果

データの分析結果を表 5 に示す．X(→ Y) は，冗長な文に含まれる冗長部 X を冗長でない表現 Y へ修正したことを表す．

一番頻度が高かったものは「によって(→で)」であった．「によって」およびそれに準じて頻度の高かった「である」「にある」についてその他の修正候補の一部を表 6 に示す．

ここで収集したデータは図 1 のアプリケーションの構築に役立つと思われる．収集したデータがどの程度役立つかの検討は今後の課題とする．

4 おわりに

本研究では冗長な文を修正する方法として，パターンを用いた手法と機械学習を用いた手法を提案した．「可能」「という」「すること」が原因となって冗長となった文の修正の実験を行った．パターンを用いる手法と機械学習を用いる手法のいずれかが，最も頻度の高いものを出力とするベースライン手法よりも同等以上の正解率であった．パターンを用いる手法と機械学習を用いる手法のいずれかで 6 割以上の正解率で冗長な文を修正できた．修正後の表現のみの推定 (修正前の表現の範囲を特定できなくてよい) では，パターンを用いる手法と機械学習を用いる手法のいずれかで 7 割以上の正解率を得た．以上により，「可能」「という」「すること」については，パターンを用いる手法と機械学習を用いる手法がある程度冗長な表現の修正に役立つことがわかった．

しかし，実際の文書の推敲での冗長な文の修正ではもっと確実な手法を用いる必要がある場合も考えられる．そのため修正をするのではなく，修正箇所の検出を自動で行い，さらに検出した冗長箇所の修正候補を頻度の高い順に並べ，ユーザーに提示するという方式を検討した．この方式では，冗長な個所とその修正候補が提示されるため，文書作成者の修正作業の負担が軽減されると思われる．この方式で必要となるデータの構築も試みた．

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 (2008). “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, 23 巻, 1 巻 , pp.25-32.
[2] Masaki Murata, Hitoshi Isahara(2002). “Automatic detection of mis-spelled Japanese expressions using a new method for automatic extraction of negative examples based on positive examples”, IEICE Transactions, VOL.E85-D, No.9, pp.1416-1424.
[3] 村田真樹, 井佐原均 (2004). “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.
[4] 内元清貴, 村田真樹, 馬青, 関根聡, 井佐原均 (2000). “コーパスからの語順の獲得”, 言語処理学会論文誌「自然言語処理」, Vol.7, No.4, pp.163-180.
[5] 村田真樹, 馬青, 井佐原均, 内元清貴 (1999). “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”, 言語処理学会論文誌「自然言語処理」, Vol.6, No.7, pp.61-73.
[6] 都藤 俊輔, 村田 真樹, 徳久 雅人, 馬 青 (2012). “冗長な文の機械的分析と機械的検出”, 第 18 回年次大会発表論文集, pp.1114-1117.
[7] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均 (2002), “SENSEVAL2J 辞書タスクでの CRL の取り組み”, 言語処理学会論文誌「自然言語処理」, Vol.10, No.3, pp.115-132.
[8] 村田真樹 (2002). “diff を用いた言語処理-便利な差分検出ツール mduff の利用”, 自然言語処理学会誌, 9 巻 , pp.91-110.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.