概要

文が読みにくい理由の一つに文が冗長であることが挙げられる [1]．文の生成や推敲の支援システムの構築には，冗長箇所を修正する技術が必要とされている．

本研究では，冗長な文を自動修正する方法を提案することを目的とする．文の改善の研究としては「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」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.「誤字の修正・適切な語の選択」では文献 [1, 2, 3] が,「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」では文献 [1, 4, 5] がある．しかし「冗長な表現の改善」を扱う研究についてはほとんどないため本研究で扱うこととした．本研究では，冗長な文の修正をするとともに，取り扱うデータの拡大し冗長な文章での検出を行う．

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用いる.「可能」「という」「すること」の存在が原因となって冗長となった文を修正する実験を行った．修正前と修正後の両方の表現を推定する場合，機械学習が関係する手法では，6 割の正解率を得た．修正後の表現のみを推定する場合においても，機械学習が関係する手法で，7 割の正解率を得た．冗長な文章での実験では，次の知見が得られた．分析により典型的な 3 種類の分類を獲得でき，また機械学習を用いる手法と，冗長度を用いる手法により冗長な文章を検出した．機械学習を用いた手法の正解率 (0.66) と，冗長度を用いる手法の正解率 (0.65) を得た．冗長度の有用性の確認として 1 文における冗長な文において，先行研究 [6] の機械学習に冗長度を素性に追加したところ性能向上が見られた．

目 次

第 1 章　はじめに　1

第 2 章　関連研究　3

2.1　冗長な文の検出 3

2.1.1　概要 3

2.1.2　知見 3

2.2　冗長な文についての知見 3

第 3 章　冗長な文の修正　5

3.1　問題設定 5

3.2　クロスバリデーション 5

3.3　データ 6

3.4　提案手法 7

3.4.1　手法 1:パターンを用いた手法 8

3.4.2　手法 2:機械学習を用いた手法 10

3.4.3　手法 3:ベースライン手法 12

3.4.4　手法 4:スタッキング手法 12

3.4.5　手法 5:最良選択手法 13

3.5　サポートベクターマシンとは 13

3.6　実験と結果 17

3.7　修正のヒント出力 19

第 4 章　冗長な文章に関する研究　22

4.1　冗長な文章に関する研究の流れ 22

4.2　冗長な文章の分析 22

4.2.1　使用データ 22

4.2.2　分析結果 24

4.3 冗長な文章の自動検出 25
4.3.1 提案手法 25
4.3.2 機械学習に基づく手法 25
4.3.3 冗長度に基づく手法 27
4.3.4 使用データ 28
4.3.5 実験 28
4.4 考察 31

第 5 章 冗長度の有効性の確認 32
5.1 提案手法 32
5.1.1 機械学習に基づく手法 32
5.1.2 冗長度に基づく手法 33
5.2 使用データ 33
5.2.1 使用データ 1(収集データ) 33
5.2.2 使用データ 2(作例データ) 35
5.3 結果 38
5.4 考察 40

第 6 章 おわりに 41

表 目 次

3.1 修正前表現と修正後表現の推定の正解率 (試行 1 回目) 17
3.2 修正前表現と修正後表現の推定の正解率 (試行 2 回目) 18
3.3 修正前表現と修正後表現の推定の正解率 (2CV の結果) 18
3.4 修正後表現の推定の正解率 (試行 1 回目) 18
3.5 修正後表現の推定の正解率 (試行 2 回目) 19
3.6 修正後表現の推定の正解率 19
3.7 正しく修正できた例 20
3.8 誤って修正した例 21

4.1 素性選択 (1 回目) 28
4.2 素性選択 (2 回目) 29
4.3 素性選択 (3 回目) 29
4.4 素性選択 (4 回目) 29
4.5 全素性を利用した場合の結果 29
4.6 冗長度に基づく手法の閾値調整 30
4.7 機械学習と冗長度に基づく冗長な文章の検出 30

5.1 同義・類義な語が重複した表現の例 35
5.2 簡潔なものへの言い換えができる表現の例 35
5.3 機械学習による検出結果 (収集データ) 38
5.4 冗長度による検出結果 (収集データ) 39
5.5 機械学習による検出結果 (作例データ) 39
5.6 冗長度による検出結果 (作例データ) 39

図 目 次

3.1 n 分割クロスバリデーション 6
3.2 データ作成 7
3.3 形態素解析の例 8
3.4 冗長箇所の例 9
3.5 SVM のプロセス 10
3.6 使用素性 11
3.7 最良選択手法 14
3.8 マージンの最大化 15
3.9 修正のヒント出力の様子 21

4.1 冗長な文章の作例データ例 23

5.1 データベース作成 34
5.2 収集データの例 36
5.3 作例データの例 37

第1章　はじめに

文が読みにくい理由の一つに文が冗長であることが挙げられる [1]．文の生成や推敲の支援システムの構築には，冗長箇所を修正する技術が必要とされている．

例文「まず初めにマシンの点検を行う。」を考えてみよう．文中の「まず」と「初め」は同じ意味を含む単語であり冗長性がある.「点検を行う」は用言性名詞「点検」が「行う」の意味を含むため冗長性がある．ここで冗長箇所を修正すると,「まずマシンを点検する。」という簡潔な文となる．この文は，例文と同じ意味を持ち，かつ，例文より簡潔な文である．本研究では，同じ意味のより簡潔な文に修正できる文を冗長な文とし研究を進める．

文の改善の研究としては「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」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.「誤字の修正・適切な語の選択」では文献 [1, 2, 3] が,「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」では文献 [1, 4, 5] がある．しかし「冗長な表現の改善」を扱う研究についてはほとんどないため本研究で扱う．

「冗長な表現の改善」には,「1．冗長な文の分析」「2．冗長な文の検出」「3．冗長な文の修正」の 3 つのプロセスが考えられる．対象となる冗長な文は大きく,「一文における冗長な文」と「冗長な文章（複数の文にまたがる冗長な文)」に分けられる．都藤らは [6]，その手始めとして,「一文における冗長な文」における「冗長な文の分析」と「冗長な文の検出」を行っている．本研究では,「一文内における冗長な文」における「冗長な文の修正」をするとともに，取り扱うデータを拡大し「冗長な文章」で「冗長な文の分析」と「冗長な文の検出」を行うことを目的とする．

この目的を達成するため，本論文では，以下の研究を行う．

1. 冗長な文の修正を行う (第 4 章)．冗長性の高いとされる「可能」「という」「すること」の表現が入った文を対象として，表現ごとに修正を行う．冗長な文とその修正文の対を照合し，修正用の文パターンや規則を取得する．これら文パターンや規則

を利用して，冗長な文の修正案を提示する技術を構築する．

2. 検出するデータ規模を冗長な文章 (複数文に渡る文) に拡大し研究する (第 5 章)．冗長な文章をウェブ上から収集，作例し修正の際と同様にパターンや規則の獲得をする．それらを利用し．文章の冗長性を判定する技術を構築する．

3. 本研究で提案する冗長度の有用性を確認する (第 6 章)．先行研究 [6] のデータセットを用い，1 文における冗長な文に役立つかを実験する．先行研究 [6] で 1 文における冗長な文の検出を行っていたが，冗長度は利用していなかった．

本研究の主な主張点は以下である．

- 一文における冗長な文の修正において，提案手法で，7 割の正解率を得た．
- 冗長な文章の分析により冗長な文章の典型的な 3 つの分類を示した．また，検出において本論文で提案する冗長度が役立つことを確認した．
- 冗長な文章において有用であった冗長度の検証として，一文における冗長な文の検出において冗長度が役立つことを確認した．

第2章　関連研究

本章は冗長な文の検出・修正に関係のある研究を紹介する．

2.1　冗長な文の検出

先行研究である都藤らの [6] 研究について述べる．

2.1.1　概要

都藤ら [6] は一文における冗長な文を対象として，冗長な文を分析する方法，機械学習を用いて自動的に検出をする方法を提案している．

2.1.2　知見

冗長な文を分析した結果「可能」や「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という知見を得ている．すべての文に対して 1 個の機械学習を利用して冗長な文の判定を行う手法で，0.52 の適合率を得てベースライン (すべてを冗長な文と判定する方法) を上回ったが，F 値ではベースラインより劣っていた．そこで特定の表現ごとに機械学習を行って冗長な文を検出する手法を利用した．この手法では,「可能」「という」「すること」の表現において 0.7 から 0.8 という比較的高い F 値で検出できている．この結果はベースラインの性能を上回った．この研究では,「可能」「という」「すること」の表現でしか実験していないが，同様の処理を行うことでこれら以外の表現についても冗長な表現の検出が期待できる．

2.2　冗長な文についての知見

要約等の研究から参考とした知見について述べる．

- 大竹ら [7] は新聞の関連複数記事を 1 つの文書へと要約するための手法を提案している．各記事の第一段落を用いて，その重複部・冗長部を削除することにより複数の関連記事をどの程度要約できるかを明らかにした．この研究での要約の手法は，

  1. 各記事の第一段落を時系列に並べる．
  2. 推量文を文末表現で特定し削除する.
  3. 詳細な住所の表現も冗長部として削除する.
  4. 記事の前提条件等を導入部とし，導入部の名詞と動詞がそれ以前の記事の文に含まれる場合削除する.
  5. 頻繁に出現する人名・地名に関してそれぞれ記事内と記事間における重複を調べ重複している部分は 1 つを残し, 他は削除する.

  以上の処理で最終的に残った文章が要約文章となる．この研究は文書要約であるが，本研究の冗長な文の判定基準の作成で参考にした．

- 原口ら [9] は開発関連文書の品質を向上させるために校正基準を定義し文書表現の記述不備を検出した．そこで開発した手法を目視による品質調査と比較を行い検出に要する時間を短縮し，検出性能も高くすることができた．この研究についても，本研究の冗長な表現の判定基準の作成で参考にした．

- 村田らは，誤った日本語文を抽出する技術 [2]，適切な英語表現に変換する文パターンを抽出する技術 [3]，語順を推定する技術 [4]，係り受けの複雑さを計量する技術 [5] を構築し誤字の修正・適切な語の選択と語順の修正・語と語の係り受けの誤り及び複雑性の修正を行った．

第3章　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，本章では，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，本章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．冗長な文の修正は文作成者の支援や，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．

3.1　問題設定

「可能」「という」「すること」の表現が入った文はを対象としてこれらの表現を含む文について冗長な文の修正を行う．また本章は，冗長な文の修正のみに重点をおくため，入力データには，冗長性を修正できる文を入力として用いる．入力データとして 3.3 節で作成する文対のうち，冗長な文 100 文を用いる．正解データとして 3.3 節で作成する文対のうち，修正文 100 文を用いる．入力データを正解データのように修正できればよいとする．評価には 2 分割のクロスバリデーションを用いる．各手法は、クロスバリデーションにおける学習データとして入力データと正解データの両方を用いることが、テストデータには入力データのみを用いることができるものとする。

3.2　クロスバリデーション

n 分割クロスバリデーションでは、標本群を n 個に分割し，そのうちの 1 つをテストデータとし，残る $n-1$ 個を訓練事例とする．そして n 個に分割された標本群それぞれをテスト事例として n 回推定を行う．そうして得られた n 回の結果を組み合わせてテスト事例全体の推定結果を得る．

クロスバリデーションの例を図 3.1 に示す．

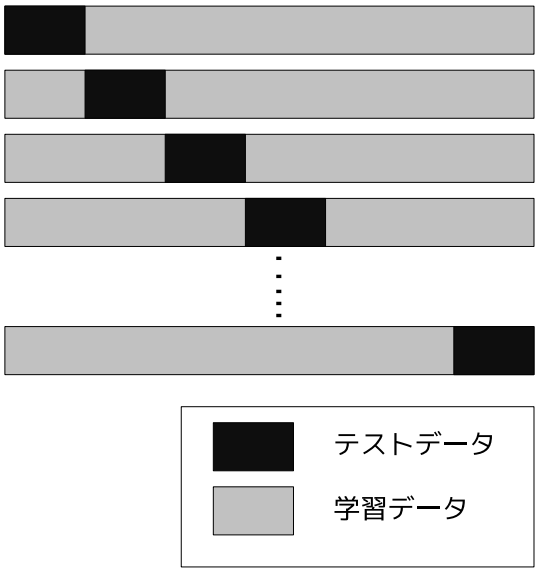

図 3.1: n 分割クロスバリデーション

3.3 データ

本章ではウィキペディア[1]の日本語ページ，2013 年 10 月 30 日のデータベースを利用する．手順を以下に示す．

1. ウィキペディアおいて,「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．

2. 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．

3. 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,「十分理解可能である。」の文からは次のような対を獲得する．

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

文対例 1

冗長な文　十分理解可能である。

修正文　　十分理解できる。

4. 「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．

図 3.2 はデータ作成の一連の流れである．

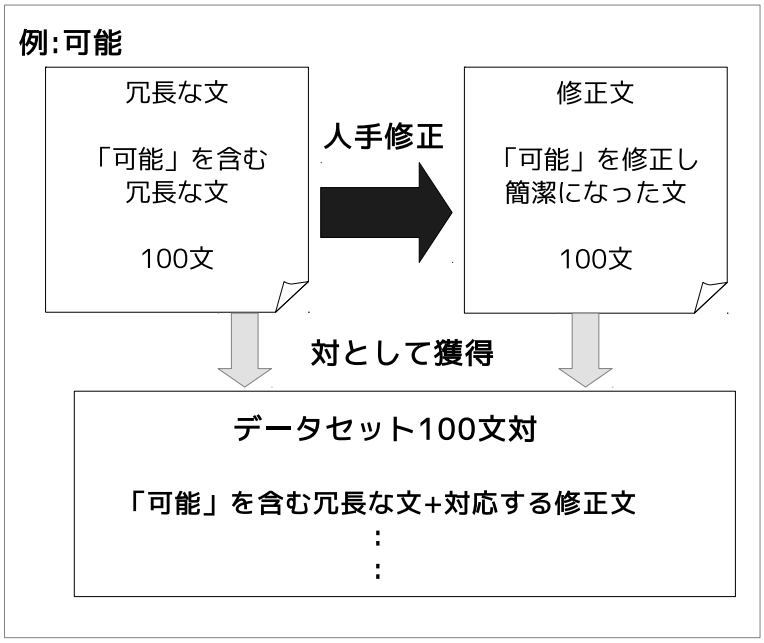

図 3.2: データ作成

3.4　提案手法

以下で述べる手法は「可能」「という」「すること」のそれぞれの表現ごとに行う．冗長な文の修正には以下の 5 手法を用いる．

- 手法 1:パターンを用いた手法
- 手法 2:機械学習を用いた手法

- 手法 3:ベースライン手法
- 手法 4:スタッキング手法
- 手法 5:最良選択手法

機械学習には，サポートベクトルマシン (SVM) を用いる．SVM の詳細は 3.5 節に述べる．

3.4.1 手法 1:パターンを用いた手法

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を次の手順で行う．

1. 「可能」について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．例えば「十分理解可能である。」という文を形態素解析にかけると図 3.3 に示した結果となる．

入力文：十分理解可能である。

```
十分    ジュウブン      十分    名詞-形容動詞語幹
理解    リカイ          理解    名詞-サ変接続
可能    カノウ          可能    名詞-形容動詞語幹
で      デ              だ      助動詞      特殊・ダ      連用形
ある    アル            ある    助動詞    五段・ラ行アル    基本形
。      。              。      記号-句点
EOS
```

図 3.3: 形態素解析の例

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 [11] を使用する．

   例えば「点検を行う」を「点検する」に修正していた場合を考えてみる．

   図 3.4 の下線部分「を行う」が「する」に修正されている．本研究ではこの「する」と修正された「を行う」が差分として検出される．

[2]ChaSen:http://ChaSen-legacy.sourceforge.jp/

入力文：　十分理解可能である。

十分	ジュウブン	十分	名詞
理解	リカイ	理解	名詞
可能	カノウ	可能	名詞
で	デ	だ	助動詞
ある	アル	ある	助動詞
。	。	。	記号-句点
EOS			
十分	ジュウブン	十分	名詞
理解	リカイ	理解	名詞
理解	リカイ	理解	名詞
できる	デキル	できる	動詞
。	。	。	記号-句点
EOS			

図 3.4: 冗長箇所の例

3. 獲得した差分部分から，パターンを作成する．例えば，先に示した文対例 1 からは次のようなパターンを獲得する．

パターン例

・パターン　可能である → できる

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．

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．

3.4.2 手法 2:機械学習を用いた手法

手法 2 では教師あり機械学習 [10] を用いて冗長な文の修正を行う．機械学習法には，サポートベクターマシン法 (以下 SVM) を用いる．(サポートベクターマシンの説明は 3.5 節で述べる）本研究では 1 次のカーネル関数を用いる．

例えば今回の機械学習は図 3.5 の流れである．

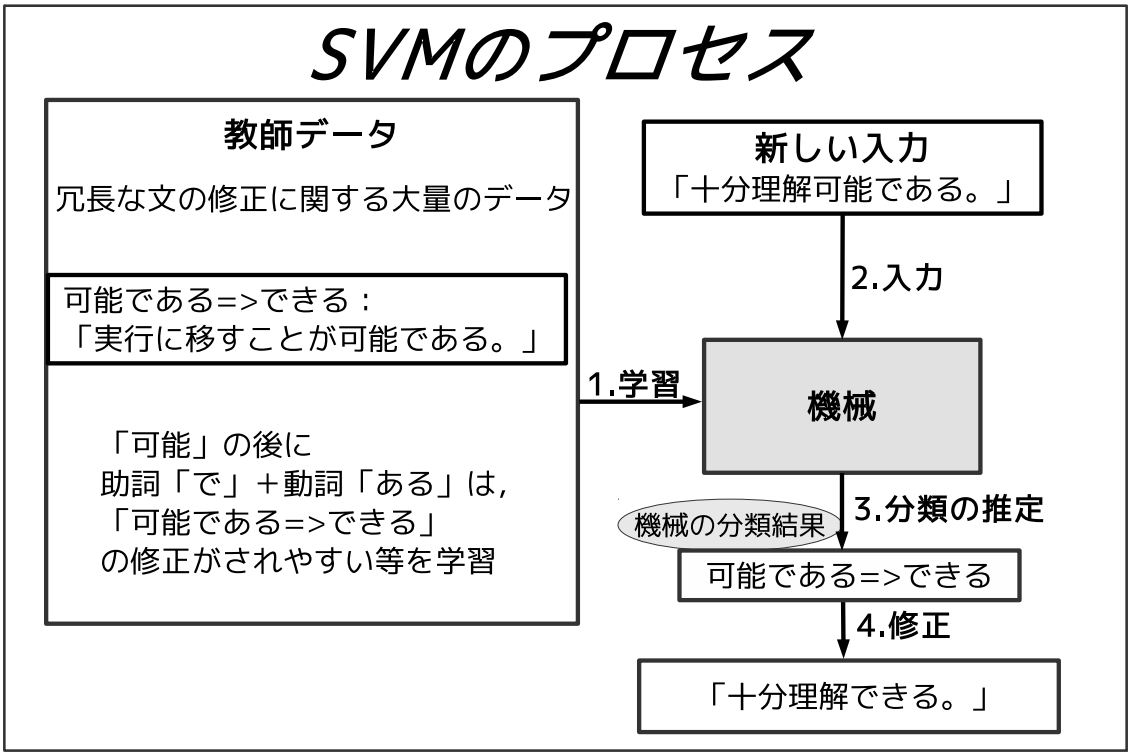

図 3.5: SVM のプロセス

「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」について学習で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 [11] 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

4. 学習データを用いて入力文から上述の分類先を推定できるように機械学習を行う．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，図 3.6 に示す素性を用いる．

○**素性番号 1(単語)** 出現単語のうち，文中に存在する対象表現 X の前後 2 単語．複数の品詞の種類がある単語を区別するため，各単語の出現形に品詞の情報を組み合わせて用いる素性である.「。」や「、」も含む．対象表現 X の前後 2 単語が存在しない場合は，前後単語がないという情報を用いる．素性の例は「名詞:日本」や「助詞:に」,「句点:。」である．

○**素性番号 2(品詞)** 文中に存在する対象表現 X の前後 2 単語の品詞．素性の例は「名詞」「動詞」等である．

図 3.6: 使用素性

上記素性の対象表現 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．次に入力文に対して, 実際に付与される素性を大まかに示す.

入力文:「迅速な対応が可能となる。」

素性番号 1:付与素性例 2 単語前単語+対応，1 単語前単語+が，1 単語後ろ単語+と，2 単語後ろ単語+なる

素性番号 2:付与素性例 2 単語前品詞+名詞，1 単語前品詞+格助詞，1 単語後ろ品詞+格助詞，2 単語後ろ品詞+動詞

上記の付与素性の例では，“+” の前の表現は素性の種類を示す記号であり，“+” の後ろの表現はその素性が持つ情報である．“+” の前の接頭語は対照表現との位置関係を表している．対象表現の 2 単語前の単語素性である場合は「2 単語前単語」という接頭語が付与され， 対象表現の 2 単語前の品詞素性である場合は「2 単語前品詞」という接頭語が付与される．

3.4.3 手法 3:ベースライン手法

比較のため，学習データに出現する中で最も頻度の高い分類を分類先とする手法をベースラインとして用いる．

分類先の作成に以下の BL1 と BL2 の方法を取る．

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

3.4.4 手法 4:スタッキング手法

スタッキングとは，他の手法の解析結果を素性に追加して用いる方法である．利用する手法それぞれの利点を引き出すために，村田 [8] らの手法を参考に手法 4 として手法 1 から手法 3 の出力結果を，手法 2 の素性に追加して用いる．各出力結果を素性として用いることで，機械学習の分類性能を高める狙いがある．

例えば，ある教師データが素性として {a,b,c} を持っており，手法の結果が “d” の場合を考えると，素性 {a，b，c，手法 1 の出力結果:d} を新しい素性として用い，機械学習する．

分類先の作成に以下の 2 つの方法を取る．

ST1 分類先に ML1 の分類を用い，素性として図 3.6 の素性と PT1，PT2，ML1，BL1 の出力結果を用いる

ST2 分類先に ML2 の分類を用い，素性として図 3.6 の素性と PT1，PT2，ML2，BL2 の出力結果を用いる

3.4.5 手法 5:最良選択手法

手法 1 から手法 4 のうち，2 分割クロスバリデーションにおける他方の分割データにおいて最も良かった手法をテストデータで用いる．

図 3.7 は最良選択手法の一連の流れである．

1. 標本群を 2 個に分割し，一つをデータ A，残りをデータ B とする．
2. データ A を学習データとし，データ B をテストデータとして用いるデータセット 1 を定義する．
3. データ B を学習データとし，データ A をテストデータとして用いるデータセット 2 を定義する．
4. 各データセットを手法 1 から手法 4 で解く．
5. 4 の結果，各データセットで最も良い結果を出した手法を用いて他方のデータセットを解く．

3.5 サポートベクターマシンとは

サポートベクトルマシン法は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負例のマージン（間隔）を大きくとるほど分類器の誤りが減少するという考えから，このマージンを最大にする超平面を求めそれを用いて分類を行なう．一般的に上記の方法の他に,「ソフトマージン」と呼ばれる学習データにお

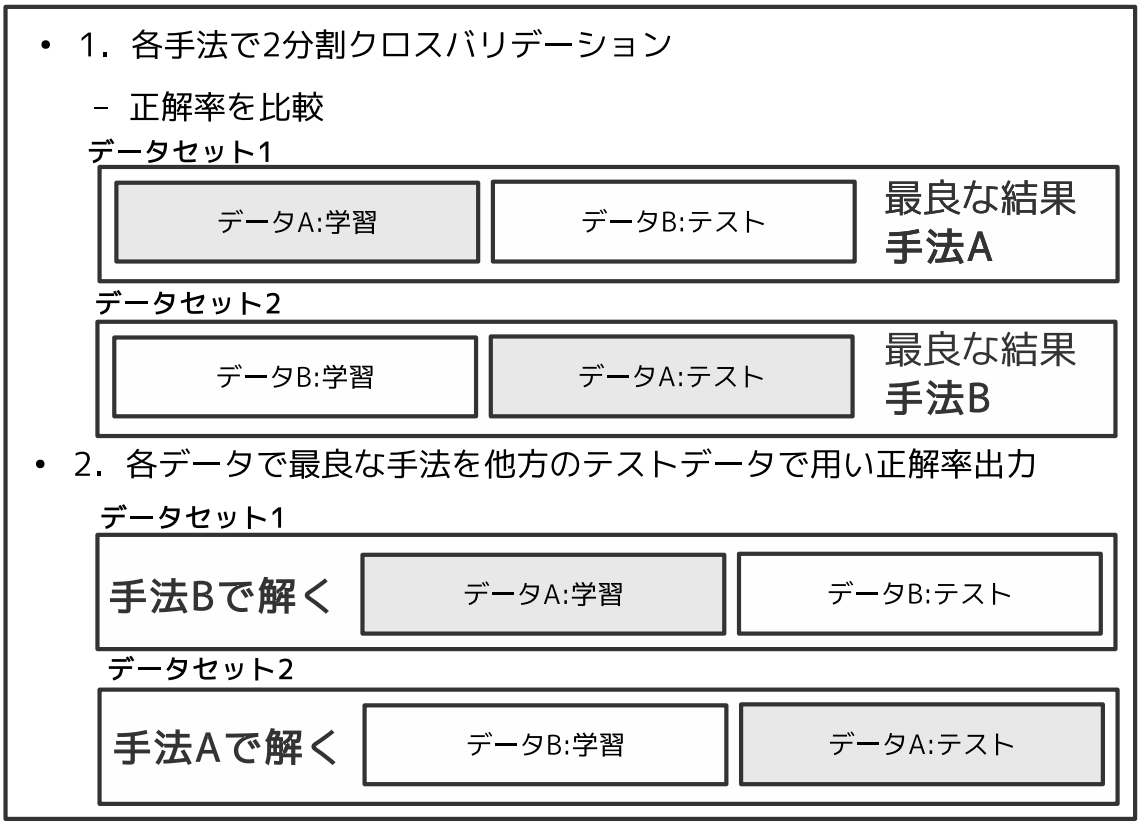
• 1．各手法で2分割クロスバリデーション
– 正解率を比較
データセット1
データA:学習
データB:テスト
最良な結果
手法A
データセット2
データB:学習
データA:テスト
最良な結果
手法B
• 2．各データで最良な手法を他方のテストデータで用い正解率出力
データセット1
手法Bで解く
データA:学習
データB:テスト
データセット2
手法Aで解く
データB:学習
データA:テスト

図 3.7: 最良選択手法

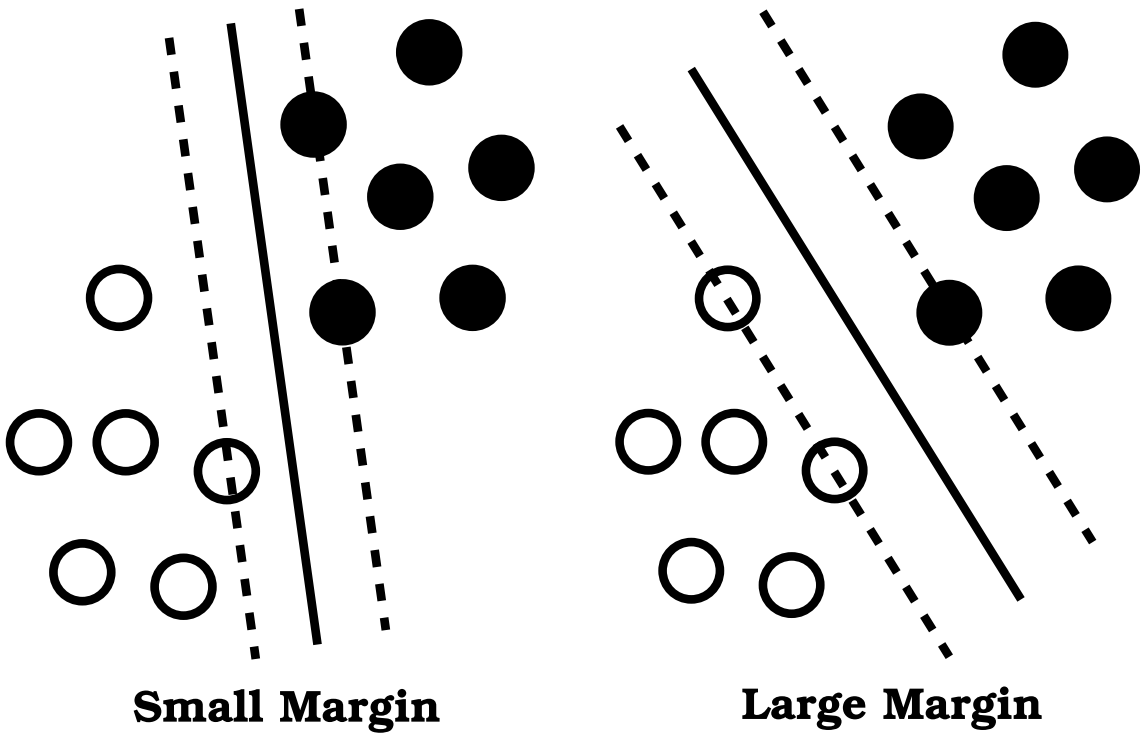

図 3.8: マージンの最大化

いてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や，線形分離が不可能な問題に対応するために，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することが可能である．

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \\
b &= -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2} \\
b_i &= \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}
\tag{3.1}
$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i (i = 1, ..., l, y_i \in \{1, -1\})$ は学習データの文脈と分類先を意味し，関数 sgn は，

$$
\begin{aligned}
sgn(x) = \quad & 1 \quad (x \geq 0) \\
& -1 \quad (otherwise)
\end{aligned}
\tag{3.2}
$$

であり，また，各 α_i は式 (3.4) と式 (3.5) の制約のもと式 (3.3) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) \quad = \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{3.3}$$

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{3.4}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.5}$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本稿では以下の多項式のものを用いる．

$$K(\mathbf{x}, \mathbf{y}) \quad = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.6}$$

C, d は実験的に設定される定数である．本稿ではすべての実験を通して C を 1 に d を 2 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (3.1) の和をとっている部分はこの事例のみを用いて計算される．

3.6 実験と結果

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)，スタッキング手法 (ST1 と ST2) による冗長な文の修正結果を示す．

ベースラインで用いられた各表現ごとの修正前と修正後の表現対 (最頻の表現対) は以下である．

表現	修正前表現	修正後表現
「可能」を含む文	可能である	できる
「という」を含む文	という	(Φ)
「すること」を含む文	することが	(Φ)

ベースラインで用いられた各表現ごとの修正後のみの表現対 (最頻の表現対) は以下である．

表現	修正後表現
「可能」を含む文	できる
「という」を含む文	(Φ)
「すること」を含む文	(Φ)

「という」「すること」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．

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正解率を表 3.1 から表 3.3 に示す．表 3.1 は 2 分割クロスバリデーション (2CV) における 1 回目の試行結果を示し，表 3.2 は 2CV における 2 回目の試行結果を示している．表 3.1 と表 3.2 を総合した結果が表 3.3 となる．

表 3.1: 修正前表現と修正後表現の推定の正解率 (試行 1 回目)

手法	可能	という	すること	合計
PT1	0.64(32/50)	0.50(25/50)	0.32(16/50)	0.49(73/150)
PT2	0.50(25/50)	0.64(32/50)	0.56(28/50)	0.57(85/150)
ML1	0.58(29/50)	0.70(35/50)	0.58(29/50)	0.62(93/150)
BL1	0.26(13/50)	0.68(34/50)	0.32(16/50)	0.42(63/150)
ST1	0.60(30/50)	0.82(41/50)	0.58(29/50)	0.67(100/150)

表 3.2: 修正前表現と修正後表現の推定の正解率 (試行 2 回目)

手法	可能	という	すること	合計
PT1	0.64(32/50)	0.62(31/50)	0.32(16/50)	0.53(79/150)
PT2	0.66(33/50)	0.68(34/50)	0.54(27/50)	0.63(94/150)
ML1	0.58(29/50)	0.80(40/50)	0.54(27/50)	0.64(96/150)
BL1	0.26(13/50)	0.64(32/50)	0.34(17/50)	0.41(62/150)
ST1	0.52(26/50)	0.70(35/50)	0.60(30/50)	0.61(91/150)

表 3.3: 修正前表現と修正後表現の推定の正解率

手法	可能	という	すること	3 表現すべて
PT1	0.64(64/100)	0.56(56/100)	0.32(32/100)	0.51(152/300)
PT2	0.58(58/100)	0.66(66/100)	0.55(55/100)	0.60(179/300)
ML1	0.56(56/100)	0.75(75/100)	0.56(56/100)	0.62(187/300)
BL1	0.26(26/100)	0.66(66/100)	0.33(33/100)	0.42(125/300)
ST1	0.56(56/100)	0.76(76/100)	0.59(59/100)	0.64(191/300)
最良 1	0.57(57/100)	0.70(70/100)	0.59(59/100)	0.62(186/300)

修正後の表現さえ推定できただけで正解とする場合の正解率を表 3.4 から表 3.6 に示す．表 3.4 は 2CV における 1 回目の試行結果を示し，表 3.5 は 2CV における 2 回目の試行結果を示している．表 3.4 と表 3.5 を総合した結果が表 3.6 となる．

表 3.4: 修正後表現の推定の正解率 (試行 1 回目)

手法	可能	という	すること	合計
PT1	0.74(37/50)	0.78(39/50)	0.34(17/50)	0.62(93/150)
PT2	0.52(26/50)	0.78(39/50)	0.58(29/50)	0.63(94/150)
ML2	0.66(33/50)	0.86(43/50)	0.66(33/50)	0.73(109/150)
BL2	0.64(32/50)	0.82(41/50)	0.44(22/50)	0.63(95/150)
ST2	0.68(34/50)	0.88(44/50)	0.64(32/50)	0.73(110/150)

表 3.3 と表 3.6 において，3 つの表現すべてを合わせたものの正解率では，機械学習が関係している ML1，ML2，ST1，ST2，最良 1，最良 2 が他手法に比べて良く，これらの手法は，表 3.3 では，約 6 割の正解率，表 3.6 では，約 7 割の正解率を得た．修正前と修正後の両方の表現を推定する場合，機械学習が関係する手法では，6 割の正解率を得た．修正後の表現のみを推定する場合においても，機械学習が関係する手法で，7 割の正解率を得た．

表 3.7 に正しく修正できた文の例を，表 3.8 に誤って修正した文の例を示す．

表 3.5: 修正後表現の推定の正解率 (試行 2 回目)

手法	可能	という	すること	合計
PT1	0.72(36/50)	0.82(41/50)	0.32(16/50)	0.62(93/150)
PT2	0.68(34/50)	0.82(41/50)	0.56(28/50)	0.67(103/150)
ML2	0.60(30/50)	0.78(39/50)	0.66(33/50)	0.68(102/150)
BL2	0.70(35/50)	0.78(39/50)	0.42(21/50)	0.63(95/150)
ST2	0.62(31/50)	0.76(38/50)	0.62(31/50)	0.67(100/150)

表 3.6: 修正後表現の推定の正解率

手法	可能	という	すること	3 表現すべて
PT1	0.73(73/100)	0.80(80/100)	0.33(33/100)	0.62(186/300)
PT2	0.60(60/100)	0.80(80/100)	0.57(57/100)	0.66(197/300)
ML2	0.65(65/100)	0.81(81/100)	0.65(65/100)	0.70(211/300)
BL2	0.67(67/100)	0.80(80/100)	0.43(43/100)	0.63(190/300)
ST2	0.65(65/100)	0.82(82/100)	0.63(63/100)	0.70(210/300)
最良 2	0.73(73/100)	0.76(76/100)	0.65(65/100)	0.72(216/300)

3.7 修正のヒント出力

「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より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 3.9 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．また頻度だけでなく，機械学習で推定した結果には，出力候補に対してそれが正解であるという確率や確信度を同時に算出できるので，算出結果をもとに値の大きい候補を順に表示させることも考えられる．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．

表 3.7: 正しく修正できた例

表現	出力文
可能	しかし、この考え方は現実的にも 適応 可能である。→ (適応できる) に修正
	理論上、感度と特異度は独立しており、共に 100 %を達成することも 可能である。→ (できる) に修正
	無泥水・無排土での 施工が 可能であり、経済的である。→ (施工でき) に修正
という	つまり動く物体の長さは縮んで 計測される という ことが分かる。→ (計測されること) に修正
	また、アジア側では赤道の北に片寄り、オセアニアからアメリカの間では赤道の南に 片寄る という 特徴が見られる。→ (片寄る特徴) に修正
	固定されていると言っても、必ずしもその値が具体的に特定されている必要はなく、特定の値をとることが決まっている という 定数の特徴 である。→ (決まっている定数の特徴) に修正
すること	以上より、問題の積分を 計算 することが できた。→ (計算できた) に修正
	所有権は、何ら人為的拘束を受けず、侵害するあらゆる他人に対して 主張 することが できる 完全な支配権であり、国家の法よりも先に存在する権利で神聖不可侵であるとする原則。→ (主張できる) に修正
	しかし、その存在は全能であるから、その存在は後からいつでも、持ち上げられる程度に石を 軽く することが できる。→ (軽くできる) に修正

表 3.8: 誤って修正した例

表現	出力文	正解文
可能	炭素源による分類も明確な区別が可能だが、混合栄養は二酸化炭素と有機物の両方を炭素源とするという特異な分類もなされる。下線部を→ (できる) に修正	炭素源による分類も明確に区別できるが、混合栄養は二酸化炭素と有機物の両方を炭素源とするという特異な分類もなされる。
という	時間計算量と空間計算量という二つの観点がある。下線部を→ (の) に修正	時間計算量と空間計算量の観点がある。
すること	上記のとおり、準拠法指定に関する立法上・解釈上の指針は、問題となる私法的法律関係に関する最密接地の法を選ぶ点にあり、そのためには、そのような地を指定することが可能となる要素を媒介とする必要がある。下線部を→削除	上記のとおり、準拠法指定に関する立法上･解釈上の指針は、問題となる私法的法律関係に関する最密接地の法を選ぶ点にあり、そのためには、そのような地を指定できる要素を媒介とする必要がある。

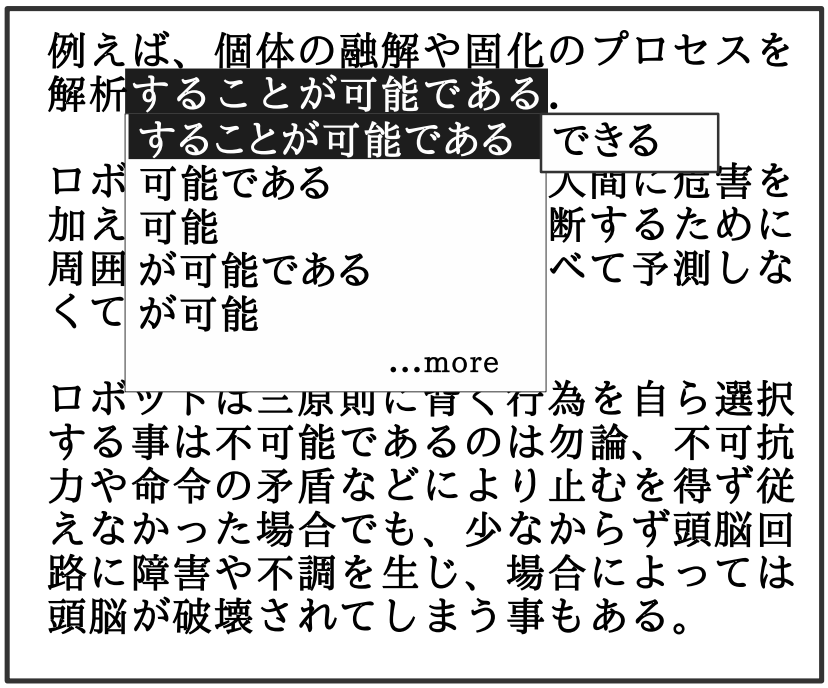

図 3.9: 修正のヒント出力の様子

第4章　冗長な文章に関する研究

先行研究 [6] では 1 文における冗長な文の分析・検出を行っている．複数文にまたがる冗長な文章に関しては行われていないため，本章で取り扱う．冗長な文章に関する研究は文章を作成する者の推敲作業を手助けするシステムの構築に役立つ．

4.1　冗長な文章に関する研究の流れ

まず，複数の文にまたがった冗長な文章に関わるデータベースを作成し，そのデータベースを分析し，どのような冗長な文章が存在するかを調査する (4.2 節)．

次に，複数の文にまたがった冗長な文章を自動検出する研究を行う (4.3 節)．検出には，教師あり機械学習や冗長度を利用する．冗長度は同じ語を多く使うほど大きな値になるものであり，この値が大きいと冗長な文章と判断するものである．

4.2　冗長な文章の分析

複数文にまたがった冗長な文章としてどのようなものがあるかの調査を行う．

4.2.1　使用データ

複数文にまたがった冗長な文章を作例する．その文章を冗長でないように適切に修正した文章も作成する．冗長な文章と修正した冗長でない文章の対のデータを格納したデータベースを作成する．このデータベースでは，冗長な文章か修正された文章のいずれかは 2 文以上からなる文章である．データの作成は，文章作成を得意とするものが行う．データは図 4.1 に示す．

作成したデータベースは，冗長な文章が 500 個，修正により作成された文章が 500 個であり，合計 1,000 個の文章である．作成したデータベースをランダムに 2 分割し，それらをデータ A(学習およびチューニング用) とデータ B(評価用) とした．

冗長な文 **1** 今日はこの間買ったばかりの新しい靴を履いていく。この靴はデパートで買ったお気に入りだ。
修正文　**1** 今日はこの間デパートで買ったばかりのお気に入りのを履いていく。

冗長な文 **2** 音楽には様式があり、それを「ジャンル」と呼んでいる。「民族音楽」「クラシック音楽」「ジャズ」「ロック」などといった名称で呼ばれているのがそれである。
修正文　**2** 音楽には「民族音楽」「クラシック音楽」「ジャズ」「ロック」など「ジャンル」と呼ばれる様式がある。

冗長な文 **3** 脳内の視床下部に、2 つの食欲中枢が存在する。1 つは視床下部腹内側核で、満腹中枢といわれる。もう一つは外側視床下野で摂食中枢といわれる。
修正文　**3** 脳内の視床下部には、満腹中枢といわれる視床下部腹内側核と摂食中枢といわれる外側視床下野という 2 つの食欲中枢がある。

冗長な文 **4** エアトレインは、ニューヨーク・ニュージャージー港湾公社と契約を執り行うボンバルディア・トランスポーテーションが運営している。この会社は、空港やエアトレイン・ニューアークも運営している。
修正文　**4** ニューヨーク・ニュージャージー港湾公社と契約を執り行うボンバルディア・トランスポーテーションは、エアトレイン、空港やエアトレイン・ニューアークも運営している。

冗長な文 **5** ぼくは酒をやらないので、この酒の口あたりのよいという味はまったくわからないが、あるときの「玄月會」で武田君にその夜の酒を選んでもらったとき、この銘柄を推薦してくれた。
修正文　**5** ぼくは酒をやらないので、この酒の口あたりのよいという味はまったくわからない。だが、「玄月會」で武田君にその夜の酒を選んでもらったとき、この銘柄を推薦してくれた。

冗長な文 **6** 近所に住む憧れの女の子 S ちゃんや、いじめっ子だが根は優しい S 夫や T 志などの友人達も交えた日常の中で、N 太は道具に頼りがちになりながらも反省し学んでいき、彼が歩んでゆく未来は少しずつより良い方向へと変わってゆく。
修正文　**6** 近所に住む憧れの女の子 S ちゃんや、いじめっ子だが根は優しい S 夫や T 志などの友人達も交えた日常の中で、N 太は道具に頼りながらも反省し学んでいく。そして、彼が歩む未来は少しずつより良い方向へと変わってゆく。

図 4.1: 冗長な文章の作例データ例

4.2.2　分析結果

前節で作成したデータを人手で分析した．データ中の冗長な文章と修正された文章の対には，以下の 3 つの分類があった．以下では，冗長な文章とそれを修正した文章の例文の対も示している．

分類 1:文単位の修正

文章を構成する各文が冗長な場合である．文ごとに修正される．

冗長な文章　　私がネット上で議論をしないもうひとつの理由は「あまりに議論効率が悪いから」です。実際に会って話し合う議論効率を 100 とすると、ネット上で議論する効率は 10 以下というのが私の印象です。

修正した文章　私がネット上で議論をしないもうひとつの理由は「あまりに議論効率が悪いから」です。私の印象では、実際に会って話し合う議論効率の 10 分の 1 以下です。

分類 2:補足文の併合による修正

この分類の冗長な文章は，先頭の文中に出現する単語を，後続する文が補足または説明をするものである．補足または説明をする文を先頭文にまとめる形で，短く簡潔な文章に修正される．

冗長な文章　　今日はこの間買ったばかりの新しい靴を履いていく。この靴はデパートで買ったお気に入りだ。

修正した文章　今日はこの間デパートで買ったばかりのお気に入りのを履いていく。

分類 3:長い文の箇条書きへの修正

この分類の冗長な文章は，雑多に書かれている長い文である．箇条書きにまとめる形で修正される．

冗長な文章　　厚化粧は、自然の肌色より大幅に明るい色のファンデーションを塗るベースリッチ型と、濃い色のアイシャドーを広い範囲に塗ったり濃い色のほほ紅、口紅を塗ったりするポイントリッチ型に分類される。

修正した文章　厚化粧は以下のように分類される。

- 自然の肌色より明るい色のファンデを塗るベースリッチ型
- 濃い色のシャドー、濃い色のほほ紅や、口紅を塗るポイントリッチ型

4.3　冗長な文章の自動検出

複数の文にまたがる冗長な文章の自動検出を試みる．

4.3.1　提案手法

提案手法には，機械学習に基づく手法と冗長度に基づく手法の 2 種類がある．

4.3.2　機械学習に基づく手法

冗長な文章と，冗長な文章を修正した文章の 2 分類のデータに対して，入力データが冗長な文章であるか，否かの 2 値分類を機械学習で行い，冗長な文章を自動検出する．機械学習法には，サポートベクターマシン法を用いる．機械学習の素性には以下を用いる．

○**素性番号 1(単語)** 文内の出現単語とその品詞．形態素解析器 ChaSen を用いて単語の情報を取得する．複数の品詞の種類がある単語を区別するため，各単語の出現形に品詞の情報を組み合わせて用いる素性である．「。」や「、」も含む．素性の例は，「名詞:日本」や「助詞:に」，「句点:。」である．

○**素性番号 2(品詞)** 文内の出現品詞．素性の例は「名詞」「動詞」である．

○**素性番号 3(冗長度)** 次式でもとめた冗長度のランク．

$$\text{冗長度 } x = \frac{N}{V}[V : \text{単語の異なり数}, N : \text{延べ単語数}] \tag{4.1}$$

最小は 1 で値が大きくなるほど冗長と考える．文ごとに素性の重なりができるように，冗長度 x を 0.1 ごとに 5 段階にランク分けして用いる．

ランク 1　$1.0 \leqq x < 1.1$
ランク 2　$1.1 \leqq x < 1.2$
ランク 3　$1.2 \leqq x < 1.3$
ランク 4　$1.3 \leqq x < 1.4$
ランク 5　$1.4 \leqq x$

○**素性番号 4(2 単語連続)** 文内に出現する 2 単語連続．文内に出現する単語を 2 単語ごとにつなげた素性である．

○**素性番号 5(2 単語連続の品詞連続)** 文内に出現する 2 単語連続の品詞連続．素性番号 4 を品詞で行った素性である．

○**素性番号 6(句点の数)** 文内に出現する句点の数．

○**素性番号 7(読点の数)** 文内に出現する読点の数．

○**素性番号 8(文長)** 文内の文字数 (句読点もカウントする)．文ごとに素性の重なりができるように，文長の値を 10 ごとに区切って素性を作成する．例えば，文字数 49 の場合「文長:40」，文字数 50 の場合「文長:50」という素性とする．

次に入力文に対して, 実際に付与される素性を大まかに示す．

入力文:「問題は、チャンスはいつ転がり込むかわからないということ。チャンスは突然にやってくる。」

素性番号 1:付与素性例 名詞+問題，係助詞+は，記号+読点，名詞+チャンス，‥

素性番号 2:付与素性例 出現品詞+名詞，出現品詞+動詞，出現品詞+格助詞，出現品詞+記号，‥

素性番号 3:付与素性例 冗長度+ランク 1

素性番号 4:付与素性例 2 単語連続+問題→は，2 単語連続+は→読点，2 単語連続+読点→チャンス，‥

素性番号 5:付与素性例 2 品詞連続+名詞→助詞，2 品詞連続+助詞→記号，2 品詞連続+記号→名詞，‥

素性番号 6:付与素性例 読点+1

素性番号 7:付与素性例 句点+2

素性番号 8:付与素性例 文長+40

上記の付与素性の例では，“+” の前の表現は素性の種類を示す記号であり，“+” の後ろの表現はその素性が持つ情報である．また以下のように接頭語を付与している．

- 単語素性はその単語の品詞が接頭語に付与される（例「名詞+問題」）
- 品詞素性は接頭語に「出現品詞」が付与される（例「出現品詞+名詞」）
- 冗長度素性は接頭語に「冗長度」が付与される（例「冗長度+ランク 1」）
- 素性は接頭語に「2 単語連続」が付与される（例「2 単語連続+問題→は」）
- 2 品詞連続素性は接頭語に「2 品詞連続」が付与される（例「2 単語連続+問題→は」）
- 読点素性は接頭語に「読点」が付与される（例「読点+1」）
- 句点素性は接頭語に「句点」が付与される（例「句点+2」）
- 文長素性は接頭語に「文長」が付与される（例「文長+40」）

学習データでの 10 分割クロスバリデーションでの性能が高い場合の素性の組み合わせを用いる．一つの素性のみを用いた推定をすべての素性で行い，性能が高かった素性を選ぶ．その素性と，残りの素性の一つを用いた推定を，残りの素性のすべての素性で行い，性能が高かった素性の組み合わせを選ぶ．上記を繰り返し行い，性能がそれ以上が上がらなくなった場合の素性の組み合わせを，テストデータでの推定に用いる．

4.3.3 冗長度に基づく手法

入力の文章において，機械学習に基づく手法の素性番号 3(冗長度) の素性の式 4.1 から冗長度をもとめ，閾値を設け冗長度が閾値以上の場合のみ冗長な文章と判定する．

閾値は学習データにおける 10 分割クロスバリデーションの正解率が高いものを用いる．閾値は 0.4 刻みで変更し，最大の正解率付近では 0.1 刻みで変更して正解率が最大になる閾値を探索する．

4.3.4 使用データ

使用するデータは 4.2.1 節で作成したデータ A とデータ B を用いる．データ A は学習データとして用いる．データ B はテストデータとして用いる．

4.3.5 実験

機械学習に基づく手法での素性選択

機械学習に基づく手法において，学習データでの 10 分割クロスバリデーションの実験により，学習データの正解率が高いときの素性の組み合わせを選択する．

1 個の素性のみを用いる実験を行った．その実験結果における正解率を表 4.1 に示す．表内の数字は，実験に使用する素性を示している．数字は，5.1.1 節での素性番号に対応している．

表 4.1: 素性選択 (1 回目)

素性	正解率
1	0.536(268/500)
2	0.482(241/500)
3	0.616(308/500)
4	0.474(237/500)
5	0.494(247/500)
6	0.582(291/500)
7	0.598(299/500)
8	0.572(286/500)

表 4.1 では，素性番号 3 が最も高い正解率を得ている．次に，素性番号 3 と残りの素性の一つを用いた機械学習をする．その結果を表 4.2 に示す．

同様にして表 4.3 と表 4.4 の実験を行った．

表 4.4 で最も性能高い場合の使用素性 [3,6,8,7] の正解率 0.634 が，表 4.3 の最高値である使用素性 [3,6,8] の 0.648 を下回ったので，素性 [3,6,8] がテストデータで利用する素性の組み合わせとなる．

参考にすべての素性を用いた場合の結果を表 4.5 に示す．

すべての素性を用いた場合の正解率 0.542 は，正解率が最大となる場合の素性の組み合わせを利用した場合の正解率 0.648 より小さいことが確認できる．

表 4.2: 素性選択 (2 回目)

素性	正解率
3,1	0.570(285/500)
3,2	0.590(295/500)
3,3	-
3,4	0.522(261/500)
3,5	0.552(276/500)
3,6	0.620(310/500)
3,7	0.610(305/500)
3,8	0.584(292/500)

表 4.3: 素性選択 (3 回目)

素性	正解率
3,6,1	0.598(299/500)
3,6,2	0.626(313/500)
3,6,3	-
3,6,4	0.540(270/500)
3,6,5	0.548(274/500)
3,6,6	-
3,6,7	0.628(314/500)
3,6,8	0.648(324/500)

表 4.4: 素性選択 (4 回目)

素性	正解率
3,6,8,1	0.582(291/500)
3,6,8,2	0.626(313/500)
3,6,8,3	-
3,6,8,4	0.538(269/500)
3,6,8,5	0.568(284/500)
3,6,8,6	-
3,6,8,7	0.634(317/500)
3,6,8,8	-

表 4.5: 全素性を利用した場合の結果

素性	正解率
1,2,3,4,5,6,7,8	0.542(271/500)

冗長度に基づく手法での閾値調整

表 4.6 は学習データにおける閾値ごとの正解率を示しており，正解率が最大であった閾値 1.4 の前後 ±0.1 の閾値の場合の正解率も確認している．表より，閾値 1.4 での正解率 0.620 が閾値 1.3，1.5 の正解率よりも大きく，最も性能が高かった．テストデータの実験に用いる閾値は 1.4 となる．

表 4.6: 冗長度に基づく手法の閾値調整

閾値	正解率
1.0	0.494(247/500)
1.3	0.590(295/500)
1.4	0.620(310/500)
1.5	0.610(305/500)
1.8	0.542(271/500)
2.2	0.504(252/500)
2.6	0.504(252/500)
3.0	0.504(252/500)

テストデータでの実験

機械学習に基づく手法と，冗長度に基づく手法でテストデータで冗長な文章の検出実験を行った．機械学習に基づく手法では，素性選択で得られた素性 [3,6,8] を利用した．冗長度に基づく手法では，閾値調整で得られた閾値 1.4 を利用した．比較として冗長度を利用しない機械学習として，素性 [6,8] を使う手法を用いる．実験結果を表 4.7 に示す．

表 4.7: 機械学習と冗長度に基づく冗長な文章の検出

手法	正解率
機械学習:素性 [6,8]	0.584 (292/500)
機械学習:素性 [3,6,8]	0.660(330/500)
冗長度	0.648(324/500)

SVM で検出できた例

例 1．デザインがベーシックだからこそ、風合いのちょっとしたニュアンスにいたるまで妥協は許されませんが、n 社は、長野県をはじめ、国内の優れた技術をもつ工場での生産にこだわりながら、自分たちにとって理想の服作りを目指しています。

例 2．参加者さんの中には、全く占星術が初めてという方、そして既に当店で占星術講座を受講されている方、有名な占星術師の方を通じて何となくはご存知で、更に詳しく聞いてみたいという方もいらっしゃいました。

冗長度で検出できた例

例 1．自然が持つ自己修復性を超えて負担をかけたり、自己修復性が損なわれたりすると、回復が遅れる。そして結果的に人類をはじめとした生物に悪影響を及ぼすことになる。

例 2.「お金」が発達するにつれ、われわれのリスクはすべからく値段に換算されることになった。いまや出産も葬式も、結婚も病気も、洗濯も食事も、教育も音楽も、おいしい水も山の空気さえ、マネーゲームに関与しないものはない。リスクはすっかり貨幣に乗っ取られてしまったのだ。

冗長度を利用しただけでも，ある程度の性能が得られた．冗長度は単純な式であるが，それが複数の文にまたがった冗長な文章の検出に役立つことがわかった．

4.4 考察

表 4.1 より「冗長度」を利用した際の正解率が最も高く正解率 0.616 であった．

素性選択を行った結果として表 4.3 より，冗長度 (素性番号 3) と共に句読点 (素性番号 6,7) や文長 (素性番号 8) の素性を追加すると正解率が向上した．句読点や文長については直接，文章の長さに関係するためだと思われる．句読点が少なく文長が長ければ，文章内の各文が長く冗長になりやすい．

表 4.7 よりテストデータにおいて機械学習を用いた手法と冗長度を用いる手法の比較をした結果，機械学習を用いた手法の正解率 (0.66) が，冗長度を用いる手法の正解率 (0.65) と同程度の正解率であった．冗長度の素性を用いずテストデータで検出を行った結果，正解率 0.584 と冗長度を用いた際より低くなった．

以上の結果により冗長度が冗長な文章の検出に役立つことがわかった．

第5章　冗長度の有効性の確認

前章のように，複数の文からなる文章では，冗長な文章の検出に，冗長度が役立つことが確認できた．本節では，冗長度が 1 文における冗長な文の検出に役立つかを調査する．冗長な文の検出の先行研究 [6] では，冗長な文の検出に冗長度は利用されていない．

5.1　提案手法

冗長な文の検出には，機械学習に基づく手法と冗長度に基づく手法の 2 つの手法を用いる．

5.1.1　機械学習に基づく手法

機械学習にはサポートベクトルマシンを用いる．機械学習では先行研究 [6] で使用した素性に冗長度を追加して用いる．使用素性の例は以下の通りである．

○素性番号 1(単語) 文内に出現する単語．

○素性番号 2(品詞) 文内に出現する単語の品詞．

○素性番号 3(3 単語連続) 文内に出現する 3 単語連続．文内に出現する単語を 3 単語ごとにつなげた素性である．

○素性番号 4(冗長度) 節同様の冗長度を用いた情報．

次に入力文に対して, 実際に付与される素性を大まかに示す．
入力文:「マシンの点検を行う。」

素性番号 1:付与素性例 名詞+マシン，助詞+の，名詞+点検，...

素性番号 2:付与素性例 出現品詞+名詞，出現品詞+助詞，出現品詞+動詞，...

素性番号 3:付与素性例 文字列+マシン，文字列+シンの，文字列+ンの点，...

素性番号 4:付与素性例 冗長度+ランク 1

上記の付与素性の例では，“+” の前の表現は素性の種類を示す記号であり，“+” の後ろの表現はその素性が持つ情報である．また以下のように接頭語を付与している．

- 単語素性はその単語の品詞が接頭語に付与される（例「名詞+マシン」）
- 品詞素性は接頭語に「出現品詞」が付与される（例「出現品詞+名詞」）
- 3 文字連続素性は接頭語に「文字列」が付与される（例「文字列+シンの」）
- 冗長度素性は接頭語に「冗長度」が付与される（例「冗長度+ランク 1」）

5.1.2 冗長度に基づく手法

各文で冗長度をもとめ，冗長度の値がある閾値以上場合に冗長な文と判定する．閾値は 1 から 2 までの値を 0.1 刻みで調整し用いた．

5.2 使用データ

実験には，5.2.1 節と 5.2.2 節で説明する 2 つのデータベースを用いる．

5.2.1 使用データ 1(収集データ)

収集データはウィキペディア[1]，解析済みブログコーパス (KNB コーパス)[2]より作成する．

図 5.1 はデータ作成の一連の流れである．

手順を以下に示す．

1. ウィキペディア・KNB コーパスおいて，冗長な文を収集する．冗長であるという判定基準に，表 5.1 と表 5.2 の「同義・類義な語が重複した表現」または「簡潔なものへの言い換えができる表現」を参考にしている．
2. 収集した冗長な文を人手で修正し，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．

[1]Wikipedia:http://ja.wikipedia.org/wiki/
[2]KNB コーパス，http://nlp.kuee.kyoto-u.ac.jp/kuntt/

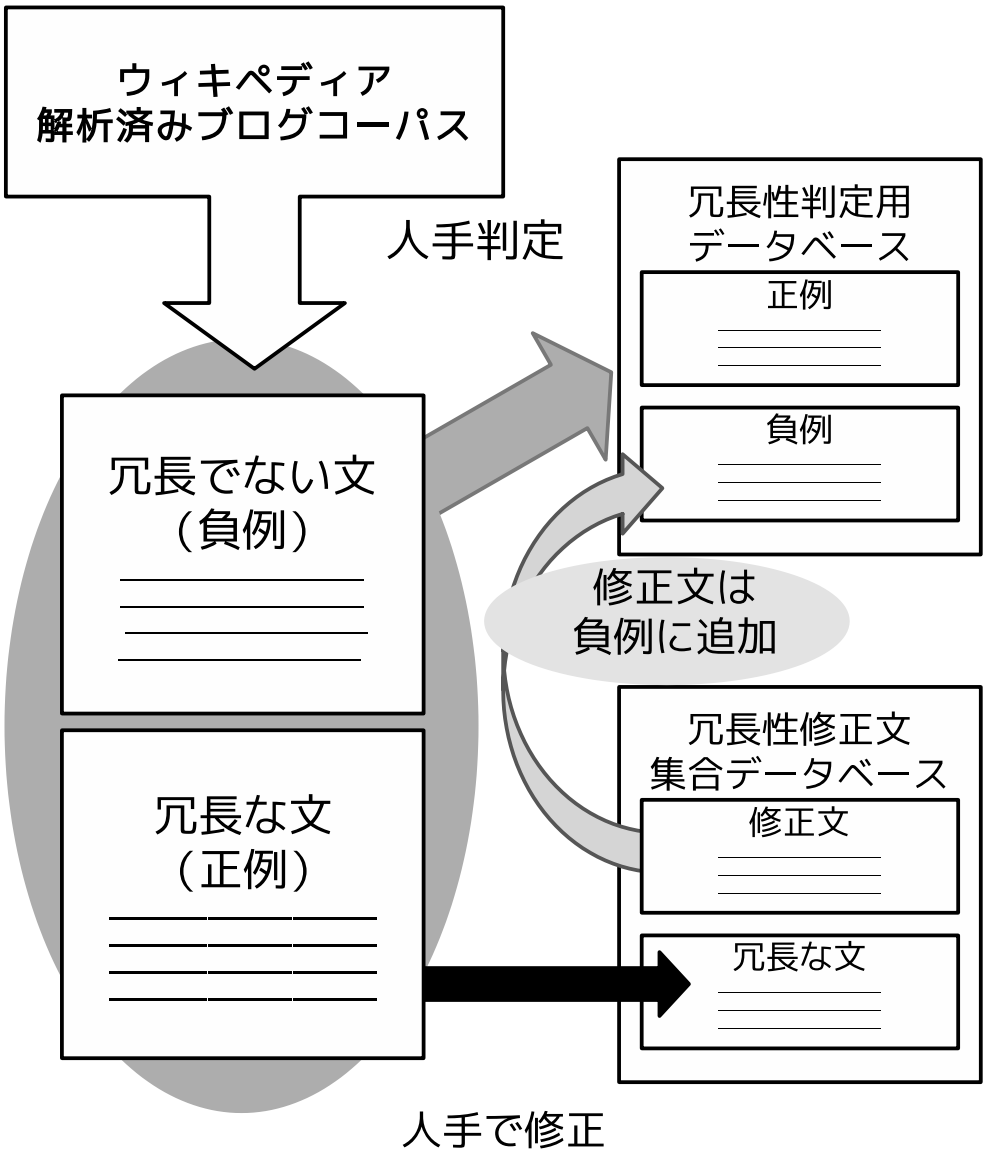

図 5.1: データベース作成

表 5.1: 同義・類義な語が重複した表現の例

表現の例	例
文意に影響しない二重の修飾	まず最初→最初
必要以上の強調	完全に一致→一致
1 文中に同じ語が近くにある表現	スポーツをしている人や散歩をしている人がいる→スポーツや散歩をしている人がいる
主語の単語を修飾語・補語・述語として同時に使用した表現	今日の天気はいい天気です→今日はいい天気です 検定方法は，○○法を使う→検定では，○○法を使う

表 5.2: 簡潔なものへの言い換えができる表現の例

表現の例	例
必要以上の漢語	存在する→ある
冗長な文末表現	～あるものである→～ている
複合語として言い換えができる表現	解決に向けた策→解決策
曖昧な表現	以下のような例→以下の例

図 5.2 は収集データの例である．収集した冗長な文には「冗長な文」のタグを付与し，その修正文には「修正文」のタグを付与している．

収集したデータは冗長な文と修正文をあわせて 800 文である．実験に利用するのはここからランダムに取り出した 400 文である．

5.2.2 使用データ 2(作例データ)

冗長な文を作例し，その適切な修正を行い対として作成したデータベースを作例データという．

データベースを作成する．データベース作成は言語データ作成に熟練したものが行った．データ例を表 5.3 に示す．作例された冗長な文には「冗長な文」のタグを付与し，その修正文には「修正文」のタグを付与している．

データ数は冗長な文 650 文，冗長でない文 650 文の合計 1,300 文である．実験に利用するのはそこからランダムに取り出した 500 文である．

冗長な文 **1** まず初めにマシンの点検を行う。
修正文　**1** まずマシンを点検する。

冗長な文 **2** まず初めに円高の解決に向けた策の検討を考えたい。
修正文　**2** まず円高の解決策を検討したい。

冗長な文 **3** スポーツをしている人や散歩をしている人，昼寝をしている人など，日常の京都人の姿を拝見できました.
修正文　**3** スポーツまたは散歩，昼寝をする人など，日常の京都人の姿を拝見できました.

冗長な文 **4** しかし、今日食べる用の方と、後日食べる用の方を見比べてみると、片方が春雨ばっか、もう片方がキクラゲばっか.
修正文　**4** しかし、今日と、後日食べる用の方を見比べると、片方が春雨、もう片方がキクラゲばっか.

冗長な文 **5** 今まで私が食べてきたパフェをパフェと呼ぶならば、このつじりのパフェは超パフェであります。
修正文　**5** 今まで私が食べてきたパフェがパフェならば、つじりのパフェは超パフェである。

冗長な文 **6** なぜか、人の作ったものは、自分で作ったものよりおいしい。
修正文　**6** なぜか、人のは、自分で作ったものよりおいしい。

冗長な文 **7** そこで、まだ短い自炊暦ですが、このブログをみて自炊を始める人がいることを願って、自炊のときに必要だと思うことことをあげてみました。
修正文　**7** そこで、まだ短い自炊暦だが、このブログをみて自炊を始める人がいることを願い、自炊のときに必要なことをあげてみました。

冗長な文 **8** 先輩は天才肌で明るい先輩である。
修正文　**8** 天才肌で明るい先輩である。

冗長な文 **9** 近くに吸っている人がいるだけでとても不愉快になります。
修正文　**9** 近くに喫煙者がいるだけでとても不愉快になります。

図 5.2: 収集データの例

冗長な文 **1** 体質を改善するというのは、一朝一夕にはいかないものです。
修正文　**1** 体質を改善するのは、一朝一夕にはいかないです。

冗長な文 **2** 目的をはっきりと明確にした上で実験を行う必要がある。
修正文　**2** 目的を明確にした上で実験を行う必要がある。

冗長な文 **3** この研究の目的は犯罪の実体を明らかにすることをねらいとしている。
修正文　**3** この研究の目的は犯罪の実体を明らかにすることである。

冗長な文 **4** 先頭騎手が馬から落馬した。
修正文　**4** 先頭騎手が落馬した。

冗長な文 **5** 多彩なメニューを彩りよく盛り込みました。
修正文　**5** 多彩なメニューを盛り込みました。

冗長な文 **6** いまこそ絶対間違いなくお得です。
修正文　**6** いまこそ間違いなくお得です。

冗長な文 **7** 別館の建築予定は 12 月末までに完成する見込みである。
修正文　**7** 別館は 12 月末までに完成予定である。

冗長な文 **8** 世界に海外展開する。
修正文　**8** 世界に展開する。

冗長な文 **9** 沖縄では夏に元気をつけると伝えられており、古くから香辛料として料理の味付けに使われてきました。
修正文　**9** 沖縄では夏に元気をつけると伝えられており、香辛料として料理の味付けに使われてきました。

冗長な文 **10** 緑茶に含まれる渋み成分のカテキンは、ポリフェノールの一種です。
修正文　**10** 緑茶の渋み成分のカテキンは、ポリフェノールの一種です。

冗長な文 **11** ペダルは逆回転もできるので、足腰をバランスよく鍛えられます。
修正文　**11** ペダルの逆回転で、足腰をバランスよく鍛えられます。

図 5.3: 作例データの例

5.3 結果

機械学習に基づく実験では，評価は 10 分割クロスヴァリデーションで行った．また素性は素性番号 4 の冗長度を用いる場合と用いない場合の 2 種類を試した．冗長度に基づく実験では，閾値は 0.1 刻みで変化させてもとめた．

評価として，正解率，再現率，適合率，F 値をもとめた．再現率と適合率は以下の式で算出される．

$$\text{再現率} = \frac{\text{システムの正解数}}{\text{テストデータ中の正解数}} \tag{5.1}$$

$$\text{適合率} = \frac{\text{システムの正解数}}{\text{システムの出力数}} \tag{5.2}$$

また (5.1) と (5.2) の値の調和平均 (5.3) を求めることで F 値を算出できる．

$$F\text{値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{5.3}$$

正解率は使用データ全体での正解の割合である．再現率，適合率，F 値は冗長な文を抽出する場合のものをもとめた．

収集データでの各手法による冗長な文の検出結果を表 5.3 と表 5.4 に示す．

表 5.3: 機械学習による検出結果 (収集データ)

素性	正解率	再現率	適合率	F 値
1,2,3,4	0.573(229/400)	0.420(68/162)	0.469(68/145)	0.443
1,2,3	0.570(228/400)	0.395(64/162)	0.464(64/138)	0.427

作例データの各手法による検出結果を表 5.5 と表 5.6 に示す．

機械学習において，素性 [1,2,3,4] を用いたとき 0.573 の正解率を得た．また F 値では，0.443 の値を得た．これは，冗長度の素性を用いなかった場合の正解率 (0.570) や F 値 (0.427) よりも高い結果である．冗長度の素性を追加で用いることで機械学習の性能が向上することが確認できた．

表 5.4，表 5.6 において，冗長度を用いる手法の性能が，収集データでは，閾値 1.5 で正解率 0.620 を得た．また作例データにおいても閾値 1.2 で正解率 0.568 を得た．機械学習の手法よりも高い場合があることがわかる．

冗長度を用いる手法が，複数の文だけでなく 1 文での冗長な文の検出にも役立つことが確認できた．

表 5.4: 冗長度による検出結果 (収集データ)

閾値	正解率	再現率	適合率	F 値
1.0	0.405(162/400)	1.000(162/162)	0.405(162/400)	0.577
1.1	0.580(232/400)	0.469(76/162)	0.481(76/158)	0.475
1.2	0.595(238/400)	0.210(34/162)	0.500(34/ 68)	0.296
1.3	0.613(245/400)	0.105(17/162)	0.630(17/ 27)	0.180
1.4	0.620(248/400)	0.080(13/162)	0.812(13/ 16)	0.146
1.5	0.620(248/400)	0.068(11/162)	0.917(11/ 12)	0.126
1.6	0.620(248/400)	0.068(11/162)	0.917(11/ 12)	0.126
1.7	0.608(243/400)	0.037(6/162)	0.857(6/ 7)	0.071
1.8	0.600(240/400)	0.019(3/162)	0.750(3/ 4)	0.036
1.9	0.600(240/400)	0.012(2/162)	1.000(2/ 2)	0.024
2.0	0.598(239/400)	0.006(1/162)	1.000(1/ 1)	0.012

表 5.5: 機械学習による検出結果 (作例データ)

素性	正解率	再現率	適合率	F 値
1,2,3,4	0.526(263/500)	0.420(97/231)	0.485(97/200)	0.450
1,2,3	0.516(258/500)	0.407(94/231)	0.472(94/199)	0.437

表 5.6: 冗長度による検出結果 (作例データ)

閾値	正解率	再現率	適合率	F 値
1.0	0.462(231/500)	1.000(231/231)	0.462(231/500)	0.632
1.1	0.550(275/500)	0.355(82/231)	0.519(82/158)	0.422
1.2	0.568(284/500)	0.139(32/231)	0.653(32/049)	0.229
1.3	0.542(271/500)	0.013(3/231)	0.750(3/ 4)	0.026
1.4	0.536(268/500)	0.000(0/231)	0.000(0/ 1)	0.000
:	:	:	:	:
1.9	0.536(268/500)	0.000(0/231)	0.000(0/ 1)	0.000
2.0	0.536(268/500)	0.000(0/231)	0.000(0/ 1)	0.000

5.4 考察

1 文における冗長な文において機械学習と冗長度を用いる手法で検出を行った．先行研究 [6] の機械学習に冗長度を素性に追加したところ収集データ，作例データの両方で，正解率，再現率，F 値で性能向上が見られた．

冗長度による検出結果においても，収集データでは正解率 0.62 という機械学習より高い値を得た．この結果により，冗長度を用いる手法の性能が機械学習の手法よりも高い場合があることがわかる．このことからも冗長度の有用性が確認できた．

第6章　おわりに

本研究では冗長な文を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」の存在が原因となって冗長となった文を修正する実験を行った．修正前と修正後の両方の表現を推定する場合，機械学習が関係する手法では，6 割の正解率を得た．修正後の表現のみを推定する場合においても，機械学習が関係する手法で，7 割の正解率を得た．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，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.

冗長な文章での実験では，次の知見が得られた．冗長な文章の分析により典型的な 3 種類の分類を示した．また機械学習を用いる手法と，冗長度を用いる手法により冗長な文章を検出した．機械学習を用いた実験では機械学習の素性として「冗長度」を利用した際の正解率が最も高かった．機械学習を用いた手法の正解率 (0.66) が，冗長度を用いる手法の正解率 (0.65) と同程度の正解率であった．1 文における冗長な文において先行研究 [6] の機械学習に冗長度を素性に追加したところ性能向上が見られた．冗長度を用いる手法が機械学習の手法より高い性能を出す場合があることが確認できた.

謝辞

本研究を進めるに当たり，終始に渡り研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科計算機工学講座 C の村田真樹教授，村上仁一准教授，徳久雅人講師に心から御礼申し上げます．また，ご多忙の中，助言をいただきました木村周平教授に厚く御礼申し上げます． 加えて，種々の御助言を龍谷大学理工学部数理情報学科の馬青教授に頂きました．ここに深く感謝いたします．その他様々な場面で御助言を頂いた計算機工学講座 C 研究室の皆様に感謝の意を表します．

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, vol.29, No.1, 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”, 言語処理学会第 18 回年次大会, pp.1114-1117.

[7] 大竹清敬, 船坂貴浩, 増山繁, 山本和英 (1999), “重複部・冗長部削除による複数記事要約手法”, 言語処理学会論文誌「自然言語処理」, Vol.6, No.6, pp.45-64.

[8] 村田真樹, 金丸敏幸, 白土保, 井佐原均 (2006), “受け身文の能動文への変換における機械学習を用いた格助詞の変換に関する実験”, 情報科学技術レターズ, Vol.5, pp.89-92.

[9] 原口智史, 坂本佳史, 中田武男, 竹内広宜, 荻野紫穂 (2011), “テキスト分析技術を用いた開発関連文書の文書品質の定量化”, 電子情報通信学会技術研究報告「思考と言語」(TL), Vol.111, No.98, pp.25-30.

[10] 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.

[11] 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”, 言語処理学会論文誌「自然言語処理」, Vol.9, No.2, pp.91-110.