概要

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．

本研究では，冗長な文の改善をするために，冗長な文の収集と分析を行い，それとともに冗長な文の自動検出を試みる．

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，ウェブ上のデータから冗長な文と冗長でない文を収集し，収集したデータに基づく冗長な文に関する分析を行った．収集したデータおよび分析結果は，すべての文を一つの機械学習で扱う方法ではそれほど良い性能を出すことはできなかったが，特定の表現を含む文の集合ごとに機械学習を行う方法 (特定の表現の種類の数だけ機械学習が必要) では，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．

目 次

表 目 次

図 目 次

第1章　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」という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，冗長な文の収集と分析を行うとともに，冗長な文の自動検出を試みる．

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．

本研究の主な主張点は以下の 3 つである．

1. 本研究では，ウェブ上のデータから冗長な文と冗長でない文を収集し，収集したデータに基づく冗長な文に関する分析を行った．収集したデータおよび分析結果は，冗長な文に関わる研究や処理のための貴重な資料となる．

2. 本研究は機械学習を用いて冗長な文の検出を行う初めての試みである．

3. 機械学習を利用した冗長な文の検出は，すべての文を一つの機械学習で扱う方法ではそれほど良い性能を出すことはできなかった．しかし，特定の表現を含む文の集合ごとに機械学習を行う方法 (特定の表現の種類の数だけ機械学習が必要) では，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．

本論文の構成は以下の通りである．第 2 章では，卒業研究である冗長な文の機械的分析と検出の全体の流れについて述べる．第 3 章では冗長な文の機械的分析の提案手法や使用データの説明を行い，分析の結果を示す．第 4 章では，冗長な文の機械的検出の提案手法とそれを用いた検出実験の結果を示す．第 5 章では本研究の関連研究を述べる．

第2章　研究の流れ

本研究では，初めに，冗長な文に関わるデータベースを作成する．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]から冗長な文と冗長でない文を収集する．冗長であると判断された文については人手で冗長でない文に修正する．これらの文から冗長な文と冗長でない文を含むデータベースを作成する．

作成したデータベースに含まれる，冗長な文とそれを修正した文を比較し冗長箇所の頻度分析をする．これにより，冗長な文に頻出する表現などの冗長な文に関わる特徴を見つける．

次に，作成したデータベースを利用して，機械学習を利用した冗長な文の検出の研究を行う．データベースの冗長な文と冗長でない文をそれぞれ学習データの正例，負例として用いる．機械学習により，冗長な文をどの程度検出できるかを調べる．

最後に，冗長な文に頻出する個々の表現に着目した，機械学習を利用した冗長な文の検出を行う．個々の特定の表現を含む文の集合ごとに機械学習を行う方法 (特定の表現の数だけ機械学習する) で入力の文が冗長な文であるか否かの判定を行う．

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

第3章　データ分析

3.1　冗長な文の収集とその分析

冗長な文と，それを冗長でないように修正した文を比較することで，冗長な文における特徴的な表現を見つけることができる．本章では，この比較に基づく分析について述べる．

3.1.1　提案手法

われわれの提案する分析手法は以下のとおりである．3.1.2 節で述べる「冗長性修正文集合データベース」にある冗長な文とその修正文をそれぞれ，形態素解析 ChaSen[1] にかけ単語単位に分割をする．例えば「まず初めにマシンの点検を行う。」という文を形態素解析にかけると図 3.1 に示した結果となる．

まず初めにマシンの点検を行う

```
まず マズ まず 副詞-一般
初め ハジメ 初め 名詞-副詞可能
に ニ に 助詞-格助詞-一般
マシン マシン マシン 名詞-一般
の ノ の 助詞-連体化
点検 テンケン 点検 名詞-サ変接続
を ヲ を 助詞-格助詞-一般
行う オコナウ 行う 動詞-自立 五段・ワ行促音便 基本形
。 。 。 記号-句点
EOS
```

図 3.1: 形態素解析の例

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

このように形態素ごとに「冗長な文」とその「修正文」を分割し，分割した各データを比較し冗長箇所の検出をする．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．

冗長箇所

点検　テンケン　点検　名詞-サ変接続
を　ヲ　を　助詞-格助詞-一般
行う　オコナウ　行う　動詞-自立　五段・ワ行促音便　基本形
。　。　。　記号-句点
EOS
点検　テンケン 点検　名詞-サ変接続
する　スル　する　動詞-自立　サ変・スル　基本形
。　。　。　記号-句点
EOS

図 3.2: 冗長箇所の例

図 3.2 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」に修正された「を行う」が冗長であると考え「冗長箇所」としている．

冗長箇所を作成データで検出し，その頻度を求める．頻度としては一単語の頻度を求めるもの (例:一単語である「行う」の頻度を求める) と，二単語連続の頻度を求めるもの (例:二単語連続である「を　行う」の頻度を求める) の二種類を行う．これによってどのような表現が冗長な文に頻出するかを調べる．また頻出表現について修正により冗長な表現がどのように変化したかを調べる．

3.1.2 データ

ウィキペディアと，解析済みブログコーパス (KNB コーパス) において冗長な文を正例，冗長でない文を負例として収集する．冗長であるという判定基準には，表 3.1 と表 3.2 を用いる．これらの表にあてはまるものを冗長な表現とし人手で判定する．

表 3.1: 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 3.2: 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

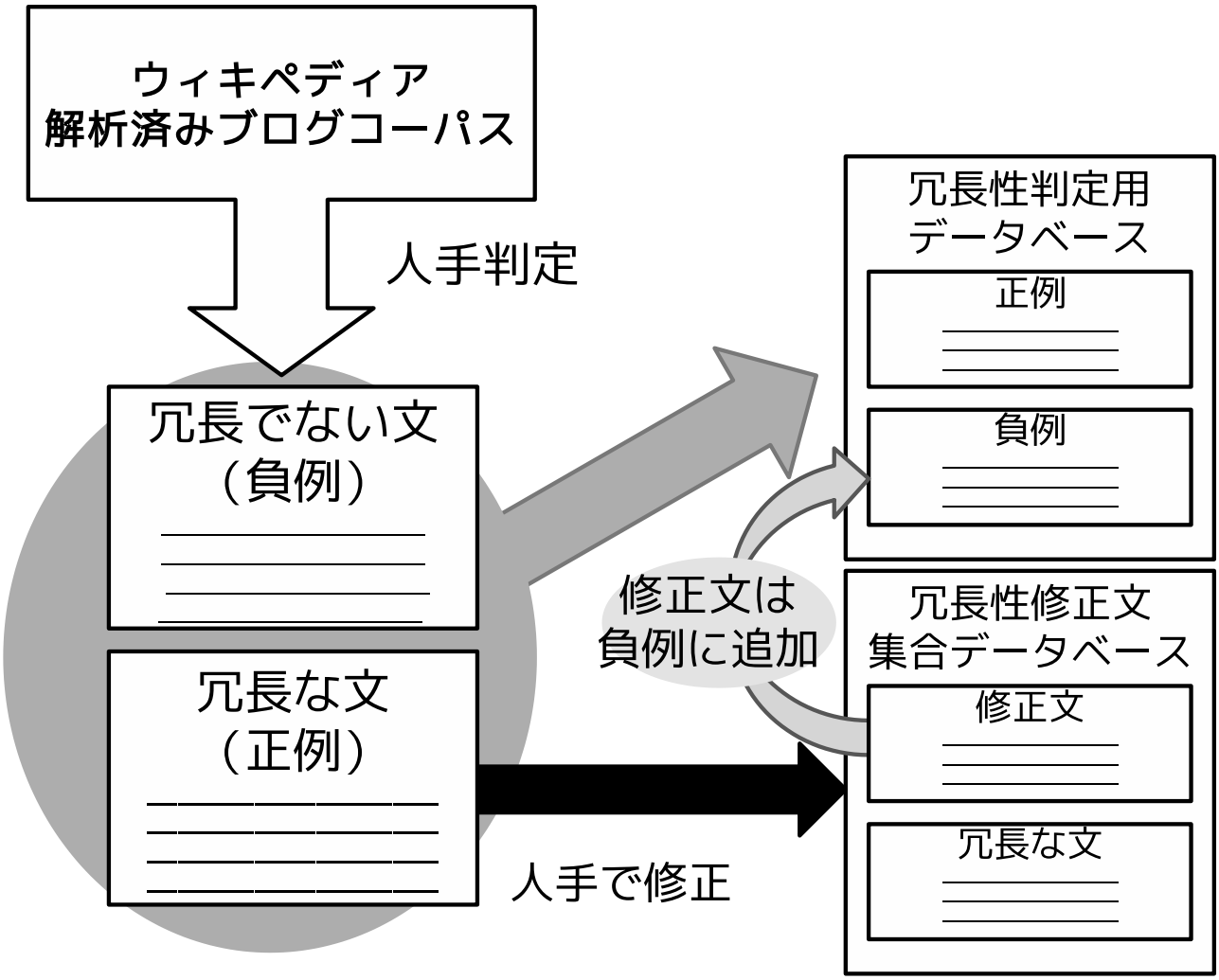

図 3.3: データベース作成

収集された正例と，負例を用いて「冗長性判定用データベース」を作成する．収集された正例については人手で冗長でない文に修正する．冗長な文と修正後の文を対として収集し「冗長性修正文集合データベース」を作成する．人手で修正した文は「冗長性判定用データベース」の負例としても用いる．

図 3.4 は冗長性判定用データベースの例である．収集してきた冗長な文には「正例」のタグを付与し，収集してきた冗長でない文には「負例」のタグを付与している．

冗長性判定用データベース

正例 まず初めにマシンの点検を行う。

正例 以下は政治哲学におけるものに絞った解説である。

負例 古代から現代の義務教育に通ずる社会制度はあった。

・
・
・
・
・
・

負例 現代の戦略にはいくつかのレベルがある。

正例 学校においては、授業料を徴収することができる。

正例 民主主義には、多数の立場や観点からの多くの評価や批判が存在している。

図 3.4: 冗長性判定用データベースの例

図 3.5 は冗長性修正文集合データベースの例である．一行目の「修正前」とタグ付けされたの文は, 収集してきた冗長な文を原文で載せており，続いて 2 行目の「修正後」とタグ付けされた文は, 表 3.1 と表 3.2 の規則を用いて冗長でない文に修正した文を表している.

冗長性修正文集合データベース

```
-----------------------------------------
修正前 まず初めにマシンの点検を行う。
修正後 まずマシンを点検する。
-----------------------------------------
修正前 以下は政治哲学におけるものに絞った解説である。
修正後 以下は政治哲学に絞った解説である。
-----------------------------------------
                                    .
                                    .
                                    .
                                    .
                                    .
                                    .
-----------------------------------------
修正前 学校においては、授業料を徴収することができる。
修正後 学校においては、授業料を徴収できる。
-----------------------------------------
修正前 民主主義には、多数の立場や観点からの多くの評価や批判が存在している。
修正後 民主主義には、多数の立場や観点からの多くの評価や批判がある。
-----------------------------------------
```

図 3.5: 冗長性修正文集合データベースの例

収集した「冗長性判定用データベース」は正例と負例をあわせて 850 文を,「冗長性修正文集合データベース」は冗長な文は 350 文であり，それを修正したものを合わせて合計 700 文を作成した.

3.1.3 実験と結果

「冗長性修正文集合データベース」を用いて修正対象になりやすい表現を分析した．表3.3と表3.4に結果を示す．表3.3の「一単語ごと」は冗長な文における修正部分に含まれる単語の頻度を求めた結果の一部を示しており,「二単語ごと」は修正部分に含まれる二単語連続の頻度を求めた結果の一部を示している．

表 3.3: 修正部分に含まれる表現の頻度

一単語ごと		二単語ごと	
頻度	単語	頻度	単語
23	もの	26	で ある
15	行う	7	と いう
10	存在	7	する こと
4	可能	6	が できる

表3.3より，例えば「もの」「行う」などの表現が冗長な表現になりやすいものであることがわかった．

表 3.4 には，冗長な文を冗長でない文に修正した際の修正例の一部を示している．

表 3.4: 冗長な表現の修正例

修正前	修正後	例文
ものである	削除	カントは宗教を、道徳の基礎の上に成り立つべきものであるとしている。
を行う	する	一人が複数票を投票する複数選挙や等級ごとに選挙を行う等級選挙がある。
存在する	ある	今日、数え切れぬほど多くの経済学が存在するものの、私は二つしか認めない。
可能である	できる	司法に該当しない国家作用であっても、法律により裁判所に権限を与えることは可能である。
という	削除	歴史的意味においてでないかぎり哲学を学ぶという ことはできない。
行わ	さ	一方、ヨーロッパ大陸ではキリスト教の精神により古くから慈善事業が行われてきた。
である	削除	一方で劣等人種の血が優勢にならないように、その流入を防ぐことも必要であるとされた。
すること	削除	民主国家では、言論の自由が与えられているため、いくらでも権力者である政治家を批判、弾劾、ときに揶揄することができる。

第4章　冗長な文の検出

本章では収集したデータまたは分析結果をもとに機械学習を用いて冗長な文の検出を行う．機械学習法には，サポートベクターマシン (以下 SVM) を用いる．(サポートベクターマシンの説明は後述)

4.1　サポートベクターマシンとは

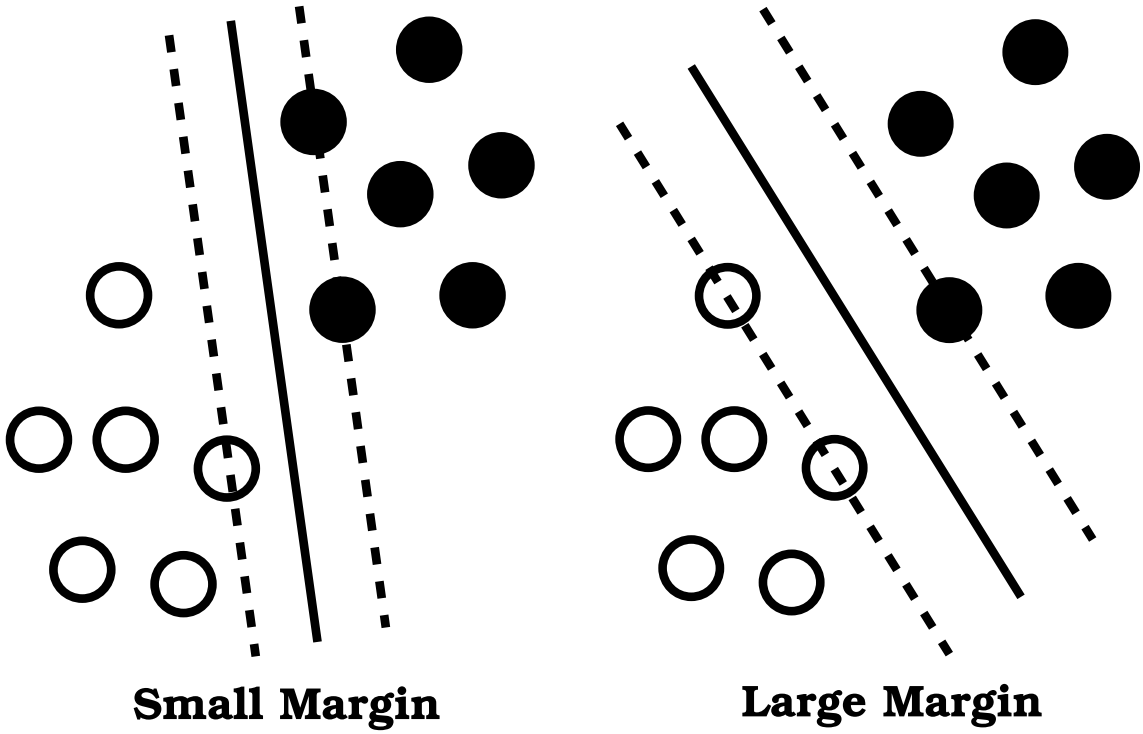

図 4.1: マージンの最大化

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．

$$\begin{aligned} f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \\ b &= -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2} \\ b_i &= \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) \end{aligned} \tag{4.1}$$

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，

$$\begin{aligned} sgn(x) = \quad & 1 \quad (x \geq 0) \\ & -1 \quad (otherwise) \end{aligned} \tag{4.2}$$

であり，また，各 α_i は式 (4.4) と式 (4.5) の制約のもと式 (4.3) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{4.3}$$

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{4.4}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4.5}$$

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．

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{4.6}$$

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 (4.1) 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．

4.2 評価方法

本章では機械学習の精度は再現率（recall），適合率（precision），F 値（F-measure）で評価している．再現率と適合率は以下の式で算出される．

$$再現率 = \frac{システムの正解数}{テストデータ中の正解数} \tag{4.7}$$

$$適合率 = \frac{システムの正解数}{システムの出力数} \tag{4.8}$$

また (4.7) と (4.8) の値の調和平均 (4.9) を求めることで F 値を算出できる．

$$F値 = \frac{2 \times 適合率 \times 再現率}{適合率 + 再現率} \tag{4.9}$$

4.3 検定方法

本研究の検定方法は 10 分割クロスバリデーションを用いている．

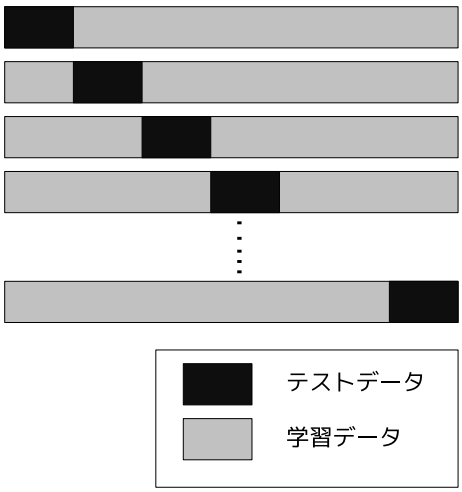

図 4.2: 10 分割クロスバリデーション

10 分割クロスバリデーションでは、標本群を 10 個に分割し，そのうちの 1 つをテストデータとし，残る 9 個を訓練事例とする．そして 10 個に分割された標本群それぞれをテスト事例として 10 回推定を行う．そうして得られた 10 回の結果組み合わせてテスト事例全体の推定結果を得る．

4.4 冗長な文の検出 1

機械学習で冗長な文と冗長でない文をどの程度検出できるのかを調査した.

4.4.1 提案手法

教師あり機械学習 [9] により各文が冗長な文か否かを判定する.

機械学習の素性として以下のものを用いる.

使用素性

素性 1 単語とその品詞

素性 2 単語の品詞

素性 3 3 文字列

これらの素性は，例えば「マシンの点検を行う」という文では表 4.1 のようになる.

表 4.1: 各素性の例 (冗長な文の検出 1)

素性名	素性の例
素性 1(単語)	マシン:名詞, の:助詞, 点検:名詞, ...
素性 2(品詞)	名詞, 助詞, ...
素性 3(3 文字列)	文字列:マシン 文字列:シンの, ...

4.4.2 データ

「冗長性修正文集合データベース」の 429 文を用いる (用いたデータの内訳は正例 170 文，負例 259 文である).

4.4.3 実験と結果

学習データとして 4.4.2 節のデータを用い，10 分割クロスバリデーションにより評価する．

ここでベースラインとして全ての文を正例と判定するものを用い，比較を行う．

結果を表 4.2 に示す．ここでの F 値は正例の文を抽出する性能を示すものである．

表 4.2: 冗長な文の検出性能

	再現率	適合率	F 値
提案手法	0.45 (76/170)	0.52 (76/146)	0.48
ベースライン	1.00 (170/170)	0.40 (170/429)	0.57

提案手法は適合率では 0.1 ほどベースラインより高かったが，F 値ではベースラインより低かった．

しかし一般に世に存在する文に含まれている冗長な文は冗長でない文に比べて量は少ないと思われる．そこで実際の出現頻度は冗長でない文が冗長な文よりも多くなると仮定し性能を算出してみた．表 4.3 に負例数を 10 倍にした場合の結果を示す．負例を 10 倍にすると F 値でもベースラインを上回った．

表 4.3: 負例数を 10 倍にした場合の検出性能

	再現率	適合率	F 値
提案手法	0.45 (76/170)	0.10 (76/ 776)	0.16
ベースライン	1.00 (170/170)	0.06 (170/2760)	0.12

結果としてベースラインを上回るものの F 値は 0.16 と，提案手法の性能は高いものではない．

4.5　冗長な文の検出2

4.4節での機械学習ではあまりよい結果は得られなかった．そこで，村田らの行った単語多義性解消問題の機械学習手法[8]を参考にし，本章では表現ごとに逐次的に機械学習を行うこととした．すべての文に対して一つの機械学習をするのではなく，特定の表現を含む文の集合に対して一つの機械学習を行う．

3.1節の分析で「もの」「である」のような表現は冗長になりやすいと確認できた．そのため分析した「冗長な文」に出やすい表現に着目する．なぜならこれらの表現は必ずしも冗長な表現になるわけではないと思われる．例えば,「すること」という表現では,図4.3のように冗長な場合と冗長でない場合がある．

必ずしも冗長でない表現

「すること」という表現について

冗長な例

　　軽く すること ができる。

冗長でない例

　　合理的発展に資 すること を定めた法律。

図 4.3: 必ずしも冗長でない表現の例

例の「軽くすることができる」の文中に含まれる「すること」は「軽く」をただ強調しているだけのため冗長な表現だといえ，例えば「軽くできる」と修正できる．しかし「合理的発展に資することを定めた法律」の文中に含まれる「すること」は資するという動詞の一部なので修正することができない．

そこでこのような必ずしも冗長ではない文について逐次的な機械学習をし冗長な文を検出する．

この検出は表現の個数分，機械学習をすることになる．例えば，特定の表現として「可能」「という」の二つがあった場合,「可能」を含む文の集合に対して一つの機械学習を行い,「という」を含む文の集合に対して一つの機械学習を行う.「可能」を含む文が冗長かいなかを判定する際には,「可能」を含む文の集合で学習した結果を利用し行う．本章ではこの考え方に基づいて行った冗長な文の検出について述べる．

4.5.1 提案手法

機械学習により特定の表現 (対象表現と呼ぶ) を含む文が冗長であるか否かを判定する．機械学習は，対象表現の種類の数だけ行う．

機械学習の素性として以下のものを用いる．

使用素性

素性 1 文中の対象表現の前後各 2 単語

素性 2 文中の対象表現の前後各 2 単語の品詞

例えばこれらの素性は「〜与えることは可能である。」の文では表 4.4 のようになる．この例での対象表現は「可能」である．

表 4.4: 各素性の例 (冗長な文の検出 2)

素性名	素性の例
素性 1(前後 2 単語)	こと, は, だ, ある
素性 2(前後 2 品詞)	名詞, 助詞, 助動詞, 助動詞

4.5.2　データ

図 4.4 は機械学習で検出する対象の表現を求める一連の流れである．

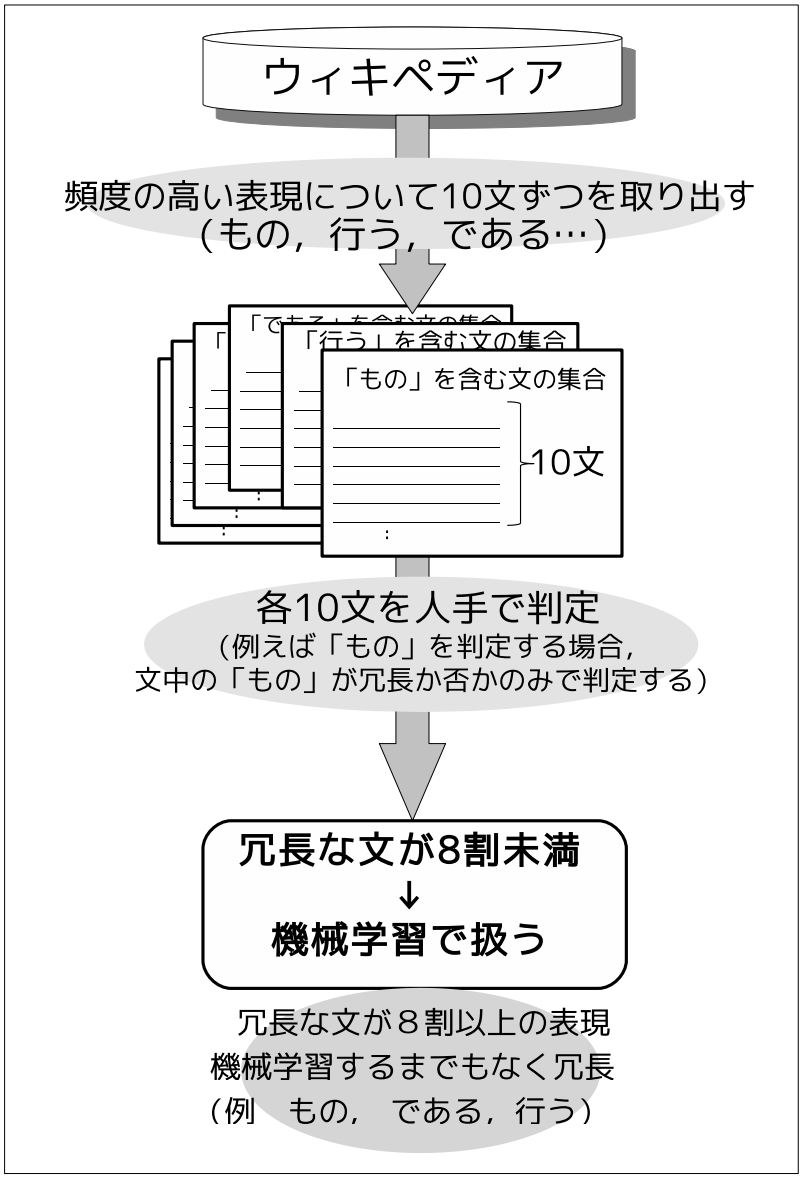

図 4.4: データベース作成

まず 3.1 節で頻度分析をした結果から修正頻度の高い表現について，各表現を含む文をウィキペディアからランダムに 10 文ずつ収集する．取り出した 10 文について手作業で判定し，冗長である文を正例，冗長でない文を負例とする．判定した結果正例の割合が 8 割未満の表現を機械学習で扱う．正例の割合が 8 割以上の表現については，機械学

習を用いるまでもなく，冗長な表現の検出に利用できる手がかりと考えることができるため，ここでの実験には用いない．正例の割合が 8 割未満の表現としては,「可能」「という」「すること」が見つかった．この表現を含む文をウィキペディアからさらにランダムに収集し，手作業で判定して冗長である文を正例，冗長でない文を負例とする.「可能」「という」「すること」のそれぞれについて 100 文ずつ合計 300 文のデータを作成する．ここでの正例と負例の判断では,「可能」などの対象表現が冗長な表現を構成する場合正例,そうでない場合負例とする．対象表現以外の箇所が冗長であるか否かはこの判断では利用しない.

4.5.3 実験と結果

4.5.2 節のデータを学習データとして 10 分割クロスバリデーションを行って評価した．ベースラインとして全て正例と判定するものを用い，比較した．

結果を表 4.5 から表 4.7 に示す．ここでの F 値は正例の文を抽出する性能を示すものである．

表 4.5: 「可能」に関する機械学習の結果

	再現率	適合率	F 値
提案手法	0.87 (47/54)	0.87 (47/54)	0.87
ベースライン	1.00 (54/54)	0.54 (54/100)	0.70

表 4.6: 「という」に関する機械学習の結果

	再現率	適合率	F 値
提案手法	0.61 (25/41)	0.83 (25/ 30)	0.70
ベースライン	1.00 (41/41)	0.41 (41/100)	0.58

表 4.7: 「すること」に関する機械学習の結果

	再現率	適合率	F 値
提案手法	0.69 (29/42)	0.74 (29/ 39)	0.71
ベースライン	1.00 (42/42)	0.42 (42/100)	0.59

表 4.5 から 4.7 のように提案手法は「可能」「という」「すること」の表現について 0.7 から 0.8 という高い F 値を得た．ベースラインよりも検出性能が高かった．

これらの表現については学習データを増やすことで性能の向上が期待できる．

表 4.8 に提案手法で正しく冗長な文であると判断できた文の例を示す.

表 4.8: 正しく判断できた正例

	正例
可能	しかし、この考え方は現実的にも 適応 可能 である。→例えば (適応できる) に修正可
	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。→例えば (できる) に修正可
	無泥水・無排土での 施工が 可能 であり、経済的である。→例えば (施工でき) に修正可
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。→例えば (計測されること) に修正可
	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。→ここでの（という）は削除可
	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 の定数の特徴である。→ここでの（という）は削除可
すること	以上より、問題の積分を 計算 すること ができた。→例えば (計算できた) に修正可
	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。→例えば (主張できる) に修正可
	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。→例えば (軽くできる) に修正可

表 4.8 より例えば「可能」についての例文の「しかし、この考え方は現実的にも適応可能である」という文を考えてみる．文中の「適応可能である」は，定義「簡潔なものへの言い換えができる表現」の「必要以上の漢語」にあてはまる．よってこの文は例えば「しかし、この考え方は現実的にも適応できる」と簡潔化ができるため，冗長であるといえる．そのため冗長な文を正しく推定できているといえる．

表 4.9 に提案手法で正しく冗長な文でないと判断できた文の例を示す.

表 4.9: 正しく判断できた負例

	負例
可能	再帰理論において原始再帰関数は、計算 可能 性の完全形式化のための重要な要素となる関数
	代数的半順序集合は、それが有限のものに制限されていても、すべての要素の近似を 可能 にするため、表示的意味論の観点からはとりわけ行儀よくふるまう。
	計算 可能 性理論の大部分はこの停止問題の結果に基づいて構築されている。
という	文の成立について、山田は「陳述」という 用語を用いた。
	居酒屋でイングラム・フライザー という 男と口論から喧嘩になり、ナイフで刺されて死亡したのち無縁墓地に埋葬された。
	地方議員の定数は、地方自治法により議員定数の上限数を定められているが、議員の定数が多いので削減すべき という 議論がある。
すること	つまり歪曲 すること を求めていないのではないか？
	漁船法は、漁船の建造を調整し、漁船の登録及び検査に関する制度を確立し、且つ、漁船に関する試験を行い、もつて漁船の性能の向上を図り、あわせて漁業生産力の合理的発展に資 すること を定めた法律。
	寄生虫病の予防及び寄生虫病患者に対する適正な医療の普及を図ることによって、寄生虫病が個人的にも社会的にも害を及ぼすことを防止し、もって公共の福祉を増進 すること を目的として制定された法律である。

表 4.9 より例えば「すること」についての例文「つまり歪曲することを求めていないのではないか？」という文を考えてみる．この文は「歪曲する」という動詞について述べている文であり,「歪曲」という名詞について述べているわけではない．よってこの文中の「すること」は必要な表現であるといえる．そのため冗長でない文を正しく推定できているといえる．

第5章　関連研究

関連研究としては以下のものがある．大竹ら [6] は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 [7] 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6章　おわりに

本研究では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た．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ことがわかった．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 値で検出できた．この結果はベースラインの性能を上回った．本研究では,「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．

今後は，特定の表現ごとに機械学習する方法を多数の表現で試すとともに，多数の特定の表現での冗長な表現の検出により任意の文でのカバー率，つまり，任意の文で冗長な表現をどの程度検出できるかを調査したいと考えている．

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 の村田真樹教授に心から御礼申し上げます．また，本研究を進めるにあたり，御指導，御助言を頂きました，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．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, 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] 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.

[7] 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.

[8] 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.

[9] 村田真樹 (2001), “機械学習手法を用いた日本語格解析-教師信号借用型と非借用型、さらには併用型-”, 情報処理学会自然言語処理研究会 2001-NL-144, pp.113-120.