冗長な文の機械的分析と機械的検出

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」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．

2　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修正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．

機械学習法には，サポートベクターマシン (SVM)[*3]を用いる．カーネル関数は一次式を用いている．

3　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．本節では，この比較に基づく分析を行う．

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

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

表 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 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

表 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 4: 冗長な表現の修正例

修正前	修正後
ものである	削除
を行う	する
存在する	ある
可能である	できる

3.1 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．3.2 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[*4]にかけ単語単位に分割をする．分割した各データを比較し修正箇所の検出をする．修正個所の頻度を求め，どのような表現が冗長な文に頻出するかを調べる．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．

3.2 データ

ウィキペディア，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 1 と表 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する．

収集された正例，負例を用いて例 1 のような「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例 2 のような「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．

例 1．冗長性判定用データベース

冗長な文：「まず初めにマシンの点検を行う。」
冗長でない文：「鳥取大学で鳥取の歴史についての講演があった。」

[*4] ChaSen:http://ChaSen-legacy.sourceforge.jp/

例 2．冗長性修正文集合データベース

冗長な文：「まず初めにマシンの点検を行う。」
修正後の文：「初めにマシンを点検する。」

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 429 文，「冗長性修正文集合データベース」は冗長な文は 175 文であり，それを修正したものを合わせて合計 350 文を作成した．

3.3 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 と表 4 に結果を示す．

表 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 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 より，「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．

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．

4.1 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．

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

表 5: 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, ...

表 6: 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

表 7: 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

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「マシンの点検を行う」という文では表 5 のようになる.

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).

4.3 実験と結果

学習データとして 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．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 6 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.

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 7 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.

とはいえ，提案手法の性能は高いものではない.

5 冗長な文の検出 2

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，本節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．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．本節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.

表 8: 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

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 8 のようになる．この例での対象表現は「可能」である.

5.2 データ

3 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．正例の割合が 8 割以上の表現は，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.

5.3 実験と結果

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.

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

表 9: 各表現ごとでの冗長な文の検出性能

「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
「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
「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 10: 正しく判断できた例

正例
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
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
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
負例
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
文の成立について、山田は「陳述」という 用語を用いた。
つまり歪曲 すること を求めていないのではないか？

結果を表 9 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 10 に提案手法で正しく判断できた文の例を示す．

表 9 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．

6 関連研究

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．

7 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] 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.

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved