機械学習と冗長度を用いた冗長な文章の検出

都藤 俊輔 [*1]　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]{s082034,murata,tokuhisa}@ike.tottori-u.ac.jp
[*2] 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「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」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．上記のような文を冗長な文とし，冗長な文の自動修正へのアプローチを試みる．

文の改善の研究としては「誤字の修正・適切な語の選択」[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」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．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，本研究で取り扱う．

本研究は文章を作成する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．

本論文の主な主張点は以下である．

- 分析により冗長な文章の典型的な 3 つの分類を示した．
- 冗長な文章の検出において本稿で提案する冗長度が役立つことを確認した．

2　本研究の流れ

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 (3 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 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．実際に実験すると，冗長な文章の検出の際に冗長度が役立つことがわかった．

最後に，1 文での冗長な文の検出において冗長度が役立つかを確かめる実験を行う (5 節)．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．

3　冗長な文章の分析

本節では，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．

3.1　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．データは下に示す．

> **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。

上記のデータベースを実際に作成した．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．

3.2　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 (冗長文章) とそれを修正した文章 (修正文章) の例文の対 (例文対) も示している．

3.2.1　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。

3.2.2　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。

3.2.3　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　冗長な文章の自動検出

本節では，複数の文にまたがる冗長な文章の自動検出を試みる．

4.1　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．

4.1.1　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，2 値分類を機械学習で行うことで，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．

$$\text{冗長度 } x = \frac{N}{V}[V:\text{単語の異なり数}, N:\text{延べ単語数}] \quad (1)$$

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 毎に 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 単語毎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 毎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．

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，性能がそれ以上があ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．

4.1.2　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 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．

4.2　使用データ

使用するデータは 3.1 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　実験

4.3.1　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 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，4.1.1 節での素性番

号に対応している．

素性番号 3 が最も高い正解率を得た．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 2 に示す．同様にして表 3 と表 4 の実験を行った．表 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 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 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 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 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 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: 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 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)

4.3.2 冗長度に基づく手法での閾値調整

表 6 は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 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)

4.3.3 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，4.3.1 節の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 4.3.2 節で得られた閾値 1.4 を利用した．実験結果を表 7 に示す．

表 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	0.660(330/500)
冗長度	0.648(324/500)

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．

5 冗長な文の自動検出

前節のように，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 単語毎につなげた素性である．
○**素性番号 4(冗長度)** 冗長度を用いた情報．

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)

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]において冗長な文を正例，冗長でない文を負例として人手で判定し収集したデータベースを収集データという．データ例を下に示す．

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

冗長な文　　学校においては、授業料を徴収することができる。
冗長でない文　意味内容の詳細については定義と特徴の項目を参照されること。

収集したデータは正例と負例をあわせて 832 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．データ例を下に示す．

冗長な文　　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。
冗長でない文　体質を改善するのは、一朝一夕にはいかないです。

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．

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．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 8 と表 9 に示す．

表 8: 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

表 9: 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

作例データの各手法による検出結果を表 10 と表 11 に示す．

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．機械学習において，冗長度の素性を追加で用いることで性能が向上することが確認できた．

表 10: 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 11: 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

表 9，表 11 において，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わかる．

6　おわりに

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．これらのことにより冗長度が有用であるとわかった．

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.