日本語文章の難易度判定における文字の視覚的複雑さの有効性について

後藤　真由子　　指導教員：山村　毅

1　はじめに

インターネット等の普及による文字情報の蓄積に伴い，様々な情報が得られるようになってきた．文章の読み手や情報の探し手は，簡単な文章で情報を得たいというニーズや難しい文章から参考を得たいなど，文章の難易度に焦点を当てた要求を持っているといえる．これに対して，文章の書き手は，読んでもいたい対象が求める難易度の文章を書く必要がある．

このように，文章を読む側，書く側の双方において，文章の難易度を判定することは必要であるといえる．

文章の難しさは，様々な要因から判断することができる．その中でも，表面上の難しさは，見た目から感じるもの，つまり視覚的な情報であり，内容から感じる難しさよりも直感的に感じる難しさである．

本研究では，文章の視覚的情報に着目した日本語文章の難易度判定を行い，それが有効であるとすることを目的としている．

2　関連研究

日本語文章の難易度判定は，これまで数多く研究されてきているが，そのほとんどは，言語情報についての統計量を用いた難易度判定である．

佐藤ら [1] は，小学校〜大学の教科書を対象に，単語ユニグラムモデルを用いて，文字の出現確率を求め，学年 (小学校〜大学の学年に対応する 1〜13 の整数値) を判定をする手法を提案している．この手法は文節や単語を分けたり，内容を見たりしていないことが特徴的である．

長谷川ら [2] は，日本語文章の難易度に影響を与えるものとして，表面上，言葉，文構造，文脈があることを述べ，これらのうちの，前の 3 つに対して，漢字の割合，漢語表現の割合，平均文節数を提案している．また，難易度ごとに有効な特徴量が異なることに着目し，マハラノビス距離による最近傍法で国語教科書を対象に難易度判定 (小学校〜高校に対応する 1〜5 の整数値) を行っている．

3　文書の難易度と視覚的複雑さ

3.1　難易度要因

文章に難しさを生じさせる要因には，様々なものがある．長谷川ら [2] は，これには，表面上の難しさと，言葉の難しさ，文構造の難しさ，文脈の難しさの 4 つがあると述べている．このうち，表面上の難しさは，文章を見た時に感じる第一印象のような難しさである．具体的なものとしては，文章の長さや漢字の割合，段落が少ない，何ページに文章がわたっているか，文字の大きさ，1 ページあたりの文字数が多いなどがある．

3.2　漢字の割合の問題点

表面上の難しさの例として，漢字の割合がある．一般的に文章を見た時に漢字が多く使われていると難解な印象を受けることから，いくつかの難易度判定の研究で，用いられている特徴量である．以下の図 1 を考える．

A:世界一美しいぼくの村。
B:最初は監獄のそばに下宿した。

図 1　異なる難易度の例

A は小学校四年生の教科書，B は高校の教科書の文章中の一文である．それぞれの文について，句読点や記号を除いた漢字の割合を計算すると，A は 50%，B は 46% なるので，漢字の割合だけでは小学校 4 年の文 A の方が，高校の文 B よりも難しいと判断されてしまうことになる．B の方が難しいのは，「監獄」という比較的難しい漢字が含まれているためであるが，単純な漢字の割合では，これを考慮することが出来ないため，誤った判定をしてしまうことがある．

3.3　先行研究

著者ら [3] は，個々の漢字の難しさを考慮するために，文字の視覚的複雑さを用いた日本語文章の難易度判定の手法を提案している．すなわち，まず，テキストデータを画像データへ変換後，文字ごとに濃度 (文字内のピクセル値の平均) を求め，テキストにおける文字濃度ヒストグラムを作る．次に，これを特徴量として，各学年の平均の文字濃度分布をプロトタイプとした最近傍分類器で難易度 (学年) を判定する．距離関数には，KL 情報量を用いている．小学校〜高校の国語の教科書の物語文と説明文を対象に評価を行い，それぞれ分類正解率 51%，45% という結果を得ている．

4　視覚的複雑さの有効性についての実験

4.1　先行研究の問題点

先行研究の結果において，精度向上につながらないと考えられる 3 つの問題点があった．

1 つは，難易度が上がると増加する連語 (連続する漢字) を考慮しなかったことである．難易度が上がると，テキスト内の漢字の出現は増えていき，連語表現も出現回数も増えていくと思われる．この特徴から，連語を考慮することで，難易度を表現することができると予測できる．

2 つ目は，用いた分類器が 1 クラス 1 プロトタイプの最近傍法であることである．1 クラス 1 プロトタイプの最近傍法では，クラスの中でデータが球状に分布し，クラス間でデータが重ならないことを仮定しているが，実際の濃度ヒストグラムはこの仮定を満たしてない．

3 つ目は，濃度ヒストグラムを作る際，文字濃度を等間隔に区切って度数を数えていた (均一ビン幅) ことである．これは，どの濃度間隔も難易度判定に同じ程度の寄与をすることを仮定している．しかし，実際には，ほとんど寄与しない濃度間隔も存在すると考えられる．

以上述べた 3 つの原因に対して次節で対処法を提案する．

4.2　問題への対処

難易度の上昇に伴って増加する連語を考慮していないという問題点に対して，n-gram(ユニグラム，バイグラム，トライグラム) を用いた手法を提案する．この手法は，文字の並びを考慮した手法で，連続する文字濃度を特徴量とするものである．先行研究で用いた，1 文字の文字濃度のヒストグラムは，ユニグラムを用いた手法に相当する．

1 つの難易度を 1 つのプロトタイプで代表していることの問題点に対して，1 クラスあたり複数プロトタイプを用いる手法で解決できるのではないかと考える．クラスあたりに複数のプロトタイプを用いることで，1 つのクラスが各プロトタイプを含む閉領域で細かく分割される．今回，この手法では，全学習データ

をそのままプロトタイプとする方法をとる．

難易度判定に寄与する異なる文字の濃度間隔 (ビン) の違いに対処するために，不均一ビン幅を用いる手法を提案する．ビンは，データの量子化の話である．濃度ヒストグラム上にある，難易度判定に大きく寄与する濃度範囲に対しては，ビン幅を小さく，寄与の小さい濃度範囲に対してはビン幅を大きくすることで難易度をより表した濃度ヒストグラムが表現できると考える．

4.3 実験

4.3.1 実験データ

小学校〜高校の国語教科書の物語文・説明文に対して，テキストデータを IPA フォントを用いて 1 文字あたり 100×100 の画像データへ変換し，そこから文字ごとに濃度 (文字内のピクセルの値の平均値) を求めたものを用いる．文字内の各ピクセルは，0〜255 の値を持つので，文字の濃度は $0〜2^{16} - 1$ となる．また，難易度は，小学校 6 段階，中学校 3 段階，高校 1 段階の 10 段階とした．

4.3.2 実験方法

前節の実験データから物語文，説明文，これらを合わせた混合文の 3 つのデータセットを用意し，実験を行う．文字濃度を濃度ヒストグラムにする際，ビン幅は，256〜9984 の範囲で 256 ずつ変化させた．なお，n-gram と複数プロトタイプの手法は，均一なビン幅とした．

不均一ビン幅の定め方は，逐次的なビン幅のクラスタリングで実現し，2,796,160 通り[*1]の不均一ビン幅の正解率を測る．

難易度判定に用いる分類器は，各ビンの生起確率を用いたナイーブベイズ分類器である．n-gram と不均一ビン幅の手法の場合は，学年と作品で，複数プロトタイプの手法の場合は，作品間で類似度を見る．

4.3.3 実験結果

提案手法の最大正解率を表 1 に示す．ユニグラム，バイグラム，トライグラムの順に精度が向上することを予測したが，実際には，これらの間にはほとんど差がなかった．評価した手法の中で最大の正解率を示したのは不均一ビン幅を用いた手法であった．

表 1　手法における最大正解率 (%)

手法	物語文	説明文	混合文
ユニグラム	53	45	45
バイグラム	52	45	45
トライグラム	51	47	43
複数プロトタイプ	43	45	44
不均一ビン幅	63	58	54

4.3.4 実験結果に対する考察

n-gram 手法は，ユニグラム，バイグラム，トライグラムの手法を用いて連語を考慮するようにしても正解率に大きな変化は見られなかった．つまり，連語は，難易度を表す特徴量として適切でない事がわかる．

複数プロトタイプ手法は，従来研究よりも低い正解率を示した．このことから同じ難易度の作品は固まって存在しておらず，難易度のクラスは互いに重なって存在していることが言える．つまり，作品の間の類似度で判定を行うことは有効でないことが分かる．

[*1] $\sum_{i=0}^{255} {}_{256-i}C_2 = 2796160$.

最後に不均一ビン幅手法は，従来研究の均一ビン幅よりも高い正解率でを達成でき，全てのデータセットで 50% 以上であった．よって，濃度ヒストグラムには，難易度判定に有効な濃度範囲が存在することが分かる．

4.3.5 テキスト情報との比較

本研究の視覚的情報を用いた手法 (不均一ビン幅) は，テキスト情報を用いた難易度判定に関する従来研究 (長谷川ら [2]，山村 [4]，山村 [5]) と同程度の性能を発揮しているということができる．

表 2 は，テキスト情報を用いた難易度判定の 1 つである山村 [5] の結果 (説明文，混合文に対しては追加の評価を行った結果である) と本研究の結果と比較したものである．この表から，どのテキストにおいても，本研究の結果がテキスト情報を用いた場合と同程度の結果が得られていることが見てとれる．

表 2　山村 [5] と本研究結果の比較 (%)

手法	物語文	説明文	混合文
山村 [5]	61	57	56
不均一ビン幅	63	58	54

テキスト情報を利用する方法では，言語解析器 (形態素解析器など) を必要とするため，分類正解率が言語解析器の性能に左右されるという問題点がある．これに対して，本研究の場合，画像情報のみを用いているため，基本的に性能が安定しているという利点がある．

5 まとめ

難易度判定において視覚的情報を用いることの有効性を示すため，先行研究の問題点から n-gram，複数プロトタイプ，不均一ビン幅と 3 つの手法を提案した．

小学校から高校までの国語の教科書から物語文，説明文，2 つを合わせた混合文の 3 つのデータで提案した手法を評価した．最も高い正解率を示したのは不均一ビン幅の手法で，物語文で 63%，説明文 58%，混合文 54% という結果が得られた．これは，従来のテキスト情報を用いた難易度判定と，同程度の性能であった．

参考文献

[1] S.Sato,S.Matsuyoshi,and Y.Kondoh, “Automaticassessment of Japanese text readabiity based on a text book corpus,” LREC08,2008.

[2] 長谷川優, 山村毅: “マハラノビス距離を用いた日本語文章の難易度判定システムの提案”, 電子情報通信学会論文誌,Vol.J94-Dno.9pp.1589-1592,2011.

[3] 後藤真由子, 伊藤徹, 山村毅: “文字の視覚的複雑さを用いた日本語文章の難易度判定”, 電気関係学会東海支部連合大会講演論文集,L31, 中京大学,2014

[4] 山村毅: “日本語文書の難易度判定におけるテキスト統計量の有効性”, 信学論,Vol.J96-Dno.8pp.1952-1955,2013.

[5] 山村毅: “複数の判断基準を用いた日本語文章の難易度判定”, 電子情報通信学会論文誌,Vol.J97-Dno.5pp.1-63-1066,2014.