文字の視覚的複雑さを用いた日本語文章の難易度判定

情報科学科　伊藤 徹　　指導教員：山村 毅

1　はじめに

近年の情報化社会の発展により，誰もが多くの種類の文章を入手することが可能となった．文章の難易度は，使用される漢字や表現方法によって様々である．そこで，個人の読解力に応じた文章を入手できれば，情報収集などをより円滑に進める手助けとなることが期待できる．

本研究では，小学校～高校の国語教科書を対象に，小学校 6 段階，中学校 3 段階，高校 1 段階の計 10 段階を難易度のランクとして，文章の難易度判定手法を提案する．

2　従来研究

文章の難易度判定に関しては，文章を構成している文や文節，単語，文字などについての統計量を用いて文章を特徴付けることにより，難易度を判定する手法がこれまでに数多く提案されている [1]～[3].

最近では，長谷川ら [2] が，難易度ごとに特徴量の分散が異なることに着目し，日本語の国語教科書の文章を対象に，平均文節数，漢字割合，漢語の割合を特徴量として，マハラノビス距離による最近傍法で難易度を判定するシステムを提案している．

3　文字の視覚的複雑さと判定手法

3.1　文字の視覚的複雑さ

前節で述べたように，従来研究では漢字の割合などの特徴量を用いて文章の難易度判定を行っているが，我々が文章を読んだときに難しいと感じる要素はそのような特徴量だけでなく，単に文章を眺めたときの視覚的な複雑さにもあるのではないかと考える．すなわち，濃いと感じた文章は難しく，薄いと感じた文章は簡単であるというものである．そこで本研究では，文章の視覚的複雑さに着目した日本語文章の難易度判定を行う．

3.2　画像処理を用いた文章の難易度判定

文字を画像に変換し，色の平均を取り，文字の濃度とする．文字画像のピクセルは，1～65535 の色（白または黒）で表されるので，文字の濃度は 1～65535 までの値になる（値が大きくなるほど，より濃度の高い文字であることを表す）．得られた文字の濃度をもとに文章の難易度を判定する．

各作品（文章）について，文字の濃度分布を求め，これを特徴量として難易度判定を行う．すなわち，ある作品のヒストグラムを各学年のヒストグラムと比較することで難易度判定を行う（似た濃度分布を示す学年の作品であると判定する）．

ヒストグラムの類似性は，次に示す KL 情報量（Kullback-Leibler divergence）を用いる．

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (1)$$

ここで P, Q はそれぞれ，作品のヒストグラム，学年の全作品のヒストグラムである．

4　評価実験

小学校～高校の物語文 145 作品を対象として，ヒストグラムの刻み幅を様々に変えて正解率を調査した．正解率の計算には，10 分割交差検定を用いている．結果を表 1 に示す．この表から分かるように，ヒストグラムの刻み幅が 5000 と 9000 のときに最大正解率は 50 %になった．

表 1　濃度の刻み幅による正解率の変化

刻み幅	正解率 (%)	刻み幅	正解率 (%)
100	32	5500	46
500	42	6000	45
1000	42	6500	43
1500	43	7000	37
2000	45	7500	39
2500	43	8000	43
3000	47	8500	46
3500	48	**9000**	**50**
4000	43	9500	46
4500	48	10000	46
5000	**50**	15000	39

山村 [5] によれば，文字の統計を用いた方法では，「漢字の割合」で 41 %，「漢字の種類数 ÷ 文字の種類数」で 44 %の精度であるので，本研究の手法が有効であることが分かる．

5　まとめ

画像処理によって文字の濃度を求め，それを用いて文章の難易度を判定する手法を提案した．

小学校～高校の国語教科書の物語文を対象に，KL 情報量を用いて文章の難易度判定を行った．ヒストグラムの刻み幅が 5000 と 9000 のときに最大正解率は 50 %になった．

今後の課題として，複数の特徴量（文字，単語などの特徴）と併用することが挙げられる．また，ヒストグラム全体ではなく，一部（濃度の高い方）を使うことが挙げられる（低学年の作品にも高学年の作品にも濃度の低い文字は存在するが，濃度の高い文字は高学年の作品に多く出現する傾向があるため）．

参考文献

[1] S. Sato, S. Matsuyoshi, and Y. Kondoh, “ Automatic assessment of Japanese text readability based on a textbook corpus, ” LREC-08, 2008.

[2] 柴崎秀子, 原信一郎,“ 12 学年を難易度尺とする日本語リーダビリティー判定式, ”計量国語学, Vol.27, no.6, pp.215-232, 2010.

[3] 建石由佳，小野芳彦，山田尚勇, “ 日本文の読みさすさの評価式, “ 情報処理学会文書処理とヒューマンインターフェース研究会資料，HI18-4, pp.1-8, 1988.

[4] 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-D No.9 pp.1589-1592，2011.

[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，印刷中