# 日本語文章の難易度判定における視覚的複雑さの有効性について

情報科学科　水谷 康太　　指導教員：山村 毅

## 1　はじめに

今日，様々な文章を誰もが容易に手に入れられるようになってきた．文章には難易度のレベルがあり，自分にあった難易度でないと，読み解くのに時間や手間を要したり，著者の伝えたいこととは異なった解釈をされてしてしまう可能性がある．そのため文章の読み手には，自分にあった難易度の文章を選びたいという要求があり，一方，文章の書き手には，書いた文章が意図した難易度になっているかを知りたいという要求がある．ここに文章の難易度判定の必要性があると言えよう．

後藤 [1] の研究では文章を読んだ時の難しさはテキスト情報だけでなく，文章を眺めたときの視覚的複雑さにあると考え，文字を画像に変換したときの濃度値を求め，各文章のヒストグラムによる難易度判定を提案している．また，水谷 [2] は，不均一のビン幅のヒストグラムを用いて難易度判定を行い，結果として物語文で 63%，説明文で 58%，全体で 54% の正解率を得ている．

本研究では先行研究の視覚的複雑さの有効性について調べる．

## 2　実験と評価

### 2.1　ひらがなの品詞の除去

後藤の研究では文章を 1 文字ごとそのまま使用していたが、ヒストグラムにおいて難易度の差が見られるのは濃度の高い漢字の部分である．全学年に共通して現れる，記号やひらがななどの濃度の低い部分を取り除くことで，より難易度を反映したヒストグラムを構成できると考えられる．そこで記号及び通常ひらがなで書かれる品詞 (以下，ひらがな品詞) である助詞，助動詞，接続詞，連体詞，感動詞を取り除き，後藤と同様の方法で判定を行う．

### 2.2　異なる分類器・濃度特徴量

後藤の研究では分類器をナイーブベイズ分類器に限定して行っており，その特徴量はヒストグラムを用いていた．いくつかの分類器で判定を行い，また複数の特徴量を組み合わせることで精度が向上することが考えられる．そこでヒストグラムに加えて，高濃度文字の割合や分散，品詞の平均濃度を利用する．

### 2.3　テキスト情報との組み合わせ

視覚的複雑さとテキスト情報を組み合わせることで精度が向上することが考えられる．ヒストグラムを一つの特徴量とし，これに山村 [3] のテキスト情報と組み合わせて判定を行う．

## 3　結果と考察

### 3.1　ひらがな品詞の除去

文章から通常ひらがなで書かれる品詞を取り除くことによって，より難易度を反映したヒストグラムになると考えたが，実際には精度の向上は見られず，5 つの品詞を取り除いた場合でも正解率は物語文で 63%，説明文で 57% だった．この理由としてこれらの品詞はどの難易度の文章でも同程度の割合に含まれていたために，除去しても精度に影響しなかったと考えられる．

### 3.2　異なる分類器・濃度特徴量

分類器を変えても精度はそれほど変化せず，物語文ではヒストグラムに対し SVM を用いたもので 48%，説明文では漢字割合等にナイーブベイズを用いたもので 47% が最大の正解率であった．各学年を表す特徴量は図 1 のように rank6〜9 の領域が完全に混ざってしまっていた．このような特徴量では分類器を変えても各学年の領域に分割することができず，精度向上にはつながらなかったと考えられる．

### 3.3　テキスト情報との組み合わせ

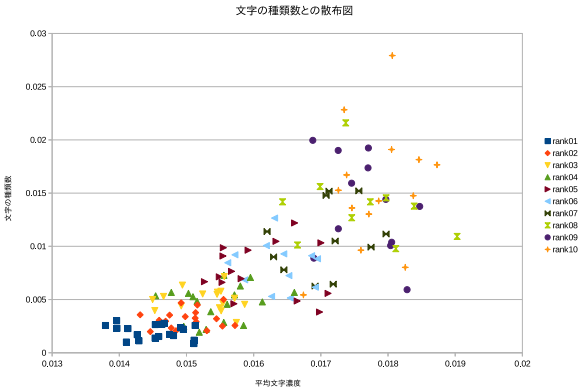


図 1　単語の種類数との散布図

テキスト情報と組み合わせても精度の向上は見られず，物語文では SVM を用いた場合で 50%，説明文ではナイーブベイズを用いた場合で 44% が最大の正解率となった．視覚的複雑さとテキスト情報の 2 つの特徴量について調べると，図 1 のように平均文字濃度といくつかの特徴量の間に正の相関があることが見受けられた．漢字の種類数との間に強い相関が見られるのは当然であるが，文字や単語の種類数との間にも強い相関が見られていた．このことから rank が上がるに連れて増えた文字や単語はそのほとんどが漢字で構成されていることが考えられる．つまり一見独立していると思われる，視覚的複雑さとテキスト情報は実際には密接に関係しているため，これらを組み合わせても精度が向上しなかったことが考えられる．

## 4　まとめ

後藤らの研究をもとに，ひらがな品詞を除いた濃度ヒストグラムを用いた方法，異なる分類器・濃度特徴量を用いた方法，及びテキスト情報を併用した方法を用いて，視覚的複雑さの有効性を検証した．視覚的複雑さは，おおよその難易度判定を行うことはできるが，細かい判定には不向きであることが分かった．

## 参考文献

[1] 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,L3-1，2014.

[2] 水谷康太, 後藤真由子, 山村毅:”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”, 電気関係学会東海支部連合大会講演論文集,C1-1,2017.

[3] 山村毅:”日本語文章の難易度判定におけるテキスト統計量の有効性”, 電子 情報通信学会論文誌,Vol.J97-D,No.5,pp.1063-1066,2014.