漢字難易度を素性としたベイズ推定による新聞記事の難易度推定

友廣　翔太[1]　　小林 伸行[2]　　椎名 広光[3]

[1] 岡山理科大学大学院 総合情報研究科 情報科学専攻

[2] 山陽学園大学 総合人間学部 生活心理学科

[3] 岡山理科大学 総合情報学部 情報科学科

i12im07ts@std.ous.ac.jp[1], koba_nob@sguc.ac.jp[2], shiina@mis.ous.ac.jp[3]

1　はじめに

複数の異なる読み手が同じ文書を読むとき，例えば，日本人学生と日本語を勉強中の外国人留学生とでは，文書に対する難易度（難しさ）の評価は異なるであろうと考えられる．また，日本人同士でも小学生と大学生，あるいは大学生同士であったとしても，文書に対する難易度の評価は少なからず異なるであろう．読み手が感じる文書の難易度は，読み手の語学力，背景知識などをはじめとした様々な要因によって異なり，また，その評価基準も人それぞれである．本研究では，こうした読み手によって異なる文書の難易度を判定する手法を提案する．本提案手法では，読み手による文書の難易度評価を事前情報として予め蓄積しておくことで，読み手の文書難易度評価の基準となる文字素性，指標を抽出する．難易度の判定では，文書分類などの目的で利用されるナイーブベイズ分類を行う．ナイーブベイズ分類では文書中に出現する単語を素性として利用することが多いが，本提案手法では文書中に出現する文字を漢字級別などの文字素性に分けることで，文字単位の素性を扱う．また，難易度の判定に利用する級別漢字などの文字素性の組み合せも個人化を図ることで，難易度判定の精度向上を図っている．

2　読み手の文書に対する難易度の評価

読み手の文書に対する難易度の評価は，語学力などの様々な要因によって異なる．本研究では，予め読み手による文書の難易度評価を 10 段階の難易度評価で行い，事前情報として取り入れた．これをナイーブベイズ分類 [1] における事前確率として難易度判定に利用した．

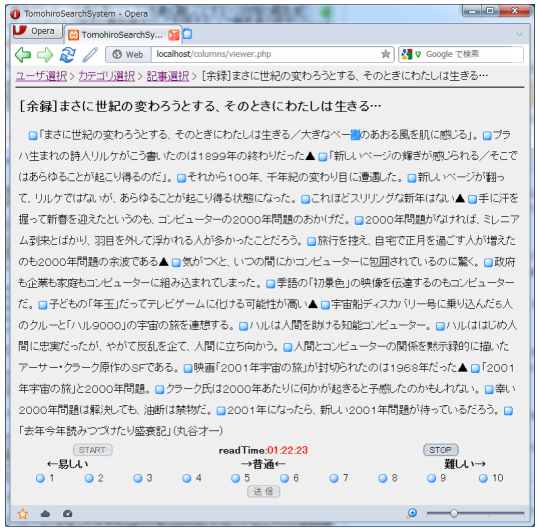

図 1: 読み手による難易度評価システムの実行画面

3　漢字の漢検級別

本研究における漢字の漢検級別には，日本漢字能力検定協会 [2] の定める漢字の級別を利用している．漢字級別は，1 級，準 1 級，…，10 級の 12 級別に分けられており，級数が小さくなるにつれて漢検が難解であると定める漢字が多く含まれている．本研究では，この漢検級別のうち，日本漢字能力検定協会が級別を公表している 2136 字（2 級から 10 級に相当）の漢検級別を素性として利用した．

4　難易度判定システムの概要

読み手による文書の難易度評価を入力するためのシステムの実行画面を図 1 に示す．読み手はこのシステムを通して文書の難易度を評価する．また，難易度判定システムの処理の流れを図 2 に示す．

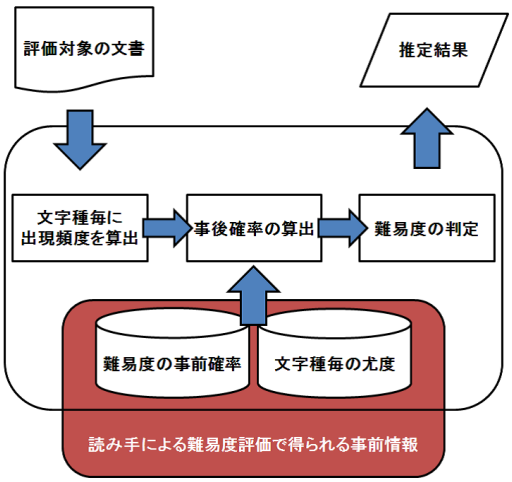

図 2: 難易度判定システムの処理の流れ

本難易度判定システムでは，読み手によって入力された文書の難易度，及びその文書中に出現する文字素性を事前情報として予め蓄積した上で難易度の判定を行う．始めに，判定対象の文書中に出現する文字を級別漢字などの文字素性に分けることで，各文字素性の出現頻度を求める．次に，蓄積された事前情報を利用して，判定対象の文書を読み手がどの難易度で評価するかを確率的に計算する．そして，すべての難易度の中で最も読み手によって評価される確率が高いと判定された難易度を文書の判定結果として出力する．

5　難易度の判定手法

本難易度判定システムは，判定対象の文書を文字単位に分割してナイーブベイズ分類を行う．本難易度判定システムにおける事前確率，尤度，難易度の判定方法について以下に述べる．

5.1　事前確率

本難易度判定システムでは，読み手による文書の難易度評価割合を事前確率として利用する．本研究では難易度のクラスとして 10 段階の難易度を用意した．各難易度を $c_i(i = 1, 2, \cdots, 10)$ として，難易度 c_i の事前確率を次のように計算する．

$$P(c_i) = \frac{\text{難易度 } c_i \text{の文書の件数}}{\text{すべての文書の件数}}$$

5.2　尤度

本難易度判定システムでは，尤度として読み手によって難易度評価をされた文書中の 2 級から 10 級の各級別漢字とひらがな，カタカナ，数字，アルファベット，その他の合せて 15 の文字素性の出現確率を用いた．15 の文字素性のうち，難易度の判定に利用する文字素性の集合を D とすると ($D \subset \{$2 級漢字，準 2 級漢字，…，その他 $\}$)，難易度 c_i の文字素性 $d(d \in D)$ の尤度を次の通り求める．

$$P(d|c_i) = \frac{\text{難易度 } c_i \text{中の文字素性 } d \text{ の総和}}{\text{難易度 } c_i \text{中の文字の総和}}$$

ナイーブベイズ分類では各素性の生起確率を独立と仮定して，文書全体の確率を各素性の尤度の積で求める．本難易度判定システムでも，各文字素性の出現確率は独立であると仮定して，文書全体の尤度を各文字素性の尤度の積で算出している．

判定対象の文書を T，文書 T 中の文字素性 d の出現頻度を t_d と定義する．文書 T が難易度 c_i と評価される尤度を次式で求める．

$$P(T|c_i) = \prod_{d \in D} P(d|c_i)^{t_d}$$

5.3　難易度の判定

ベイズの定理より，文書 T が難易度 c_i と判定される確率 $P(c_i|T)$ を，事前確率と尤度を用いて次式求める．

$$P(c_i|T) = \frac{P(c_i)P(T|c_i)}{P(T)}$$

分母の $P(T)$ は，すべての難易度において共通の定数であるため難易度の判定の過程では無視する．事後確率を最大にする難易度 $\tilde{c}$ を次式で求める．

$$\tilde{c} = \arg\max_{c_i} P(c_i)P(T|c_i)$$

以上の過程により求められた難易度 $\tilde{c}$ を本難易度判定システムの判定結果とする．

6　文字素性の組み合せ

本難易度判定システムでは漢字級とひらがな，カタカナ，数字，アルファベット，その他の計 15 の文字素性の尤度を任意に組み合わせて判定対象の文書の尤度を求める．そのため，本難易度判定システムでは全部

表 1: 順位平均によって決定した文字素性の組み合せと各指標

被験者	文字素性の組み合せ	正答率	順位	二乗誤差平均	順位	JS ダイバージェンス	順位	平均順位
学生 A	3 級，5 級，10 級，ひらがな，その他	0.54	9	1.12	235	0.581969804	441	228.34
学生 B	2 級，準 2 級，5 級，6 級，8 級，10 級，カタカナ，アルファベット	0.32	382	2.62	59	2.52231291	367	269.34
学生 C	2 級，準 2 級，7 級，9 級，10 級，アルファベット	0.52	2	1.4	7	0.136395697	1	3.34

で $2^{15}-1$ 通りの判定結果が存在しうる．しかしながら，本難易度判定システムの判定結果は 1 つでなければならないので，$2^{15}-1$ 通りの組み合せの中からより判定の精度が高い組み合せを選択する必要がある．そこで，本研究では,「正答率」,「二乗誤差平均」,「JS ダイバージェンス」の 3 つの指標の順位を利用することで使用する文字素性の尤度を選択する．各指標の詳細を説明する．

6.1 正答率

本研究では，正答率を難易度判定システムによる判定が読み手による難易度評価と一致する確率としている．正答率はその判定対象の文書によって,「未知の文書の正答率」と「既知の文書の正答率」の 2 種類に分かれる.「既知の文書の正答率」は，事前情報として読み手の難易度評価を入力された文書を本難易度判定システムで再判定したときの正答率であり，文字素性の組み合せの選択にはこちらの正答率を用いる．一方,「未知の文書の正答率」は，事前情報として入力されていない未知の文書に対して難易度判定を行ったときの正答率であり，この正答率を高めることが本難易度判定システムの目標である．

6.2 二乗誤差平均

正答率が難易度判定の一致の割合を測るための指標であったのに対して，二乗誤差平均は，難易度判定システムによる判定のズレを測る指標である．この値が小さい文字素性の組み合せほど，難易度判定の精度が高い．なお，二乗誤差平均を計算する際，難易度 c_i と c_{i+1} 間の距離を 1 として扱う．

6.3 JS ダイバージェンス

より読み手の評価に近い難易度の判定を行うためには，判定の難易度分布も読み手の評価に近づける必要がある．JS ダイバージェンス [3]（ジェンセン・シャノン・ダイバージェンス）は，2 つの確率分布がどれくらい違うかを測るための指標である．本研究においては，読み手が事前情報として入力した難易度評価の分布と難易度判定システムが判定する難易度の分布の差を測るのに利用した．比較する 2 つの確率分布を P と Q として，2 つの確率分布の JS ダイバージェンスを次式に示す.

$$D_{JS}(P \parallel Q) = \frac{1}{2}\left(\sum_x P(x)\log\frac{P(x)}{\frac{P(x)+Q(x)}{2}} + Q(x)\log\frac{Q(x)}{\frac{P(x)+Q(x)}{2}}\right)$$

6.4 文字素性の組み合せの選択

文字素性の組み合せの選択には,「正答率」「二乗誤差平均」「JS ダイバージェンス」の順位を利用する．この順位は，$2^{15}-1$ 通りの文字素性の組み合せと比較した順位であり，正答率は指標値の降順に，二乗誤差平均と JS ダイバージェンスは指標値の昇順に順位付けした．本研究では，3 つの指標の順位の平均値が最も良いものを最も精度の高い文字素性の組み合せとして選択した．

7 読み手の違いによる難易度判定の差異

読み手の違いによる難易度判定の差異を確かめるために 2 つの難易度判定の実験を行った．なお，本実験で扱う文書は，毎日新聞 2000 年版 [4] に掲載されたコラムから選んでいる．

7.1 実験 1

実験 1 では，事前情報として入力された文書の難易度を本難易度判定システムで再判定することで，最も精度の高い文字素性の組み合せを決める．本実験では，学生 3 人を被験者として，50 件の文書の難易度を事前情報として入力した．そして,「正答率」「二乗誤差

表 2: 未知の文書に対する難易度判定結果

	正答率	二乗誤差	JS ダイバージェンス
学生 A	0.32	1.48	0.530128371
学生 B	0.1	4.88	3.248624412
学生 C	0.42	1.48	1.179140185

表 3: 読み手の難易度評価の分布

難易度	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
学生 A	0	0	0	9	13	18	8	2	0	0
学生 B	0	1	5	9	10	7	13	3	2	0
学生 C	0	0	3	14	22	3	7	1	0	0

平均」「JS ダイバージェンス」の平均順位を求めることで文字素性の組み合せを決定した．本手法により選択された文字素性の組み合せと各指標値の順位を表 1 に示す．

学生 A，学生 C の正答率は，それぞれ 0.54，0.52 と 5 割を越えており，二乗誤差平均，JS ダイバージェンスに関しても良好な指標値を示す文字素性の組み合せが選択された．一方，学生 B は 3 つの指標すべてが他の 2 人の値と比較して低めな文字素性の組み合せが選択されている．

7.2 実験 2

実験 2 では，実験 1 で選択した文字素性の組み合せを利用して，事前情報として入力されていない未知の文書 50 件を判定対象として難易度判定した．実験 2 の結果を表 2 に示す．難易度が既知の文書を判定した実験 1 と比較すると 3 人とも正答率が下がる結果となった．また，学生 A と学生 C に関しては，二乗誤差平均と JS ダイバージェンスともに実験 1 の結果と比較してそれほど大きな差は無かった．しかしながら，学生 B は 3 つの指標ともすべて悪くなるという結果となった．

7.3 考察

実験 1，実験 2 の結果より，学生 A と学生 C に関しては実験 2 で正答率が下がったものの，二乗誤差平均と JS ダイバージェンスについては大きな差は無く精度の高い文字素性の組み合せを選択することができた考える．一方，実験 1 の時点で精度の低かった学生 B の文字素性の組み合せは実験 2 ですべての指標値とも大きく下げる結果となった．この結果から事前情報として入力した文書の難易度評価の分布について調査した．その結果，学生 A は c_4 から c_8 の 5 段階，学生 C は c_3 から c_8 の 6 段階の難易度の範囲で評価しているのに対して，学生 B は c_2 から c_9 の広い範囲で難易度評価をしていることが分かった．事前情報として入力された学生ごとの難易度評価の分布を表 3 に示す．他の 2 人と比較して難易度選択の範囲が広く，難易度の事前確率がバラついてしまったことが学生 B の難易度判定の精度低下の原因であると考えられる．

8 まとめと今後の課題

本論文では，読み手による文書の難易度評価に基づく難易度判定システムの作成手法について提案した．本難易度判定システムを用いた判定実験では，使用する文字素性以外に事前情報として入力する難易度評価の範囲にも依存しており，判定精度に大きな影響する可能性があることが分かった．

本実験では，文書の難易度として 10 段階の難易度を想定したが，被験者への聞き取り調査の中には，10 段階は多く評価基準が決めづらいため統一的な難易度評価が出来ないという意見もあった．また，読み手によって難易度評価の幅にも差異が生じ，難易度判定の精度の大きな影響を及ぼした．今後の課題としては，難易度範囲を 3 段階から 5 段階程度に減らすことで難易度判定の精度向上を図り，読み手の違いによる難易度判定の違いについても研究を進める必要がある．

参考文献

[1] *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] 財団法人 日本漢字能力試験 : http://www.kanken.or.jp

[3] 高村大也，奥村学，言語処理のための機械学習入門，コロナ社，2010.

[4] 毎日新聞社，CD- 毎日新聞 2000 データ集，日外アソシエーツ，2000.