教科書コーパスを用いた日本語テキストの難易度推定

近藤　陽介[†1]　　松吉　俊[†1,†2]　　佐藤　理史[†1]

[†1] 名古屋大学大学院 工学研究科, [†2] 京都大学大学院 情報学研究科

1. はじめに

テキストは，人々の生活の中で幅広く用いられている情報伝達手段である．情報伝達に用いられるテキストの中には，例えば，人の生死や安全に関わる情報を伝えるテキストのように,「分かりやすい」ことがとりわけ求められるものがある．しかし，情報をどの程度の難しさで記述するかは，テキストの書き手に委ねられており，難しさの基準も人によって異なると考えられる．そこで，計算機を用いて日本語テキストの難しさを推定することを考える．テキストの書き手に「難しさの客観的評価」を提供することは，円滑な情報伝達を実現するための計算機支援の一形態となる．このような支援を実現するためには，日本語テキストの難しさを測るための基準や方法が必要となる．

テキストの難しさに関する研究は，英語に対しては1920年代から，日本語に対しては 1940 年代から行われている [1)]．英語に対しては，これまでに，Flesch Reading Ease や Kincaid Grade Level など，多くの難易度算定公式が提案されており，実際に，読解教材の難易度の推定などに広く用いられている．一方，日本語に対しては，これまでにいくつかの難易度算定公式が提案されているが [2)]，実用には至っていない．

このような背景から，我々は昨年より日本語テキストの難易度を推定する方法の研究を行っている．昨年は，社会科の教科書と新聞の社説を規準コーパスとして用い，中学・高校・一般の 3 つの区分の難易度を推定するシステムを構築した [3)]．しかし，中学・高校の規準コーパスの内容が社会科に偏っており，テキストの内容（分野）がテキストの難易度以上に強く影響を及ぼすことが分かった．そこで，新たに英語を除く全教科の教科書をテキスト収集源として規準コーパスを作成し，それを利用して難易度推定システムを再構築した．

2. 難易度区分と規準コーパス

2.1 難易度の区分

本研究では，テキストの難易度を表す区分として，学年区分を使用する．これは，Reading-age という考え方に基づくものであり，テキストの難易度のレベルが直感的に分かりやすく，実用的な区分であると考えられる．今回作成したシステムでは，小学 1 年から高校 3 年までの 12 学年に，大学を加えた 13 の区分で，テキストの難易度を表すこととした．

2.2 規準コーパス

難易度推定を実現するためには，その規準となるコーパスが必要である．これを規準コーパスと呼ぶ．規準コーパスに含まれるそれぞれテキストは，その難易度が既知でなくてはならない．さらに，規準コーパスは，前述の 13 の難易度区分のテキストを含んでいることが必要である．本研究では，規準コーパスのテキスト収集源として，教科書を用いる．小・中・高で用いられる各教科書は，その対象学年が比較的明確である．さらに，それらの教科書は文部科学省の検定を受けており，用いられている言語表現も，対象学年に適した難しさに調整されていると考えられる．すなわち，教科書を利用することで，小学校から高校までの幅広い学年において，難易度が既知であるテキストが入手可能となる．

このような考えに基づき，本研究では，規準コーパスとして，小学校から大学の教科書を収録対象とした教科書コーパスを設計・構築した [4)]．具体的には，愛知県名古屋市で使用されている小・中・高の全学年・全教科の教科書を 1 冊づつ入手し，英語を除く全教科の 111 冊の教科書からサンプルテキストを抽出し，1167 サンプル，728,002 文字のテキストを電子化した．大学の規準テキストには，実際に大学の教養課程の講義で使用されている指定教科書を採用した．国語や数学など高校の各教科に対応する分野の教科書 16 冊を選定し，それらから，311 サンプル，345,261 文字のテキストを抽出して電子化した．

3. 多項ナイーブベイズ分類を用いた難易度推定

3.1 フレームワーク

Collins-Thompson ら [5)] は，文書分類の枠組を応用した，英語テキストの難易度推定手法を提案している．この方法では，あらかじめ，N 個の難易度クラス $G_i(i = 1, 2, ..., N)$ のそれぞれに対応する，N 個の言語モデル M_i を構築しておく．ここで，言語モデル M_i は，難易度が G_i であることが既知であるテキストの集合から構築した，単語 unigram モデルである．テキスト T の難易度を推定する際は，まず，各言語モデル M_i に対して，次式で定義される尤度を計算する．

$$L(M_i|T) = \sum_{w \in T} C(w) \log P(w|M_i) \qquad (1)$$

ここで，w はテキスト T に含まれる異なり単語，$C(w)$

はテキスト T における単語 w の出現回数，$P(w|M_i)$ は言語モデル M_i における単語 w の生起確率である．

こうして得られた N 個の尤度のうち，最大の尤度をとる言語モデル M_i を求め，これに対応する難易度 G_i を，推定結果として出力する．

3.2 日本語テキストの難易度推定

日本語テキストの難易度推定では，単語 unigram モデルに代えて，文字 unigram モデルを用いる [3]．この理由として，日本語と英語の差異に起因する，次の 2 点が挙げられる．

(1) 日本語ではテキストが単語に分かち書きされていない．日本語テキストの単語を計算機で扱うには，形態素解析器を利用して単語分割を行う必要がある．しかし，常に正しい単語分割結果が得られるという保証は無く，単語 unigram を用いる場合は，解析誤りの影響を受けることになる．それに対し，文字 unigram を用いれば，解析誤りを考慮しなくてよい．

(2) 日本語では，漢字 1 字を疑似的な単語とみなすことができる．

4. 日本語テキストの難易度推定システム

本研究で作成した難易度推定システムの概要を図 1 に示す．システムは，入力された日本語テキスト T に対して，小学 1 年から大学までの 13 の難易度クラス $G_i(i = 1, 2, ..., 13)$ の中から，適当と判断される難易度クラスを 1 つ決定し出力する．

4.1 言語モデルの構築

2 章で述べた教科書コーパスを用いて，13 の難易度クラスのそれぞれに対して文字 unigram モデルを構築した．本研究では，コーパス中に現れる文字のうち，ひらがな・カタカナ・漢字のみを対象とし，それ以外の数字や記号，アルファベットは全て無視する．言語モデル M_i における，文字 x が生起する確率 $P(x|M_i)$ は次式で求める．

$$P(x|M_i) = \frac{C(x, D_i)}{\sum_{z \in D_i} C(z, D_i)} \tag{2}$$

ここで，D_i は，難易度クラス G_i が付与されたテキストの集合（学習テキスト）であり，z は学習テキスト D_i に含まれる異なり文字（ひらがな・カタカナ・漢字のみ），$C(z, D_i)$ は D_i における文字 z の出現回数である．

4.2 尤度計算と難易度の決定

テキスト T の難易度を推定するため，各難易度クラス G_i の言語モデル M_i に対し，次式を用いて尤度を計算する．

$$L(M_i|T) = \sum_{z \in T} C(z, T) \log P(z|M_i) \tag{3}$$

こうして得られる 13 個の尤度のうち，最大の尤度をとる言語モデル M_i を求め，これに対応する難易度 G_i を，推定結果として出力する．

4.3 未知文字の生起確率

言語モデル M_i の学習テキストに出現しない文字がテキスト T に含まれている場合，言語モデル M_i におけるその文字の生起確率は 0 であるので，式 (3) の尤度を計算することができない．そこで，学習テキストに出現しない文字に対して，次の 2 つの方針を定め，対処した．

方針 1 13 の難易度クラスの全ての学習テキストに出現しない文字は，尤度計算の対象から除外する

方針 2 いずれかのクラスの学習テキストに出現する文字は，生起確率の値を補正する

方針 2 の補正方法として，次の 2 つの手法を実装した．

補正手法 1：小さな定数 文字 x の生起確率 $P(x|M_i)$ が 0 となる場合，その値を，小さな正の定数 ε で置き換える．

補正手法 2：線形補間 難易度クラスには，小学 1 年のクラスが最も易しく，大学のクラスが最も難しいという順序関係が存在する．さらに，隣接するクラス間では文字の出現傾向が似ていることが期待される．このような理由から，確率 0 となる値を，隣接する難易度クラスの生起確率の平均値で置き換える．

$$P(x|M_i) = \frac{P(x|M_{i-1}) + P(x|M_{i+1})}{2} \tag{4}$$

$P(x|M_i)$ が 0 で，かつ $P(x|M_{i-1})$ と $P(x|M_{i+1})$ のいずれかが 0 でない言語モデル M_i に対して，学年が高いほうから順に式 (4) により生起確率を補間する．この操作を，生起確率が 0 であるクラスが無くなるまで繰り返し適用する．

4.4 難易度クラス間でのスムージング

先述のように，難易度クラスには順序関係が存在し，かつ隣接するクラス間では文字の統計的特徴が近いことが予想される．これらのことから，生起確率 $P(x|M_i)$ や尤度 $L(M_i|T)$ の値を，クラス間でスムージングすることにより，推定精度が向上する可能性がある．そこで，次の 2 つのスムージング手法を実装した．

スムージング手法 1：生起確率のガウス回帰

補正手法 1 を採用した場合，生起確率 $P(x|M_i)$ に対して，Gaussian-kernel を用いた回帰を適用する．すなわち，次式を用いて補正値 $\hat{P}(x|M_i)$ を計算する．

$$\hat{P}(x|M_i) = \frac{\sum_{j \in A_i} K(i, j) P(x|M_i)}{\sum_{j \in A_i} K(i, j)} \tag{5}$$
$$(A_i = \{k : | \, k - i \, | \leq h, 1 \leq k \leq N\})$$

ここで，h はバンド幅，N は難易度クラスの数を表す．$K(i, j)$ は次式で定義されるガウス関数である．

$$K(i, j) = \exp\left(\frac{-(i - j)^2}{2\sigma^2}\right) \tag{6}$$

この式で，σ はパラメータである．

スムージング手法 2：尤度の多項式回帰

補正手法 2 を採用した場合，尤度 $L(M_i|T)$ に対して，2 次または 3 次の多項式を用いた回帰を適用す

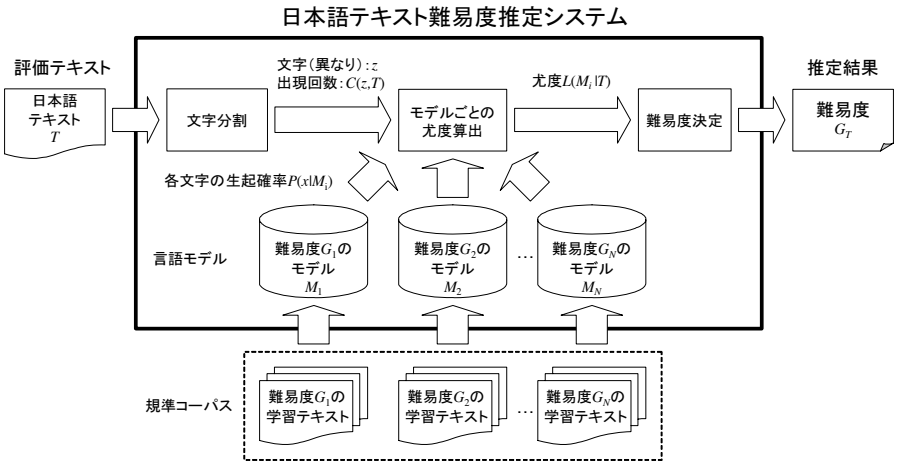

図 1 難易度推定システムの概要

表 1 補正・スムージング手法ごとの相関係数と RMSE

補正手法	スムージング手法	相関係数	RMSE
小さな定数	なし	0.895	1.638
	Gaussian-kernel 回帰	**0.915**	**1.482**
線形補間	なし	0.900	1.626
	2 次曲線回帰	0.886	1.794
	3 次曲線回帰	0.898	1.701
建石らの公式		-0.793	-
建石らの公式（簡略版）		-0.783	-

る．なお，難易度クラス間の距離は全て等間隔であると仮定する．

どちらのスムージング手法も 13 の難易度クラスに対するスムージングではあるが，スムージング手法 1 は生起確率 $P(x|M_i)$ に対するスムージング，スムージング手法 2 は尤度 $L(M_i|T)$ に対するスムージングである．

5. 実　　験

5.1 評　　価

本研究で提案する難易度推定手法の評価実験を行った．評価実験では，先述の教科書コーパスを用いて，Leave-one-out 交差検定を行った．評価指標として，テキストに付与されている難易度と推定された難易度との相関係数及び RMSE(Root Mean Square Error，平均二乗平方根誤差) を用いた．補正手法 1 において未知文字の生起確率として与える定数 ε としては，最も大きなサイズの学習テキストを持つ大学の言語モデルにおける文字の最小の生起確率を，100 で除した値を用いた．スムージング手法 1 を実行するには，2 つのパラメータ h と σ を指定する必要がある．予備実験において，これら 2 つのパラメータを変えて 10 分割交差検定を行い，RMSE が最小となるパラメータの組 $(h = 2, \sigma = 0.9)$ を求め，この値を評価実験に用いた．

実験の結果を表 1 に示す．この表には，補正手法とスムージング手法の 5 種類の組み合わせの他に，建石らが提案した 2 つの評価式 [2] の，教科書コーパスにおける相関係数を示した．

表 2 3 つの手法の出力の中間値を用いた場合の相関係数と RMSE

組み合わせた 3 手法	相関係数	RMSE
定数, 線形+2 次, 線形+ 3 次	**0.920**	1.413
定数, 定数+ GK 回帰, 線形+ 2 次	0.919	**1.412**
定数+GK 回帰, 線形+ 2 次, 線形+ 3 次	0.919	1.436

この表に示すように，建石らの 2 つの公式の相関係数の絶対値が 0.8 程度なのに対し、5 種類のいずれの方法も，0.9 に近い相関係数を示している．5 種類の組み合わせのなかでは，補正手法 1（小さな定数）とスムージング手法 1（生起確率のガウス回帰）の組み合わせが，相関係数と RMSE のいずれの指標においても，もっとも良い値を示した．

さらに，補正・スムージング手法の組み合わせが異なる 3 つの難易度推定器を使用し，3 つの出力の中間値を最終的な推定値とする場合の性能を調べた．この実験で性能が高かった 3 つの組み合わせを表 2 に示す.「定数」は補正手法 1 のみ,「定数+GK 回帰」は補正手法 1 とスムージング手法 1 の組み合わせ,「線形+2 次」は補正手法 2 とスムージング手法 2（2 次多項式回帰）の組み合わせ,「線形+3 次」は補正手法 2 とスムージング手法 2（3 次多項式回帰）の組み合わせを表す．この結果から，複数の手法を組み合わせることによって，推定精度を向上させることができることが分かった．

以上のように，教科書コーパスを用いた交差検定において，作成したシステムは非常に良好な性能を示した．

5.2 教科書以外のテキストへの適用

我々の最終的な目標は，多様なテキストに対して安定して難易度を推定するシステムを実現することにある．本節では，教科書以外のテキストに対して，作成したシステムがどのような挙動を示すのかを調べた．

評価テキストとして，ウェブ上から,「小学生向け」「中学生向け」「高校生向け」と対象読者が明記されたテキストを 4 サンプルずつ計 12 サンプルを入手した．さらに，一般向けのテキストとして，医療に関するテキストを 4 サンプル収集した．これらのサンプルテキストは，企業

表 3 教科書以外のテキストに対する難易度推定の結果

難易度 (学校区分)	分野	推定結果			
		小学	中学	高校	大学
小学	Web	4	0	0	0
中学	防災	0	4	0	0
高校	特許	0	0	4	0
一般	医療	0	0	3	1

や行政機関のウェブページ上に掲載されているテキストであり，それぞれの区分（小・中・高・一般）において異なる分野のものを選んだ．

これら 16 サンプルに対するシステムの難易度推定結果を表 3 に示す．この表に示すように，ラベルづけされた学校区分と推定結果の学校区分は概ね一致した．実験の規模が小さいので，さらなる評価が必要ではあるが，作成したシステムは，教科書以外のテキストに対しても，妥当な難易度を出力することがわかった．

6. 関 連 研 究

建石ら [2] は，日本語テキストの難易度を測定するため，文字の種類や文長等を用いた複数の公式を提案している．これらの公式によって得られる値は，複数のテキストの難易度を比較することには有効であるが，その値が具体的にどの程度の難易度に対応するものであるかは明らかではない．

川村 [6] は，テキスト中に存在する各語に対して日本語能力試験出題基準の級（1〜4 級）を出力するシステムを提案した．このシステムは，テキストに各級の語がどの程度の割合で含まれているかを出力する．しかし，各級の語の割合からテキスト全体の難易度を算出する方法は提供していない．

柴崎ら [7] は，小学校の国語の教科書を規準テキストとして採用し，ひらがなの含有率，文の総数に対する単文数，内容語の漢語率から、テキストの読みやすさを算出する公式を提案している．しかし，国語科以外の教科や他分野のテキストへの適用については考慮されていない．

7. お わ り に

本論文では，教科書コーパスを用いた日本語テキストの難易度推定手法を提案した．この手法は，文字 unigram の言語モデルを用いた分類問題として，難易度推定を定式化する．教科書コーパスを用いた交差検定実験において，あらかじめ付与された難易度と推定した難易度の相関係数は 0.92 であり，非常に高い相関を示すことがわかった．

本手法は，テキスト中に含まれるひらがな・カタカナ・漢字のみから難易度を推定する．そのため、テキスト中に数式や記号等の非テキスト要素が含まれていたとしても問題は生じない．また，形態素解析を前提としない方法なので，ウェブページのように，文を認定することが難しいテキストに対しても，問題なく適用できるという長所を持つ．

今後は，教科書以外のテキストに対して広範囲な実験を実施し，提案手法の性能，安定性および限界を明らかにするとともに，人間が実際に感じる難易度との比較検討を行っていきたいと考えている．

謝辞 本研究は、科学研究費補助金 基盤研究 (A) 「円滑な情報伝達を支援する言語規格と言語変換技術」(課題番号 16200009) の支援を受けた．

参 考 文 献

1) 高木裕子: 速読用読解教材開発に向けて　ーリーダビリティー研究を基礎にして, 関西外国語大学留学生別科目日本語教育論集，Vol.1, pp. 66–85 (1991).
2) 建石由佳, 小野芳彦, 山田尚勇: 日本文の読みやすさの評価式, 情報処理学会研究報告 1988-HI-018, pp. 1–8 (1988).
3) 近藤陽介, 佐藤理史: 多項ナイーブベイズ分類を用いた日本語テキストの難易度判定手法の検討, 言語処理学会第 13 回年次大会発表論文集, pp. 534–537 (2007).
4) 松吉俊, 近藤陽介, 橋口千尋, 佐藤理史: 全教科を収録対象とした日本語教科書コーパスの構築, 言語処理学会第 14 回年次大会発表論文集,D3-2 (2008).
5) Collins-Thompson, K. and Callan, J.: Predicting Reading Difficulty with Statistical Language Models, *Journal of the American Society for Information Science and Technology*, Vol. 56, No. 13, pp. 1448–1462 (2005).
6) 川村よし子: 語彙チェッカーを用いた読解テキストの分析, 早稲田大学日本語研究教育センター講座日本語教育, 第 34 分冊, pp. 1–22 (1998).
7) 柴崎秀子, 沢井康孝: 国語教科書コーパスを応用した日本語リーダビリティー構築のための基礎研究, 信学技報 NLC2007-32(2007-10), pp. 19–24 (2007).