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

近藤　陽介[†]　　佐藤　理史[†]

[†] 名古屋大学大学院 工学研究科

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，最も基本的な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，読者に内容が誤解なく伝わること，すなわち,「分かりやすいテキスト」であることが特に強く求められるものがある．例えば，災害現場や医療現場における生死に関わる情報を含むテキスト，教育現場における教科書や参考書に記述されるテキストなどである．しかし,「テキストの分かりやすさ」はテキストの作者の意識やスキルに委ねられており，その基準も作者によって異っていると考えられる．

そこで，計算機によって日本語テキストの分かりやすさを判定することを考える．テキストの作者に「分かりやすさに関する客観的評価」を提供することができれば，円滑な情報伝達の計算機による支援の一形態となるが，そのためには，日本語テキストの分かりやすさを測るための基準や手法が必要となる．

テキストの分かりやすさに影響する要素は，以下のように分類することができる．

(1) **情報内容の難しさ**
　記述される事柄や説明方法に起因する
(2) **言語表現の難しさ**
　(a) **構成要素の難しさ**
　　単語や文字の難しさに起因する
　(b) **構成方法の難しさ**
　　構造的複雑さや文長などに起因する

このうち,「情報内容の難しさ」はテキストの内容に強く依存する．そこで,「言語表現の難しさ」のみを判定の対象とし，これを**テキストの難易度**と定義する．

テキストの分かりやすさに関する研究は，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，様々な難易度測定の手法が提案されている．Flesch-Kincaid のように，文長やシラブル数などを手がかりとした，公式に基づく手法が代表的である．その他にも，語彙リストと照らし合わせる手法や，言語モデルを用いた手法の検討も行われている．日本語に対しても，難易度測定に関する検討は行われているが，実用的な手法は確立されていない．

本研究は，計算機によって日本語テキストの難易度を判定するため手法を確立することを目的とする．

本論文は，以下のように構成される．まず，第 2 章で難易度判定に関する本研究の方針について述べ，次に，第 3 章において、関連研究を述べる．第 4 章で難易度判定の手法について説明し，第 5 章で判定手法の実験とその結果及び考察を述べる．最後に，第 6 章でまとめを述べる。

2. 難易度判定に関する方針

2.1 対象範囲

先に示したように，テキストの難易度を決める要因を,「構成要素の難しさ」と「構成方法の難しさ」に分類する．このうち，本研究では，構成要素のみに注目し，構成要素からテキストの難易度を判定すること考える．

2.2 難易度の区分

本研究では，テキストの難易度の表す区分として,「中学生」や「高校生」などの学年区分を使用する．知識や読解力は人それぞれ異なり，1 つのテキストから感じる「難しさ」にも個人差が, このような個人差は考慮しない．

3. 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手がかりに難易度と相関のあるスコアを求める公式を導出している．このスコアを用いることにより，複数のテキストの難易度を相対的に比較することは可能である．しかし，そのスコアの値が具体的にどの程度の難易度に対応するものかは明らかになっていない．

川村 [3] は，テキスト中に用いられている語彙を日本語能力試験出題基準と照合するという検討を行っている．日本語能力試験には 1〜4 級のレベルが設定されており，各級の語彙がどの程度の割合で含まれているかを求めるシステムを提供している．しかし，各級の語彙の割合からテキスト全体の難易度をどの程度なのかを測る方法は提供していない．

4. 日本語テキストの難易度判定手法

本研究で作成した日本語テキストの難易度判定システムの概要を図 1 に示す． 入力は日本語テキストであり，これを被判定テキスト T と呼ぶ．ここで，難易度の種類数を N，各難易度を $G_i(i = 1, 2, ..., N)$ とする．システムは入力された T に対して，N 種類の難易度の中から適当と判断される難易度 G_T を 1 つ決定し出力する．

難易度の決定には，多項ナイーブベイズ分類を用いる．この手法を用いた，英語テキストの難易度判定に関する

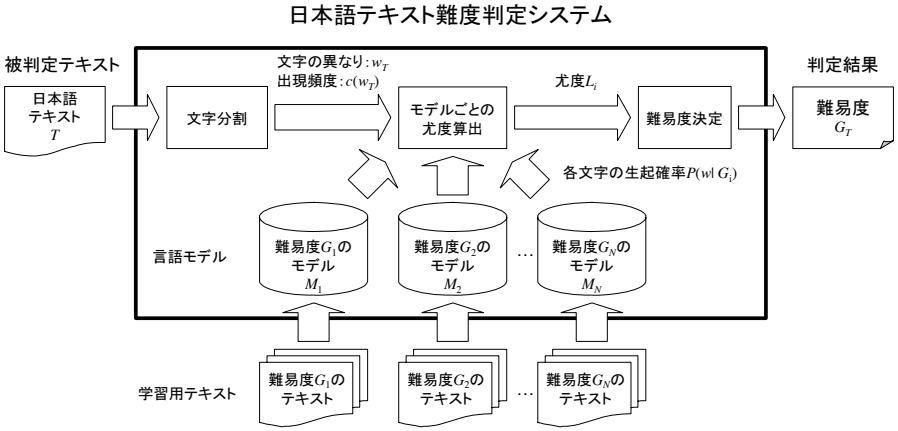

図 1 難易度判定システムの構成

研究 [4] が行われている．本研究の手法は，この先行研究を基礎としている．

4.1 難易度判定の手順

4.1.1 準　　備

難易度判定の前に，予め各難易度 G_i に対応する N 個の言語モデル M_i を構築しておく．言語モデル M_i は，難易度が G_i 相当であると思われるテキストを用いて構築する．言語モデルの構築に用いる学習用テキストと呼ぶ．言語モデル M_i は文字に関する Unigram モデルであり，学習に用いられたテキストに含まれる文字の異なり w_i と，その生起確率 $P(w_i|G_i)$ で定義される．

4.1.2 判　　定

入力したテキスト T の難易度を次の 3 つの手順で判定する．

1．文字分割　テキスト T に含まれる文字の異なり w と，各異なりの出現回数 $C(w)$ を求める．

2．尤度の算出　テキスト T に対する，各言語モデル M_i の尤度 L_i を求める．尤度 L_i の定義は次節で示す．

3．難易度の決定　最も尤度の高いモデル M_I を求め，それに対応する難易度 G_I をテキスト T の難易度 G_T と決定する．

4.2 尤度の定義

本研究では，先行研究で提案されている尤度をそのまま用いる．以下ではその詳細を述べる．

テキスト T の作者は，次のような手順でテキストを書くという生成モデルを仮定する．

(1) 作者は，言語モデルの集合 $M = \{M_1, .., M_N\}$ の中から，あるモデル M_i を確率 $P(M_i)$ で選択する．さらに，各言語モデルは語彙 V 上で多項分布をとるものとする．

(2) テキストの長さ（文字数）L を，確率 $P(L|M_i)$ で選択する．

(3) 先に選択した言語モデル M_i から，多項分布に従って L 個の文字を取り出す．テキスト T 中の各文字は，言語モデル M_i から互いに独立に取り出されるものとする．

以上の仮定から，言語モデル M_i が選ばれた場合に，テキスト T が生成される確率 $P(T|M_i)$ は次式で得られる．

$$P(T|M_i) = P(L|M_i)L! \prod_{w \in T} \frac{P(w|M_i)^{C(w)}}{C(w)!} \quad (1)$$

テキスト T が与えられた場合に，それが言語モデル M_i から生成されたテキストである確率 $P(M_i|T)$ は，ベイズの定理より次式である．

$$P(M_i|T) = \frac{P(M_i)P(T|M_i)}{P(T)} \quad (2)$$

ここで．さらに次の 2 つを仮定する．

(1) 各言語モデル M_i は，等しい確率 $P(M_i) = 1/N$ で選択される．

(2) テキストの長さが L である確率は $P(L|M_i)$，言語モデルに対し独立である．

これらの仮定と式 (1) 及び式 (2) より，$\log P(G_i|T)$ は次式で表される．

$$\log P(M_i|T) = \sum_{w \in T} C(w) \log P(w|M_i) + Z \quad (3)$$

ここで，言語モデル M_i 依存しない項を Z にまとめた．

テキスト T に対して尤もらしい難易度の言語モデル M_I を決定することは，言語モデルの集合 M から式 (3) が最大値をとるような言語モデル M_I を探索することに等しい．さらに，Z の項は言語モデル M_i に因らず一定である．これらのことから，テキスト T に対する各言語モデルの尤度を次式で定義する．

$$L_i = \sum_{w \in T} C(w) \log P(w|M_i) \quad (4)$$

4.3 先行研究との相違点

先行研究と本研究との大きな相違点は，先行研究が文字に関する言語モデルを用いているのに対し，本研究では文字に関する言語モデルを用いる点である．日本語は

表 1 ひらがな及び漢字の割合の比較 [%]

	中学校教科書	高校教科書	社説
ひらがな	52.85	48.57	47.44
学習漢字	36.91	39.20	41.10
学習漢字以外の常用漢字	2.46	3.34	4.24
常用漢字以外の漢字	0.35	0.44	0.26

英語と異なり，文字が意味を持ちうる最小単位であり種類も多い．文字を用いることにより，単語を用いた場合に比べ少ない量の学習用テキストで判定が可能であると思われる．さらに，同じ言葉でも，漢字と仮名のどちらを用いるかに自由度がある．5 章で述べる実験に用いた中学・高校の社会科教科書及び社説のテキストについて，記号を除く文字に対する漢字とひらがなの割合を表 1 に示す．漢字とひらがなの割合は，各テキストに想定される読者の違いによって異なる．このことから，文字が難易度を判定する手がかりになり得るのではないかと考えられる．

5. 実　　験

先述の手法を用いた難易度判定について，2 つの実験を行った．2 つの実験とも，難易度は中学生・高校生・一般の 3 種類とした．

5.1 実 験 1

実験 1 では，学習用テキストと被判定テキストが，同種のテキストであるという条件の下で，難易度判定の実験を行った．各難易度の学習用及び被判定テキストとして，以下のテキストを用意した．

- **中学生**
  - 中学生用の社会科（地理・日本史・世界史・公民の 4 分野）教科書で各分野 2 冊ずつ
  - 各教科書から 6 サンプル，計 48 サンプルを抽出
  - 全 73908 文字，1 サンプル平均 1539.8 文字
- **高校生**
  - 高校生用の社会科（ 地理・日本史・世界史・現代社会・倫理・政治経済の 6 分野）教科書で各分野 2 冊ずつ
  - 各教科書から 6 サンプル，計 72 サンプルを抽出
  - 全 168420 文字，1 サンプル平均 2339.2 文字
- **一般**
  - 2000 年毎日新聞の社説
  - 学習用として 1 月分・2 月分の計 146506 文字
  - 評価用として 3 月分から 20 サンプル，1 サンプル平均 1277.5 文字

ここで，中学生向けおよび高校生向けテキストについては，次のような操作を行った．まず，同一分野同一教科書の 6 サンプルの中から 1 サンプルずつ取り出す．取り出された中学生向けテキスト 8 サンプル，高校生向けテキスト 12 サンプルを被判定テキストとする．それ以外のサンプルを，各モデルの学習用テキストとする．同様の手順を，全てのサンプルが 1 度は被判定テキストとなるように 6 回行い，被判定テキストと学習用テキストのパターンを 6 種類作成した．各パターンの学習用テキストの文字数と各分野の比率の平均値を表 2 に示す．

表 2 学習用テキストの平均文字数と各分野の比率

難易度	平均文字数	分野	平均比率 [%]
中学	61590.0	地理	21.71
		日本史	25.88
		世界史	27.32
		公民	25.09
高校	140350.0	地理	14.09
		日本史	17.80
		世界史	17.62
		倫理	18.02
		現代社会	18.77
		政治経済	13.70

学習用テキストに無い文字がテキスト T 内に現れた場合，その文字にも生起確率を割り当てる必要がある．そこで，各言語モデルごとに Simple Good-Turing によるスムージングを行った．

判定結果を表 3 に示す．一般の被判定テキストは，20 サンプルを 6 回判定したため，表中では 120 サンプルとしている．中学が 87.5%，高校が 70.0%と高い正答率であった．社説は 100.0%であったが，これは難易度ではなく，社説の文字の統計的な傾向が，教科書と異なるために正しく判定されたと思われる．中学生向けのテキストは，公民のみが高校や社説と誤って判定され，他の 3 分野のテキストは正しく判定された．高校向けのテキストは，地理・日本史・世界史のテキストが，中学と判定される場合がみられた．これらは，中学の学習テキスト内に含まれる地理・日本史・世界史のテキストは割合が，高校の学習テキストに比べ高いことに起因すると思われる．一方，高校生向けの現代社会のテキストは，一般（社説）と誤って判定される場合が見られた．これは，内容の類似性により，出現傾向が近い文字が多く存在するためと思われる．

以上のことから，より難易度を反映した判定を行うためには，出現傾向が内容や分野に強く依存する文字を，判定の対象から外す必要があると思われる．そこで，以下の 2 つの条件を満たす文字のみを尤度計算に用いる実験を行った．

- 各難易度の学習用テキストにおいて，過半数を超える分野で出現する
- 全ての難易度の学習用テキストに出現する

その結果を，表 4 に示す．一部で判定が誤り転じるサンプルはあるが，全体として正答率は向上した．多くの各分野に共通に使われる文字のみを用いた尤度を用いることで，分野に依存する文字の影響を軽減できたものと思われる．

5.2 実 験 2

実験 2 では，被判定テキストが、学習用テキストとは

表 3 実験 1 の判定結果

想定難易度	分野	総数	判定結果			正答率 [%]
			中学	高校	一般	
中学生	地理	12	12	0	0	87.5
	日本史	12	12	0	0	
	世界史	12	12	0	0	
	公民	12	6	4	2	
高校生	地理	12	5	7	0	70.0
	日本史	12	3	9	0	
	世界史	12	2	10	0	
	倫理	12	0	11	1	
	現代社会	12	1	7	4	
	政治経済	12	1	10	1	
一般	社説	120	0	0	120	100.0

表 4 実験 1 で文字種を制限した場合の判定結果

想定難易度	分野	総数	判定結果			正答率 [%]
			中学	高校	一般	
中学生	地理	12	12	0	0	91.6
	日本史	12	12	0	0	
	世界史	12	12	0	0	
	公民	12	8	3	1	
高校生	地理	12	1	11	0	87.5
	日本史	12	2	10	0	
	世界史	12	3	9	0	
	倫理	12	0	12	0	
	現代社会	12	1	11	0	
	政治経済	12	1	10	1	
一般	社説	120	0	0	119	99.1

別種のテキストであるという条件の下で，難易度判定の実験を行った．学習用テキストとして，実験 1 で用意した各難易度のすべてのテキストを用いた．被判定テキストとして，以下のテキストを用意した．

- **小学生向けウェブサイトのテキスト（各 6 サンプル）**
  - 電波に関するサイト:平均 1525.2 文字
  - 古墳に関するサイト:平均 1840.5 文字
  - リサイクルに関するサイト：平均 2828.8 文字
- **中学生向けウェブサイトのテキスト（8 サンプル）**
  - 金融に関するサイト：平均 1737.4 文字
- **高校生向けウェブサイトのテキスト（7 サンプル）**
  - 金融に関するサイト：平均 2300.7 文字
- **一般向けテキスト**
  - 電子情報通信学会 2006 年 12 月論文誌内の論文のまえがき：12 サンプル，平均 2236.4 文字
  - 平成 4〜6 年通産省通商白書内のテキスト：10 サンプル，平均 2058.2 文字

判定結果を表 5 に示す．想定される難易度に対し，特定の分野のテキストは判定結果が大きく異なることが分かる．小学生向けテキストでは，古墳やリサイクルといった社会科に関連のあるテキストは中学生と判定される傾向があった．加えて，通信白書は，高校生と判定される傾向があった．一方で，小学生の電波に関するテキスト，中学・高校の金融に関するテキスト及び論文は，一般と判定される傾向があった．これらのことから，被判定テキストの分野が学習用テキストとは大きく異なる場合，判定結果は，難易度よりも被判定テキストの分野に強くされる．さらに，先ほどと同様に文字種を制限した場合の実験も行ったが，改善は見られなかった．学習テキスト内の過半数を超える分野で出現する文字に尤度計算の対象を制限した場合でも，分野に影響を受ける文字の選別ができていなかった．

表 5 実験 2 の判定結果

想定難易度	内容	総数	判定結果		
			中学	高校	一般
中学生（小学生）	電波	6	0	1	5
	古墳	6	6	0	0
	リサイクル	6	5	0	1
中学生	金融	8	0	0	8
高校生	金融	7	0	0	7
一般	論文	12	0	1	11
	通商白書	10	0	8	2

6. おわりに

日本語テキストの難易度を判定する手法として，文字に関するナイーブベイズ推定を用いる手法の検討を行った．

学習用テキストと被判定テキストが同種のテキストの場合は，高い精度で分類が可能であった．さらに，多くの分野のテキストで出現する文字のみ用いて尤度を算出することで，分類の精度を高めることが可能であることも分かった．一方で，被判定テキストが，学習用テキストとは全く別種のテキストである場合は，難易度よりも分野に依存してしまうことが分かった．

より幅広い分野のテキストに対応した難易度判定を実現するため，学習に用いるテキストの分野を広げることが必要である．あわせて，分野に依存する文字と難易度に依存する文字を選別し，難易度判定において前者の影響を軽減させる手法の検討が必要である．

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) 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).
4) 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).