文字 bigram モデルを用いた日本語テキストの難易度推定

小島　健輔 [†1]　　佐藤　理史 [†2]　　藤田　篤 [†2]

[†1] 名古屋大学工学部 電気電子・情報工学科, [†2] 名古屋大学大学院 工学研究科

1. はじめに

現代社会の複雑化と流動化により，情報伝達・意思疎通の重要性はますます高まっている．1990 年代のインターネットの発展とそれに伴う電子メールやウェブの普及により，テキストは再び情報伝達の最重要メディアに返り咲いた．一方で，高齢化社会の到来や在日外国人の増加等により，人々の日本語能力は多様化してきている．このような背景により，情報を「平易でわかりやすいテキスト」で書き表すことの重要性が高まっている．

我々は,「平易でわかりやすいテキストを増やすためには，まず，テキストの難易度を簡単に調べられるツールが必要である」という考えに基づき，2003 年から日本語テキストの難易度を推定する方法について検討を行ってきた．2007 年には，約半年をかけて約 100 万字の教科書コーパス [1] を編纂し，これを規準コーパスとして難易度を推定するツール『帯』を作成した [2),3)]．『帯』では，難易度を推定するために文字 unigram モデルを用い，相関係数 0.919 の性能を達成した．

本研究では，文字 unigram モデルの代わりに文字 bigram モデルを用いるシステムを提案する．提案システムと，unigram モデルを用いた既存システムの難易度推定性能を比較し，2 つのシステムの優劣を実験的に明らかにする．

2. 文字 bigram モデルを用いた難易度推定

2.1 規準コーパスと難易度区分

本研究の主眼は，文字 unigram を用いた既存システムと文字 bigram を用いた提案システムの比較にある．このため，提案システムの規準コーパスとして，既存システムが規準コーパスとして採用した教科書コーパス [1] をそのまま用いる．

教科書コーパスは，愛知県名古屋市で使用されている小・中・高の全学年・全教科の教科書のうち，英語を除く全教科の 111 冊の教科書と，大学の教養課程の教科書 16 冊の計 127 冊の教科書から抽出した，総計 1,478 件 (約 105 万字) のサンプルテキストから構成されている．このコーパスにおいて，難易度は，小・中・高の学年区分 12 段階に大学を加えた 13 クラスに区分されている．教科書コーパスの概要（各難易度クラスに対するサンプルテキスト数と文字数）を表 1 に示す．

表 1　教科書コーパスの概要

難易度クラス	サンプル数	文字数
1	37	10,451
2	45	17,407
3	70	31,204
4	65	21,555
5	67	40,523
6	62	30,020
7	89	70,400
8	93	71,107
9	78	67,243
10	143	94,445
11	225	134,326
12	193	122,209
小計	1167	710,890
13	311	341,016
合計	1,478	1,051,906

2.2 bigram 言語モデルの構築

上記の教科書コーパスを用いて，13 の難易度クラスそれぞれに対して文字 bigram 言語モデルを構築する．コーパス中に現れる文字のうち，ひらがな，カタカナ，JIS 第一水準の漢字のみを有効文字とし，それ以外の数字や記号，アルファベット等はすべて無視する．言語モデル M_i における文字 bigram $x_{j-1}x_j$ の生起確率 $P_i(x_j|x_{j-1})$ を，次式で求める．

$$P_i(x_j|x_{j-1}) = \frac{f(x_{j-1}x_j, D_i)}{f(x_{j-1}*, D_i)} \tag{1}$$

ここで，x は有効文字，$*$ は任意の有効文字を表す．すなわち，文字 bigram $x_{j-1}x_j$ は，x_{j-1} と x_j の両方が有効文字であるもののみを用いる．これを有効 bigram と呼ぶ．D_i は難易度クラス i が付与されたテキストの集合（学習テキスト）を表し，$f(x_{j-1}x_j, D_i)$ は 学習テキスト D_i における文字 bigram $x_{j-1}x_j$ の出現回数を表す．

次節で定義する尤度は，$P_i = 0$ の場合には計算不能となるので，文字 unigram を用いる既存システムと同じ方法で補正する．すなわち，補正には以下の式を用い，すべての P_i が 0 でなくなるまで，この式を繰返し適用する．

$$P_i(x_j|x_{j-1}) = \frac{P_{i-1}(x_j|x_{j-1}) + P_{i+1}(x_j|x_{j-1})}{2} \tag{2}$$

なお，$P_i(x_j|x_{j-1})$ が全ての難易度クラス i に対して 0 の場合は，その文字 bigram $x_{j-1}x_j$ を尤度計算に使用する bigram（有効 bigram）から除外する．

2.3 尤度計算と難易度の決定

テキスト T の難易度を求めるために，まず，各言語モ

デル M_i における尤度 $L(M_i|T)$ を，次式を用いて計算する．

$$L(M_i|T) = \sum_{x_{j-1}x_j \in T} f(x_{j-1}x_j, T) \log P_i(x_j|x_{j-1}) \quad (3)$$

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 i を推定結果として出力する．

2.4 尤度のスムージング

難易度クラスには，順序関係（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）が存在する．このため，上記で計算した尤度は，難易度クラスに対して緩やかな曲線を描くことが期待される．このような考えに基づけば，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．

文字 unigram を用いた既存システムでは，まず，尤度計算で得られた 13 個の値を 2 次曲線および 3 次曲線に当てはめてスムージングし，スムージング後の尤度値が最大となるクラスを求める．次に，こうして得られた 3 つの値（スムージングなし，2 次曲線スムージング，3 次曲線スムージング）の中央値を最終的な難易度推定値として採用する．本論文で提案する文字 bigram を用いたシステムでも，この方法を踏襲する．但し，4 次曲線スムージングを新たに導入する．すなわち，次の 6 種類の方法を試す．

(1) スムージングなし
(2) 2 次曲線スムージング
(3) 3 次曲線スムージング
(4) 4 次曲線スムージング
(5) 中央値（スムージングなし，2 次曲線，4 次曲線）
(6) 中央値（スムージングなし，3 次曲線，4 次曲線）

3. 実　　験

3.1 既存システムとの比較

ここでは，文字 bigram を用いる提案システムの性能と，文字 unigram を用いる既存システムの性能を，Leave-one-out 交差検定によって比較する．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，二乗平均平方根誤差 (Root Mean Square Error; RMSE) を用いる．

Leave-one-out 交差検定の結果を表 2 に示す．この表において，「± 0」の欄はテキストに付与されている難易度が推定結果と一致した割合（的中率）を，「± 1」の欄は前後 1 クラスのずれを許容した場合の的中率を示している．

文字 unigram を用いる既存システムで性能が良いのは，中央値（スムージングなし，2 次曲線，3 次曲線）の場合で，このときの相関係数は 0.919，RMSE は 1.441 である．一方，文字 bigram を用いる提案システムで性能が良いのは，4 次曲線スムージングの場合で，このときの相関係数は 0.940，RMSE は 1.216 である．中央値を用いる 2 つの方法も，それとほぼ同等の性能である．文字 bigram を用いる提案システムは，2 次曲線スムージングの場合を除いて，すべて既存システムの性能を上回っている．これらの事実から，文字 bigram を用いる提案システムは文字 unigram を用いる既存システムより優れていると判断できる．

3.2 有効 bigram の制限

次に，有効 bigram を制限する実験を行った．この実験を行った理由は，出現回数が低い文字 bigram は尤度計算にはあまり貢献せず，逆にノイズとなっている可能性があると考えたからである．ここでは，規準コーパス全体における出現回数が N 回未満の文字 bigram を有効 bigram から除外した場合の性能を，leave-one-out 交差検定により求めた．なお，スムージング方法には，先の実験で最も性能が良かった 4 次曲線スムージングを用いた．

実験結果を表 3 に示す．この表において，「異なり」と「トークン数」は，それぞれの場合の有効 bigram の異なり数とトークン数を表す．この表では，比較のために文字 unigram を用いる既存システムのデータも併記した．

この表より，出現回数が 1 回の文字 bigram を有効 bigram から除外した場合，RMSE が向上することがわかる（相関係数は変わらない）．このときの文字 bigram の異なり数は，36,814 である．

一方，文字 unigram の異なり数は 2,532 である．これは，上記の数と比較すると 1 桁少ない．提案システムにおいて，50 回未満の文字 bigram を削除した場合，有効 bigram の異なり数は 2,653 となり，文字 unigram の異なり数と近い値となる．このときのトークン数は 50.2 万であり，全体の 37%を削除したことになる．このように有効 bigram を大幅に削除した場合でも，提案システムの性能は，文字 unigram を用いる既存システムとほぼ同等である．以上の事実は，文字 bigram は，難易度の特徴量として，文字 unigram よりも優れていることを示唆する．

出現回数が 1 回の文字 bigram を有効 bigram から除外した場合の，各スムージング方法に対する提案システムの性能を表 4 に示す．この表は，表 2 とほぼ同じ傾向を示しており，4 次曲線スムージングが最も優れた性能を示している．

3.3 教科書以外のテキストへの適用

我々の最終目標は，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（論文 [3] で実験に使用されているものと同一のテキスト）に対し，提案システムによる難易度推定実験を行い，提案システムの推定結果と既存システムの推定結果を比較する．なお，本節の実験では，学習用コーパスとして教科書コーパス全体を用い，出現回数が 1 回の文字 bigram を有効 bigram から除外して作成したモデルを用いた．尤度のスムージングには

表 2 leave-one-out 交差検定による比較

言語モデル	スムージング	相関係数	RMSE	± 0	± 1
bigram	(1) なし	0.927	1.323	**0.509**	0.799
	(2) 2 次曲線スムージング	0.912	1.513	0.245	0.664
	(3) 3 次曲線スムージング	0.934	1.277	0.419	0.778
	(4) 4 次曲線スムージング	**0.940**	**1.216**	0.452	**0.815**
	(5) 中央値（なし，2 次，4 次）	0.937	1.236	0.449	0.811
	(6) 中央値（なし，3 次，4 次）	0.938	1.235	0.465	0.809
unigram	なし	0.900	1.620	0.440	0.721
	2 次曲線スムージング	0.885	1.794	0.167	0.522
	3 次曲線スムージング	0.900	1.686	0.368	0.676
	中央値（なし，2 次，3 次）	0.919	1.441	0.398	0.738

表 3 使用する文字 bigram を制限したときの性能の変化

N 回未満削除	異なり	トークン数	相関係数	RMSE	± 0	± 1
削除しない場合	62,467	796,813	**0.940**	1.213	**0.455**	0.814
2	36,814	771,160	**0.940**	**1.207**	0.453	**0.817**
3	27,386	752,304	0.938	1.237	0.453	0.811
4	22,217	736,797	0.934	1.279	0.449	0.805
5	18,718	722,801	0.932	1.302	0.447	0.804
6	16,269	710,556	0.932	1.304	0.448	0.798
10	10,871	671,405	0.928	1.343	0.443	0.798
20	6,053	606,335	0.927	1.361	0.448	0.793
30	4,246	563,312	0.922	1.417	0.436	0.790
40	3,245	529,216	0.921	1.421	0.434	0.788
50	2,653	502,368	0.921	1.421	0.425	0.781
unigram	2,532	912,999	0.919	1.441	0.398	0.738

表 4 出現回数 1 回の文字 bigram を削除した場合

言語モデル	スムージング	相関係数	RMSE	± 0	± 1
bigram	(1) なし	0.930	1.298	**0.515**	0.802
	(2) 2 次曲線スムージング	0.912	1.504	0.246	0.669
	(3) 3 次曲線スムージング	0.935	1.260	0.422	0.784
	(4) 4 次曲線スムージング	**0.940**	**1.207**	0.453	**0.817**
	(5) 中央値（なし，2 次，4 次）	**0.940**	1.213	0.453	0.816
	(6) 中央値（なし，3 次，4 次）	**0.940**	1.211	0.472	0.813

4 次曲線スムージングを用いた．

3.3.1 NHK 週刊こどもニュース

NHK で放送されているテレビ番組「週刊こどもニュース」のウェブページにある「今週の大ハテナ」のテキスト 389 サンプルに対する難易度推定結果を表 5 に示す．これらのテキストの正解難易度は不明であるが，おおよそ小学校高学年から中学生（難易度 5-9）と考えられる．

この表が示すように，提案システム (bigram) の推定結果は，難易度 7 から 9 の間に集中しており，平均は 8.01，分散は 0.58 である．一方，既存のシステム (unigram) の推定結果は，平均こそ 8.43 であるが，分散は 1.83 であり，bigram の場合と比較して，かなり散らばっていることがわかる．

ここで使用したテキストは，注意深く編集されたテキストであり，難易度はそれほどばらつかないはずである．この推測が正しいとすれば，分散がより小さい提案システムの方が，より正しく難易度を推定できていると考えられる．

3.3.2 ウェブページ

対象とする学生（小学生，中学生，高校生）が明記されているウェブページ 29 サイト，268 ページに対する難易度推定結果を表 6 に示す．提案システム (bigram) と既存システム (unigram) の結果を比較すると，推定された難易度の平均値にそれほど差はない．一方，分散は提案システムの方が小さい値となり，特に，小学生用ページにおいて，その傾向が顕著である．しかしながら，この実験結果から，提案システムと既存システムの優劣を読みとることはできない．

3.4 短いテキストへの適用

本節では，テキストがどの程度の長さであれば安定して難易度推定を行うことができるのかについて述べる．

実験では，1,478 サンプルの教科書テキストのうち，有効 bigram が 250 回以上出現する 1,194 サンプルのみを利用し，leave-one-out 交差検定を行った．モデルの作成では，出現回数が 1 回の文字 bigram を有効 bigram から除外した．難易度を推定する際は，対象テキストの先頭から N 個の有効 bigram のみを用い，尤度のスムージングには 4 次曲線スムージングを用いた．

実験結果を表 7 に示す．ここでは，論文 [3] に示されている既存システム (unigram) の性能値を参考のため併記した．但し，既存システムの実験では，有効文字数 250 以上の 1,286 サンプルを使用しており，完全に同一条件

表 5 週刊こどもニュースの難易度推定結果

言語モデル	スムージング	ページ数	推定結果										
			5	6	7	8	9	10	11	12	13	平均	分散
bigram	4 次曲線スムージング	389	5	2	89	210	82	5	0	0	1	8.01	0.58
unigram	中央値（なし，2 次，3 次）	389	1	34	46	106	165	10	15	5	7	8.43	1.83

表 6 ウェブページの難易度推定結果

言語モデル	対象学生	ページ数	推定結果												
			3	4	5	6	7	8	9	10	11	12	13	平均	分散
bigram	小学生 (難易度 1-6)	135	0	1	5	26	65	33	3	2	0	0	0	7.04	0.87
	中学生 (難易度 7-9)	78	0	0	0	0	11	25	28	11	3	0	0	8.62	1.03
	高校生 (難易度 10-12)	55	0	0	0	0	1	10	16	19	3	0	6	9.67	2.11
unigram	小学生 (難易度 1-6)	135	2	16	29	21	14	29	16	2	1	3	2	6.73	4.27
	中学生 (難易度 7-9)	78	0	0	0	1	5	11	30	17	6	7	1	9.38	1.90
	高校生 (難易度 10-12)	55	0	0	0	0	0	2	15	20	0	9	9	10.47	2.32

表 7 短いテキストに対する難易度推定結果

言語モデル	評価値	有効 bigram 数 (有効文字数)								
		10	15	20	25	50	100	150	200	250
bigram	相関係数	0.830	0.875	0.888	0.907	0.920	0.936	0.939	0.946	0.947
	RMSE	1.935	1.641	1.547	1.395	1.284	1.149	1.111	1.046	1.026
unigram	相関係数	0.750	0.806	0.810	0.829	0.857	0.883	0.897	0.907	0.907
	RMSE	2.308	2.039	2.009	1.918	1.777	1.617	1.509	1.428	1.411

の比較実験とはなっていない．

この表より，有効 bigram が 25 以上あれば，相関係数が 0.9 以上と高い相関を示すことがわかる．また，unigram の場合と比較すると，より少ない数の bigram で，高い相関係数，低い RMSE 値を達成していることがわかる．

4. 関 連 研 究

我々が先に提案した難易度推定システム『帯』以外に，日本語テキストの難易度を推定する方法として，難易度公式を用いる方法がある．その先駆的な研究に，文字の種類や文長等を用いた公式を提案した建石ら [4] の研究がある．最近の柴崎ら [5] の研究は，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，内容語の漢語率を特徴量として利用した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．

寺田・田中 [6] は，2 つのテキストの難易順序判定器を構成し，これを用いてテキスト集合を難易度順にソートする方法を提案している．この方法では，難易度はスコアではなく，ソートされたテキスト集合の相対順位として表現されることとなる．

5. お わ り に

本論文では，文字 bigram モデルを用いた日本語テキストの難易度推定システムを提案し，文字 unigram モデルを用いた既存システムと比較した．提案システムを leave-one-out 交差検定で評価した結果，相関係数および RMSE ともに既存システムよりも上回る性能が得られた．また，より短いテキストに対して，安定して妥当な難易度が推定できることがわかった．

参 考 文 献

1) 松吉俊, 近藤陽介, 橋口千尋, 佐藤理史. 2008. 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, pp.520-523.
2) 近藤陽介, 松吉俊, 佐藤理史. 2008. 教科書コーパスを用いた日本語テキストの難易度推定. 言語処理学会第 14 回年次大会発表論文集, pp.1113-1116.
3) Satoshi Sato, Suguru Matsuyoshi, and Yohsuke Kondoh. 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *Proceedings of the 6th International Language Resources and Evaluation* (LREC2008).
4) 建石由佳, 小野芳彦, 山田尚勇. 1988. 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.
5) 柴崎秀子, 沢井康孝. 2007. 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.
6) 寺田博視, 田中久美子. 2008. 文書の難易順序判定法. 言語処理学会第 14 回年次大会併設ワークショップ「教育・学習を支援する言語処理」論文集, pp.59-62.