日本語テキストの難易度判定ツール『帯』

名古屋大学大学院工学研究科教授
佐藤　理史

PROFILE 京都大学大学院工学研究科電気工学第二専攻博士課程研究指導認定退学。博士（工学）。北陸先端科学技術大学院大学、京都大学を経て、2005年6月より名古屋大学大学院工学研究科電子情報システム専攻教授。

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100万字の教科書コーパスを編纂し、このコーパスを規準として難易度を推定するルール『帯』を作成した[1]。『帯』の主目標は、ウェブページの難易度の推定であり、それを実現するために、文字の出現確率に基づく推定法を採用している。本稿では、この『帯』について紹介する。

2 難易度判定ツール『帯』

ウェブのブラウザを通して利用できる『帯』の入力画面を図1に示す。この画面のボックスに難易度を測定したいテキストを入力し、「難易度を推定」のボタンを押

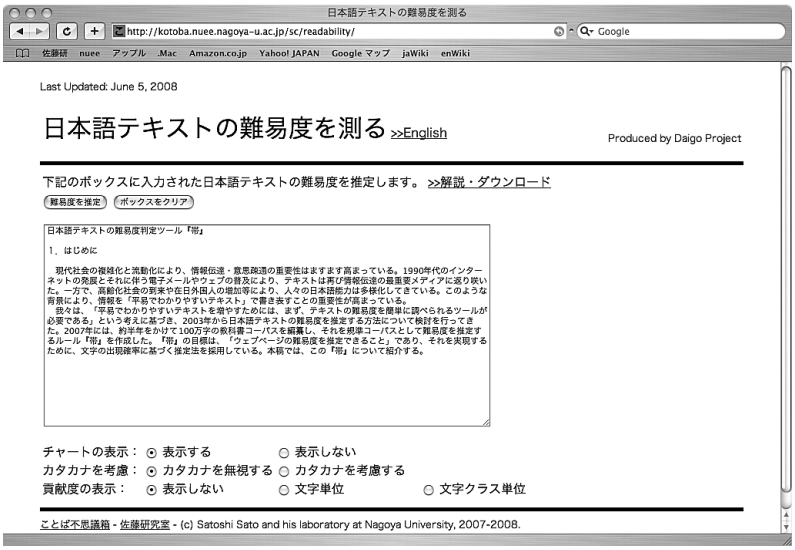

図1 『帯』の入力画面

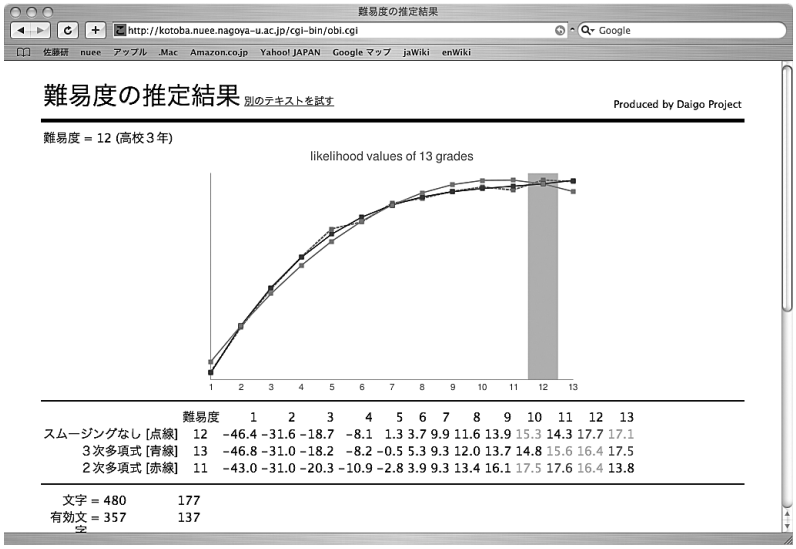

図2 『帯』の出力画面

すと、図2のような画面が表示される。出力される難易度は学年区分に対応しており、小学1年（難易度1）から高校3年（難易度12）までの12学年に、大学（難易度13）を加えた13段階で表示される。

『帯』の最大のポイントは、対象テキストを選ばないという点である。文章は文の並びと定義できるが、実際のテキストでは句点で終わる文以外に、タイトルや見出し、箇条書きといった句点を伴わないテキスト要素や、数式や記号、図表、URLなどの非テキスト要素が出現する。ウェブページは、特にその傾向が顕著である。『帯』はそのようなテキストであっても問題なく動作するので、テキストをそのままコピーアンドペーストで入力することができる。

『帯』は、2008年5月よりhttp://kotoba.nuee.nagoya-u.ac.jpで公開しており、誰もが自由に利用することができる。また、プログラミング言語Rubyが動作するUnix環境があれば、プログラムをダウンロードして、ローカルコマンドとして利用することもできる。

3 教科書コーパス

どのような難易度推定法を採用するにしても、まずは、難易度の規準となるテキスト集合が必要不可欠である。これを**規準コーパス**と呼ぶ。規準コーパスは、つぎの条件を満たす必要がある。

（1）コーパスに含まれるそれぞれのテキストには、難易度が付与されていること。

（2）コーパスは、多様な難易度のテキストをカバーしていること。

（3）コーパスは、多様なジャンルのテキストをカバーしていること。

近年、多くのテキストコーパスが電子的に利用可能になってきたが、残念ながら難易度が付与されたコーパスは存在しない。そこで、我々は、小中高の教科書を利用して規準コーパスを編纂することにした。

小中高の教科書は、文部科学省の検定により内容が統制されているが、それとともにテキストの難易度も統制されていることが期待できる。大部分の教科書は、使用する学年が明示されているので、これを難易度レベルと

して利用できる。わかりやいウェブページを作成する指針である“Web Content Accessibility Guideline 2.0”（http://www.w3.org/TR/WCAG20/）は、9年間の教育を受けた人が理解できないような難しいテキストには、よりやさしいテキストを併記することを求めているが、このための難易度測定にも、難易度レベルが学年として出力されると都合がよい。

我々は、名古屋市で使われている、小学1年から高校3年までの全ての教科書を1セット購入し、そこからサンプルを抽出して電子化データを作成した。ここで、上記の条件（3）を満たすために、国語、数学、社会、理科の主要4教科からだけでなく、美術、音楽、技術・家庭、情報、保健体育、書字など全ての科目（但し、英語と地図帳を除く）からサンプルを抽出した。

小中高の教科書の入力作業が一段落した後、高校よりも難しいレベルもほしいということになり、高校の教科に対応する大学の教養課程の教科書16冊を選び、これらからもサンプルを抽出して電子化した。こうして出来上がったのが、総文字数約100万文字の通称「教科書コーパス」である。教科書コーパスの概要を表1に示す。

4 難易度推定法

伝統的な難易度推定法は、リーダビリティ公式と呼ばれる数式を用いる方法である。たとえば、英語のリーダビリティを測る最も有名な公式の一つであるFlesch-Kincaidの公式は、次の式で与えられる。

$$G=0.39sl+11.8wl-15.59$$

ここで、slは1文当たりの語数、wlは1語当たりのシラブル数を表し、Gはそのテキストが理解できる学年を表す。この式は次のように理解できる。リーダビリティを決定する2つの主要な因子は、使われている語の難しさと文の構造の難しさであり、wlは前者を（英語の単語は、一般にシラブルの数が増えるほど難しい）、slは後者を（1文当たりの語数が増えるほど、文の構造は複雑になる）を見積もる指標となっている。

この公式を適用するためには、テキストを文に区切り、文を語に区切る必要がある。前述の通り、ウェブページには句点を伴わないテキスト要素や非テキスト要素が多数含まれるため、文区切りの認定がそれほど容易ではない。テキサス大学のサイト（http://webapps.lib.utexas.edu/TxReadability/app）では、ウェブページのリーダビリティの測定に、文区切りを必要としないForcast Grade Levelを用いているが、この方法でも語の区切りは必要となる。

Forcast Grade Level＝

20－（150語中に含まれる1シラブル語の数）/10

一方、最近提案されたテキスト分類に基づく手法は、あらかじめ難易度別のコーパスを準備しておき、判定対象テキストが、それらのうちのどの難易度のコーパスに最もよく似ているかをもって、難易度を推定する方法である。具体的には、まず、各難易度Giに対して尤度$L(Gi|T)$を次式で定義する[2]。

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

表1　教科書コーパスの概要

	難易度	教科書数	サンプル数	文字数
小　学	1-6	53	346	151,160
中　学	7-9	25	260	208,750
高　校	10-12	33	561	350,980
大　学	13	16	331	341,016
合　計	1-13	127	1,478	1,051,906

ここで、Tは難易度を推定したいテキスト、wはTに含まれる語、$C(w)$はTにおけるwの頻度、$P(w|Gi)$は難易度Giのコーパスにおけるwの出現確率を表す。こうして得られた尤度のなかで、最大の尤度値をとる難易度を、推定結果として出力する。

この方法を、語ではなく文字を使って計算しようというのが、我々の基本的なアイディアである。日本語テキストの場合、文を語に区切るためには、形態素解析を適用する必要がある。文が正しく区切られていれば、形態素解析の精度も高いが、ウェブページを対象として想定するならば、そのような楽観的な状況は仮定できない。となれば、残された選択肢は文字ベースの方法しかない。

日本語の漢字は表意文字であり、それ自身で固有の意味を持つ。JIS第1水準に含まれるだけでも2965文字あり、擬似的な語とみなすことができる。さらに、テキストの難易度を制御するために、我々は、漢字とかなを使い分けたり（あるいは、ルビを振ったり）、漢語と和語を使い分けたりする。このような理由により、文字ベースの方法でも高い精度で難易度を推定できる可能性がある。

『帯』の難易度推定法の概略を図3に示す。まず、教科書コーパスの13の難易度レベルのそれぞれにおいて、文字xの出現確率$P(x|Gi)$を求めておく。次に、難易度を推定したいテキストTが与えられると、次式を用いて各難易度Giの尤度$L(Gi|T)$を計算する。

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

ここで、zはTに出現する有効文字、$C(z,T)$はその出現回数を表す。有効文字とは、難易度を推定する際に考慮する文字のことであり、現在は、ひらがな83文字とJIS第1水準の漢字2965文字を用いている。

こうして得られた13個の尤度値からさらに次の3つの値を求め、その中央値を最終的な難易度の推定値として出力する。

（1）13個の尤度値のうち、最大の尤度値をとる難易度Gi

（2）13個の尤度値に2次曲線をあてはめ、それぞれの難易度に対して平滑化された尤度$L'(Gi|T)$を求める。この尤度値が最大となる難易度Gi'

（3）13個の尤度値に3次曲線をあてはめ、それぞれの

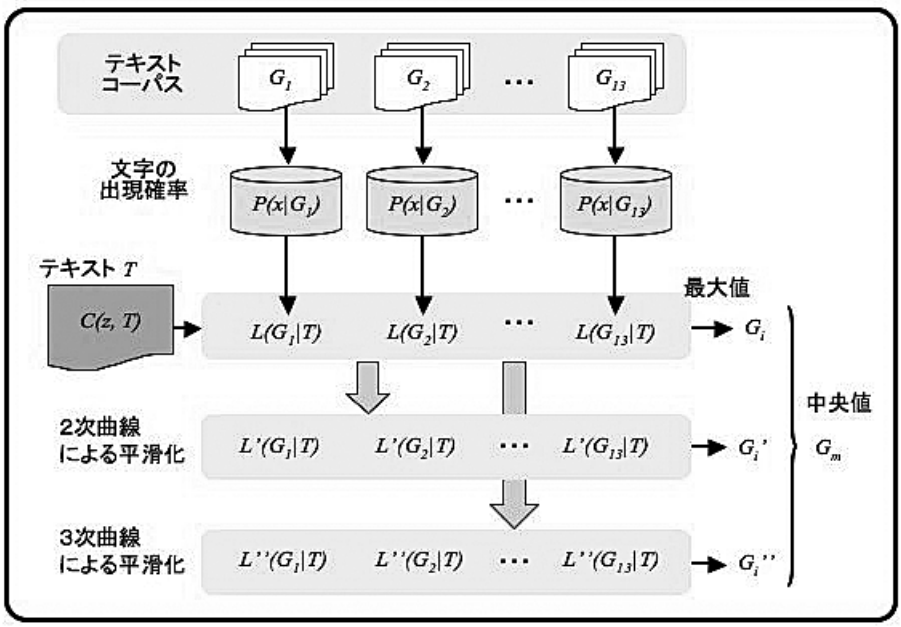

図3 『帯』の難易度推定法

難易度に対して平滑化された尤度 $L''(Gi|T)$ を求める。この尤度値が最大となる難易度 Gi''

ここで、2次曲線や3次曲線を用いて平滑化するのは、13個の尤度値は難易度に対して連続的に変化することが期待されるからである。なお、図1で難易度の下に表示される3本のグラフは、それぞれ $L(Gi|T)$, $L'(Gi|T)$, $L''(Gi|T)$ の値を示している。

5 『帯』の性能

『帯』の性能を、leave-one-out交差検定法で評価した結果を表2に示す。この表で、相関係数は、規準コーパスで定義される真の難易度と、『帯』の推定値との間にどのくらいの相関があるかを示す指標であり、絶対値が1.0に近いほど強い相関があることを表す。一方、Root Mean Square Error（RMSE）は、真の値と推定値との誤差を表す指標で、小さいほど好ましい。この表では、前章の（1），（2），（3）のそれぞれ単独の場合と、それらの中央値をとった場合を示しているが、中央値をとる方法がどちらの指標でも優れていることがわかる。

『帯』が、どれくらい短いテキストに適用可能かを調べるために、各テキストの難易度を、最初のN個の有効文字のみを用いて推定した場合の結果を表3に示す。この表より、25文字で相関係数が0.8を超え、100文字で0.883となることがわかる。これより、有効文字が100文字程度あれば、かなり安定して難易度を推定できるということがわかる。

上記の交差検定の他に、実際のウェブページの難易度を推定する実験も行った。たとえば、NHK週刊こどもニュースのウェブサイトにある原稿389件は、おおむね小学6年から中学（難易度6から9）と判定された。また、ウェブで見つけた小中高向けの計268ページの推定結果も、小中高のそれぞれで平均を取ると、その値はページ作成者が想定している読者の範囲にほぼ入ることがわかった。

表2　Leave-one-out交差検定による評価

方法	相関係数	RMSE
（1） 平滑化なし	0.898	1.632
（2） 2次曲線による平滑化	0.882	1.817
（3） 3次曲線による平滑化	0.898	1.691
上記3つの値の中央値	0.916	1.469
建石[3] の公式	-0.758	N/A

6 おわりに

本稿では、日本語テキストの難易度判定ツール『帯』について述べた。『帯』では、ウェブページを含むあらゆるテキストに適用可能であることを目指したが、この目標はほぼ達成できた。さらに、交差検定による評価でも、かなり良い性能を示すことがわかった。

『帯』の実現により、日本語テキストの難易度判定は一つの壁を超えたと考えるが、まだまだ課題も多い。社会的需要が高いと考えられるのは、成人の母国語話者が感じる「やさしい・ふつう・むずかしい」の判別であるが、そのためには、現在の難易度13（大学）を、複数のレベルに分ける必要がある。また、適切な難易度レベルの教材の選定といった、非母国語話者への学習支援という応用も重要であるが、非母国語話者にとっての難易度は、母国語話者が感じる難易度とは一致しない可能性

表3　有効文字の制限

有効文字数	10	25	50	100	200
相関係数	0.750	0.829	0.857	0.883	0.907
RMSE	2.308	1.918	1.777	1.617	1.428

もある。今後は、これらの応用を見据え、『帯』を改良していく予定である。

参考文献

[1] Satoshi Sato, Suguru Matsuyoshi and Yohsuke Kondoh. Automatic assessment of Japanese text readability based on a textbook corpus. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[2] Kevyn Collins-Thompson and Jamie Callan. A language modeling approach to predicting reading difficulty. *Proceedings of the HLT/NAACL 2004 Conference, pp.193-200, 2004.*

[3] 建石由佳，小野芳彦，山田尚勇．日本語文の読みやすさの評価式．情報処理学会 文書処理とヒューマンインタフェース研究会，18-4, 1988.