TRF: テキストの読みやすさ解析ツール

渡邉 亮彦[†,§] 村上 聡一朗[†,§] 宮澤 彬[‡,¶,§] 五島 圭一[†,§] 柳瀬 利彦[*] 高村 大也[†,§] 宮尾 祐介[‡,¶,§]
[†]東京工業大学 [‡]総合研究大学院大学 [*]日立製作所 [¶]国立情報学研究所 [§]産業技術総合研究所

{watanabe,murakami}@lr.pi.titech.ac.jp, goshima.k.aa@trn.dis.titech.ac.jp,
toshihiko.yanase.gm@hitachi.com, takamura@pi.titech.ac.jp,
{miyazawa-a,yusuke}@nii.ac.jp

1 はじめに

テキストの読みやすさを解析する技術は，小論文の採点に関する研究 [1] や，金融テキストの読みやすさと企業の収益の関係を分析する研究 [7] など，多くのタスクで活用されている．テキストの読みやすさを表す絶対的な指標は存在せず，これまでに，Gunning Fog Index [2] や，文の容認度 [6] など，多くの間接的な指標が提案されてきている．そのため，どのような指標を採用するかは，研究の応用によって変える必要がある．しかしながら，それらをサポートするツールは整備されておらず，多くの場合，先行研究で提案されている手法を再実装するなどの労力を要する．そこで，本稿では，テキストの読みやすさに関する指標を自動的に計算し，効率的に実験を行うためのツール，Text Readability Features (TRF) を提案する．TRF を用いれば，分析を行いたいテキストを入力するだけで，種々の読みやすさに関する指標のスコアを計算し，出力することができる．本稿では，TRF がサポートする指標，および TRF を用いた実験について述べる．TRF は，https://aistairc.github.io/plu/で公開予定である．

2 代表的読みやすさ指標と既存研究

代表例な指標の一つとして，Gunning Fog Index [2] が挙げられる．Gunning Fog Index は，英語テキストを最初に読解したときに，内容の理解に必要な学校教育年数を表した指標であり，文の単語数の平均と 3 つ以上の音節を持つ単語の割合から求められる．また，単語の平均音節数を用いる Flesch-Kincaid スコア [5] も提案されている．これらの指標は，読みやすさ指標の研究の初期に提案されたものであり，単語やフレーズの難易度と構文の複雑さに着目した指標である．

一方，近年では，機械学習を用いてテキストの読みやすさを推定する研究が行われている [11, 4]．Schwarm ら [11] は，対象学年別に記述されたニュース記事の分類を行っており，言語モデルのパープレキシティや，構文木の深さなどの素性が有効であったと報告している．Kate ら [4] は，より多様な素性を用いて実験を行い，パープレキシティに加えて，文中の名詞節や動詞節数の平均や動詞節を含まない文の割合などの統語情報に基づく素性，Out-of-vocabulary (OOV) の割合や代名詞，機能語の割合などの語彙的な素性を組み合わせることで，読みやすさ推定の精度が向上したことを報告している．これらの研究は，語彙に基づく指標，統語情報に基づく指標，言語モデルに基づく指標を用いる点で共通している．

Tanaka ら [12] は，テキスト中で使用される単語の具体度からテキストの具体性を推定する手法を提案しており，テキストの具体性が分かりやすさに影響を与えることを示している．また，Lau ら [6] は，言語モデルによって算出された文の生成確率を，文長，単語の出現頻度に基づいて，文の容認度に変換する手法を提案している．ここで文の容認度とは，ある文が与えられたときに，その文を正しい文として容認できる度合いのことであり，文の読みやすさを文法性の観点から捉えることが可能となる．これらの研究は，テキストの読みやすさを直接推定するものではないが，テキストの読みやすさを測る指標として有用なものである．

TRF では，上述した読みやすさ推定の研究において広く用いられる，テキストの長さなどの基本的な指標，語彙や統語情報，言語モデルに基づく指標だけでなく，テキストの具体性や文の容認度に関する指標など，幅広い指標をサポートするように設計した．

3 TRF がサポートする指標

TRF では，大きく分けて 4 種類の，テキストの読みやすさに関する指標をサポートしている．表 1 に，TRF がサポートする指標の一覧および指標のスコア例を示す．ここで，スコア例は，表 2 中のテキストを解析した結果である．また，表 2 中のテキストは，BCCWJ コーパス [1] から引用したものであり，文書 1 は教科書，文書 2 は白書に関するテキストである [2]．以下，各指標について種別ごとに説明する．

[1] http://pj.ninjal.ac.jp/corpus_center/bccwj/
[2] サンプル ID はそれぞれ OT31_00007，OW4X_00076 である．

表 1: TRF がサポートする指標一覧および指標のスコア例

通番	指標名	指標の説明	対象	スコア (文書 1)	スコア (文書 2)
1	平均文長	各文に含まれる形態素数の平均		21.0	48.7
2	文数	テキストに含まれる文の総数		6	3
3	トークン数	テキストに含まれる単語のトークン数		126	146
4	タイプ数	テキストに含まれる単語のタイプ数		65	79
5	語彙の難易度	テキストに含まれる単語の難易度の割合	初級前半	0.048	0.000
			上級前半	0.000	0.075
6	語種	テキストに含まれる単語の語種の割合	漢語	0.198	0.377
			外来語	0.000	0.014
7	品詞	テキストに含まれる単語の品詞の割合	名詞	0.222	0.315
			動詞-非自立可能	0.032	0.089
8	語彙の具体度	テキストに含まれる名詞の上位語数の平均		6.14	5.36
9	jReadability	文書難易度判定システムに基づくスコア		10.5	8.91
10	仮定節	仮定節が含まれる文の割合		0.333	0.000
11	係り受け木の深さ	各文の係り受け木の深さの最大値の平均		2.5	6.0
12	モダリティ	各種モダリティが含まれる文の割合	認識-推量	0.167	0.000
13	言語モデルの尤度	各文の言語モデルの対数尤度の平均		-2.48	-1.77
14	容認度	各文の容認度スコアの平均		5.06	5.75

表 2: 平易なテキストと難解なテキストの例

文書 1: 平易なテキスト（207 字）
安全な食料は，どのように選べばよいのでしょうか。のりおさんの学校では，給食に使う材料を同じ市の農家から買っています。地域の農業をさかんにするためですが，それだけでなく，安全で，新せんな材料が手に入るからです。農家の伊藤さんと栄養士の後藤さんから話を聞きました。給食は，安全性が高い材料を使ってつくります。近くの農家にたのめば，給食用に，安全な方法でつくってもらうことができます。
文書 2: 難解なテキスト（235 字）
健康や安全についての国民の関心の高まりを背景に，健全な食生活や安全な食料品を志向する動きが強まっている。成人病や肥満を防ぐ観点から低脂肪，低塩分の食品の購入が増加しており，カルシウムやミネラル等微量栄養素や植物繊維を強化した食品も多く出回るようになっている。また，いわゆる有機栽培による農産物等への関心が高まっているほか，食料品の輸入が増加していることから，その残留農薬や食品添加物等輸入食品を含めた食料品の安全性の確保が強く求められるようになってきている。

3.1 基本指標

表 1 中の，通番 1–4 に対応する指標は，読みやすさの指標として広く用いられる基本的な指標である．

「平均文長」は，構文の複雑さを測るための簡易的な指標として，Gunning Fog Index [2] や，読みやすさ推定の研究において古くから利用されている指標である．また,「トークン数」,「文数」,「タイプ数」は，テキスト全体の複雑さを簡易的に測る指標であり，特に「トークン数」は，Pitler ら [10] によってテキストの読みやすさと有意に相関があることが示されている．「トークン数」と「タイプ数」を組み合わせることで単語の重複度を算出し，Pitler らのように語彙の結束性に関する指標として利用することもできる．

3.2 語彙に基づく指標

表 1 中の，通番 5–9 に対応する指標は，語彙の情報を利用して算出される指標である．

「語彙の難易度」は，Gunning Fog Index などでも考慮される要素の一つであり，古くから利用される指標である．また,「語種」は語彙の語種の分布に着目した指標である．TRF では，語彙の難易度と語種の辞書として，日本語教育語彙表 [3] を利用した．日本語教育語彙表には，約 18,000 語からなる語彙が収録されており，各単語には，初級前半から上級後半までの 6 段階の難易度と，外来語や漢語といった計 5 種類の語種がラベル付けされている．解析を行う際は，これらの難易度や語種を持つ単語の割合を計算し出力する．

「品詞」は，Kate ら [4] において利用されている指標である．品詞の情報を得るために，TRF では MeCab[4] を用いて形態素解析を行い，テキストに含まれる単語の品詞ごとの割合を計算し出力する [5]．

「語彙の具体度」は，Tanaka ら [12] と同様に，テキスト中に具体度が高い単語が頻出した場合，内容の具体性が高く，読者が内容を理解しやすいという考え方に基づいた指標である．単語の具体度を測る際には，

[3] http://jhlee.sakura.ne.jp/JEL.html
[4] http://taku910.github.io/mecab/
[5] 辞書は UniDic を用いた．従って，TRF は UniDic の品詞体系に従い各品詞の割合を出力する．

日本語 WordNet [3] を利用した．TRF では，テキスト中の名詞に対して，日本語 WordNet を用いて上位語がいくつあるかを再帰的に計算し，単語の具体度とする．最終的に，テキスト中のすべての名詞に対して単語の具体度を計算し，これらの平均を出力する．

「jReadability」は，日本語文章難易度判別システム jReadability[6] に基づいた指標である [13]．jReadability では，テキストの平均文長や助詞率などの指標を，日本語の教科書データと BCCWJ コーパスから求めた重みで組み合わせることで，テキストの難易度をスコアリングしている．このとき，スコアが高いテキストほど，難易度が高いテキストとみなす．TRF では，テキストの jReadability スコアを算出し，出力する．

3.3 統語情報に基づく指標

表 1 中の，通番 10–12 に対応する指標は，テキストの統語情報に基づいて算出される指標である．

「仮定節」は，テキスト中の仮定節の情報に着目した指標であり，このような節の情報に着目した指標は，Kate ら [4] においても利用されている．TRF では，仮定形をとる形態素が含まれる文の割合を出力する．また,「係り受け木の深さ」は，文の係り受け構造に基づいた指標であり，Schwarm ら [11] の研究で有効であった指標である．TRF では，係り受け解析器として CaboCha[7] を用いて，テキスト中の各文の係り受け木の深さの最大値を求め，その平均値を出力する．

上記に加えて，TRF ではモダリティに基づく指標もサポートしている.「モダリティ」は，テキストに含まれる文末モダリティに着目した指標であり，たとえば，テキストの蓋然性などの情報を捉えることができる．TRF では，KNP[8] を用いてモダリティの検出を行い，モダリティの種別ごとに，対応するモダリティを持つ文の割合を出力する．

3.4 言語モデルに基づく指標

表 1 中の，通番 13–14 に対応する指標は，言語モデルに基づいて算出される指標である．

「言語モデルの尤度」は，Pitler ら [10] によってテキストの読みやすさと有意に相関があることが示されている指標であり，このような言語モデルの尤度に基づく指標は，Schwarm ら [11] も提案している．TRF では，テキスト中の各文の対数尤度を計算し，その平均を出力する．また,「容認度」は Lau らの手法 [6] に基づいて文の容認度スコアを計算し，テキスト中の文の容認度の平均を出力する．両指標を計算する際には，言語モデルとして，Recurrent Neural Network に基づく言語モデル [9] を用いた．このとき，学習データとして，日本語 Wikipedia の本文データ [9] を利用した．

4 実験

本節では，TRF を利用した実験について述べる．Loughran ら [8] は，Form 10-K[10] の読みやすさに着目し，読みにくい Form 10-K が提出された後は不確実性が大きくなるため，株価のボラティリティ[11] が大きくなることを報告している．そこで，本実験では TRF を用いて行える実験の応用例の一つとして，Loughran らの研究を日本語の有価証券報告書に対して適用し，有効性の検証を行った．

4.1 実験設定

実験では，有価証券報告書提出後の企業の株価のボラティリティが，大きくなるか否かを二値分類する．このとき，有価証券報告書提出後にボラティリティが大きくなる事例を正例，小さくなる事例を負例とする．素性としては，TRF を用いて算出した有価証券報告書の読みやすさに関する指標を利用する．また，株価のボラティリティは，企業の財務状況やマーケットの影響を受けて変動すると考えられるため，時価簿価比率などの財務指標，および提出日の時価総額などのマーケット指標を，基本素性として利用する．基本素性を利用したモデルをベースラインとし，ベースラインに対して，TRF による読みやすさ指標を追加したモデルとの比較を行う．今回のタスクは，株価のボラティリティが大きくなる場合に投資リスクが高まるため，正例の分類精度が高い分類器を構築できたほうが望ましい．そのため，モデルの比較を行う際は，正例と負例の F 値による比較を行う．

有価証券報告書データとして，有報リーダー [12] から収集した，2011 年 7 月から 2016 年 6 月までの有価証券報告書提出会社のデータを用いた．このとき，有価証券報告書のうち，投資家の投資判断に特に影響を与えると考えられる「企業の概況」,「事業の状況」,「提出会社の状況」,「経理の状況」の 4 セクションに対して，個別に読みやすさ指標を算出し，セクションごとに分類器を学習した．また，財務指標およびマーケット指標，株価のボラティリティを計算するために，日経 NEEDS 株式日次収益率データ（2016 年 6 月版）と，日経 NEEDS 一般事業会社企業財務データ (2016 年 10 月 12 日版) を用いた．分類器としては，Support Vector Machine (SVM) を用い，SVM の実装として LIBLINEAR[13] を用いた．また，素性の入力方法とし

[6] http://jreadability.net/
[7] https://taku910.github.io/cabocha/
[8] http://nlp.ist.i.kyoto-u.ac.jp/?KNP
[9] https://dumps.wikimedia.org/jawiki/
[10] 米国における有価証券報告書に相当する文書である．
[11] ボラティリティとは，価格の変動の激しさのことである．
[12] http://www.uforeader.com/v1/
[13] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

て，各指標の値を 10 段階に離散化して二値素性として入力する方法 (disc)，平均が 0，分散が 1 となるように標準化して入力する方法 (std)，トークン数などの大きさを表す指標を log スケールに変換し入力する方法 (log) の 3 種類を用いて実験を行った．

実験では，2011 年から 2014 年までの有価証券報告書提出企業のデータを訓練データ，2015 年から 2016 年までのデータをテストデータとし，企業単位で 5 分割交差検定を行った．データセット中の企業数の合計は 2,785 件，訓練データとテストデータの事例数は，それぞれ 7,149 事例，2,752 事例である．また，開発データにおいて正例の F 値が最大となるように，SVM の正則化パラメータ $C(\in [0.001, 0.002, ..., 16400, 32800])$ をチューニングした．

4.2 実験結果

表 3 に基本素性のみを使用したモデルの F 値，表 4 に読みやすさ素性を追加したモデルの F 値を示す．読みやすさ素性を追加したモデルでは,「事業の概況」と「経理の状況」のテキストを用いた場合の結果を記載している．また，表中の数値は F 値の平均である．

まず,「事業の概況」において，基本素性のみを使用したモデル (Base) と読みやすさ素性を追加したモデル (Read) を比較すると，素性の入力方法を disc および log とした場合に正例の F 値が向上している．しかしながら，両モデルの正例の F 値の最大値を比較すると，Read は Base を上回っておらず，対応する負例の F 値は，Read では減少していることが分かる．このような傾向は,「企業の概況」,「提出会社の状況」を使用した場合にも見受けられた．一方,「経理の状況」では，Read の正例の F 値の最大値は Base を上回っている．しかしながら，対応する負例の F 値は大幅に減少していることから，正例ばかりを出力する分類器が学習されていることがわかる．

本実験結果から，Loughran らが示したテキストとボラティリティの関係は，日本の有価証券報告書では明確な傾向が見られなかった．このような実験を行うためには既存の読みやすさ指標の調査や再実装が必要であるが，TRF を利用することで，実験を効率的に行うことができるといえる．

5 おわりに

本稿では，様々なテキストの読みやすさに関する指標を自動的に計算するツール，TRF を提案した．また，TRF によって計算した読みやすさ指標を利用して，企業における有価証券報告書提出後の株式のボラティリティを予測する実験を行った．TRF を利用することで，このような読みやすさ指標を利用した実験を，効率的に行うことができる．今後の課題としては，EntityGrid モデルを利用した指標などを実装することが期待される．

表 3: 基本素性のみを使用したモデルの F 値

入力方法	正例	負例
disc	0.352	**0.680**
std	**0.504**	0.678
log	0.385	0.653

表 4: 読みやすさ素性を追加したモデルの F 値

入力方法	事業の概況		経理の状況	
	正例	負例	正例	負例
disc	0.470	0.557	0.441	0.561
std	0.353	**0.688**	0.374	**0.634**
log	**0.504**	0.397	**0.543**	0.097

謝辞 この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (ＮＥＤＯ) の委託業務の結果得られたものです．

参考文献

[1] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, Vol. 4, No. 3, 2006.

[2] Robert Gunning. *The technique of clear writing*. McGraw-Hill, New York, 1952.

[3] Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. Development of the japanese wordnet. In *Proc. of LREC'08*, 2008.

[4] Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, and Chris Welty. Learning to predict readability using diverse linguistic features. In *Proc. of COLING'10*, pp. 546–554, 2010.

[5] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas for navy enlisted personnel. Technical report, Memphis Tenn: Naval Air Station, 1975.

[6] Jey Han Lau, Alexander Clark, and Shalom Lappin. Unsupervised prediction of acceptability judgements. In *Proc. of ACL-IJCNLP'15*, pp. 1618–1628, 2015.

[7] Feng Li. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, Vol. 45, No. 2, pp. 221–247, 2008.

[8] Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *The Journal of Finance*, Vol. 69, No. 4, pp. 1643–1671, 2014.

[9] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech'10*, pp. 1045–1048, 2010.

[10] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proc. of EMNLP'08*, pp. 186–195, 2008.

[11] Sarah E. Schwarm and Mari Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proc. of ACL'05*, pp. 523–530, 2005.

[12] Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proc. of WSDM'13*, pp. 475–484, 2013.

[13] 李在鎬. 大規模テストの読解問題作成過程へのコーパス利用の可能性. 日本語教育, Vol. 148, pp. 84–98, 2011.