日本語の難易度に関する特徴分析

前川 絵吏　　村尾 元
神戸大学 国際文化学研究科
{eri.maekawa, hajime.murao} @ mulabo.org

1 はじめに

近年，日本で生活する外国人が増加し，やさしい日本語が注目されている．やさしい日本語を自動生成するため機械翻訳の技術を利用した研究が活発である．機械翻訳では翻訳精度を BLEU [1]や SARI [2]で比較することが多い．BLUE や SARI は入力文，参照文との比較でスコアを計算するが，直接的にテキストの難易度を測定しない．そこで，本研究ではテキストの難易度を測るべく日本語の特徴量を分析し，分類モデルを構築した．将来的には，テキスト平易化の評価指標とすることを検討している．

キーワード：やさしい日本語，ランダムフォレスト，Permutation Impmortance

2 先行研究

劉ら [3]は語彙レベルと構文の複雑さからテキストの難易度を判定する手法を提案した．語彙レベルを表す尺度には，読解学習支援システムである「リーディング・チュウ太・語彙チェッカー」 [4]を使用している．リーディング・チュウ太・語彙チェッカーは旧日本語能力試験の出題基準に基づいて語彙レベルを判定する Web システムである．構文の複雑さを表す尺度には係り受け距離を採用し，これが長い文は難易度が高いとしている．実験では，日本語能力試験の問題集をコーパスとして利用し，重回帰分析によって難易度算定公式を提案している．

張ら [5]は日本人が定義する難易度ではなく，外国人の感覚に合った日本語の難易度を自動推定する手法を提案した．テキストの特徴量を抽出し，外国人が付与した難易度スコアを利用して，特徴量が有効かどうかを調査した．特徴量として，単語数，品詞の数，品詞の割合，分節数，係り受けの距離，係り受けの回数を抽出した．難易度自動推定では，線形回帰モデルにより難易度スコアを算出する式パラメータを推定している．これにより，任意の文から難易度スコアを推定することを可能にした．

本研究では，外国人にとっての難易度に着目している点では先行研究と同様であり特徴量の参考にした．しかし，現行の日本語能力試験では出題基準を公開していないため，本研究では語彙レベル等を特徴量とはしていない．それに相当する特徴量として出現頻度を設定した．

3 難易度推定のための特徴量の抽出

3.1 特徴量の求め方

テキスト 1 文から特徴量を抽出し，普通の日本語とやさしい日本語の文で比較する．特徴量として以下の値を用いる．

単語数　テキストに含まれる形態素の数．

漢字率　テキストに含まれる漢字の個数を文字数で割った値．

外来語率　全ての文字がカタカナである形態素の数．これを単語で割った値．

受身率　接尾語の「れる」もしくは「られる」の形態素の数．これを単語数で割った値．

サ変接続名詞率　品詞が「サ変接続名詞」である形態素の数．これを単語数で割った値．

副詞率　品詞が副詞である形態素の数．これを単語数で割った値．

読点率　読点の個数．これを単語数で割った値．

否定率　品詞が助動詞の「ない」「ぬ」「ん」の数．これを単語数で割った値．

出現頻度最大値　品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の数のうちの最大値．

出現頻度平均値　品詞が「名詞」「動詞」「形容動詞」「形容詞」である形態素の数のうちの平均値．

係り受け平均距離　文節単位での修飾文節と被修飾文節間の分節数を係り受け距離としたとき，1 文内で修飾−被修飾関係にある全ての文節間の係り受け距離の平均値．

係り受け最大距離　1 文内で修飾−被修飾関係にある全ての文節間の係り受け距離の最大値．

表 1 特徴量の統計量

特徴量	やさしい日本語				通常の日本語			
	平均値	最大値	最小値	標準偏差	平均値	最大値	最小値	標準偏差
単語数	24.281	60.000	6.000	8.362	38.471	285.000	3.000	17.096
漢字率	0.261	0.714	0.000	0.091	0.352	0.750	0.000	0.096
外来語率	0.029	0.250	0.000	0.038	0.021	0.429	0.000	0.030
受身率	0.001	0.100	0.000	0.007	0.009	0.125	0.000	0.016
サ変接続名詞率	0.023	0.188	0.000	0.032	0.068	0.500	0.000	0.045
副詞率	0.009	0.167	0.000	0.021	0.010	0.333	0.000	0.020
読点率	0.044	0.308	0.000	0.030	0.052	0.368	0.000	0.031
否定率	0.008	0.182	0.000	0.021	0.006	0.167	0.000	0.015
出現頻度最大値	3055.456	9046.000	8.000	2147.463	4720.032	9046.000	0.000	2127.299
出現頻度平均値	507.638	5145.125	4.333	386.579	707.079	3688.797	0.000	441.815
係り受け平均距離	1.990	4.125	1.000	0.565	2.454	7.324	0.000	0.797
係り受け最大距離	6.312	20.000	1.000	3.526	10.940	120.000	0.000	6.796
係り受け被修飾数	2.754	7.000	1.000	0.930	3.453	11.000	0.000	1.235

係り受け被修飾数　1 文内における全ての文節についての被修飾数の最大値.

3.2 実験データ

3.2.1 データの取得

実験では，通常の文として NHK NEWS WEB[i]と，やさしい日本語の文として NEWS WEB EASY[ii]の記事からテキストを抽出した．2020 年 7 月 9 日から 2020 年 12 月 8 日までに投稿された 329 件の記事を Web スクレイピングで収集した．収集するテキストは Web ページのニュース本文のみである．タイトル，記事の投稿日時，注目ワード，別の記事へのリンクテキストやバナーの文字などは含まない.

得られたテキストから，NEWS WEB EASY 2,600 文，NHK NEWS WEB 2,600 文をランダムに選び，合計 5,200 文を教師データとした．このうち学習データは 5,000 文で，NEWS WEB EASY の文が 2,501，NHK NEWS WEB の文が 2,499，検証データは 200 文で，NEWS WEB EASY の文が 109，NHK NEWS WEB の文が 91 とした.

3.2.2 特徴量の計算

NEWS WEB EASY のテキストから HTML タグとルビを取り除き，特徴量を計算した．求めた特徴量ベクトルにその文が通常文かやさしい日本語かを示すラベルを付与し，ラベルを推定するようモデルを学習した.

特徴量の基本統計量を表 1 に示す．なお，特徴量の標準化はしていない．特徴量のうち，最小値・最大値・平均値など基本統計量に差が見られるものは，「単語数」「漢字率」「サ変接続名詞」「副詞率」「係り受け最大距離」「出現頻度最大値」が該当する.

反対に，「外来語率」や「否定率」はほとんど差がないということがわかる.

4 特徴量に基づく分析

4.1 分析手法

4.1.1 ランダムフォレスト

ランダムフォレストは機械学習の手法のひとつで，分類や回帰問題に適用できる．ランダムフォレストは複数の決定木で学習する．その手順は次の通りである．まず，全ての学習データからそれぞれの決定木を学習するためのデータをランダムに選択する．これを用いてそれぞれの決定木を学習する．学習後のランダムフォレストを分類に利用する場合はそれぞれの決定木の出力からもっとも多い出力をランダムフォレストの出力とする.

決定木のアルゴリズムは特徴量と閾値を調整して決める．例えば「単語数は 20 より大きいかどうか」という条件で分岐した結果が，ラベル 0 と 1 のグループにクラス分けできていれば精度がよいと言える．今回の実験のように 2 値に分類する場合，条件によって分割したデータ群にラベルが混ざり合っている状態は不純度が高く，一方のラベルが集まっている状態を不純度が低いとする．親ノードより子ノードの不純度が低くなるよう学習していく.

[i] https://www3.nhk.or.jp/news/

[ii] https://www3.nhk.or.jp/news/easy/

4.1.2 Permutation Importance

Permutation Importance は特徴量の重要度を測る方法の一つで，それぞれの特徴量がどれくらいモデルの予測精度に貢献しているかを重要度とする．重要な特徴量の場合，その特徴量をランダムに並べ替えて学習すると正解率が低くなる．

計算の手順を示す．

1.データセットDを，学習済モデルで分類したときの正解率をsとする．

2.特徴量jごとに以下を計算する．

2.1 特徴量jをランダムに並べ替えて，データセットD_jを生成する．

2.2 データセットD_jを，学習済モデルで分類したときの正解率をs_jとする．

3.特徴量がK個あるとしたときの特徴量jの重要度PI_jを次の計算式で定義する．

$$PI_j = s - s_j \ (1 \leq j \leq K)$$

4.2 分析結果

4.2.1 ランダムフォレスト

ランダムフォレストでの学習には Python のパッケージである scikit-learn[iii]を使用した．パラメータは，決定木の数を 100 とし，それ以外は scikit-learn のデフォルトを使用した．

正解率は次の式で求める．

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

TP は真陽性，FP は偽陽性，FN は偽陰性，TN は真陰性とする．5,000 件のデータを学習させた結果，正解率(Accuracy)の平均は 0.825，分散は 6.44×10^{-5} であった．

4.2.2 PermutationImportance による特徴量の重要度

次に，どの特徴量がモデルの正解率に影響を与えているかを調べるために Permutation Importance を測定する．全ての特徴量について計算した結果を図 1 に示す．

図 1 より，「サ変接続名詞率」，「単語数」，「漢字率」，「受身率」，「出現頻度最大値」の 5 つの特徴量で特に誤差が大きくなった．これ以外の特徴量は，ランダムにシャッフルして学習しても精度は0.01以下しか変わらないことから，学習済モデルの推定精度に大きな影響を与えないと言える．つまり，先に挙げた 5 つの特徴量が重要であることが分かる．

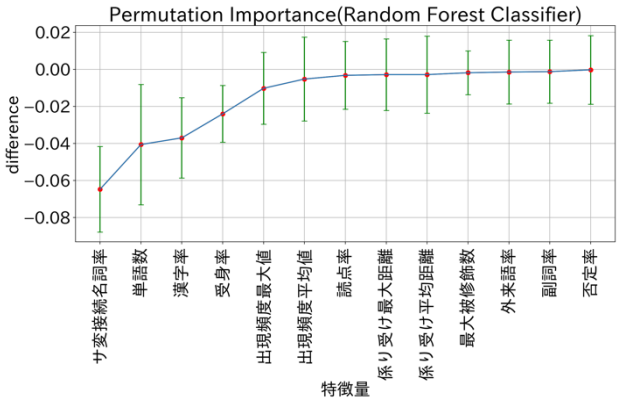

図 1 Permutation Importance の結果

5 分類モデルの実験

5.1 実験の目的

本研究では，やさしい日本語と通常の日本語のテキストデータから特徴量を抽出し，それぞれの特徴量が文の難易度にどれくらい影響を与えているかを分析した．分析では特徴量とラベルをランダムフォレストで学習し，分類モデルを生成した．分類モデルはテキストの難易度を評価するための判断として利用できることを期待している．それを確認するため，モデルの学習に使用していないコーパスでモデルを実験する．

5.2 実験に利用するコーパス

4.2.1 で生成したランダムフォレストの学習済モデルを，他の 4 つのコーパスから収集したテキストデータに適用し，ラベルを推定した．それぞれのコーパスの特徴と文例を示す．

(a) こどもコーパス　甲南大学知能情報学部の研究チームが集めたコーパス [6]である．10 歳～11 歳のこどもが本を紹介するブログに書き込んだテキストを収集したものである．

(b) 青空文庫　青空文庫で公開している作品である．宮沢賢治の「銀河鉄道の夜」を選んだ．

(c) ウェブ文書リードコーパス　京都大学の研究室で公開しているコーパスの 1 つで，Web 文書の冒頭 3 文を収集したコーパスである．ニュース記事，ブ

[iii] https://scikit-learn.org/stable/

ログ，商用ページなどさまざまな文書を含む．

(d) Wikipedia コーパス　共同作業で執筆されるインターネット上の事典である．日本語 Wikipedia のダンプデータからテキストを抽出した．Wikipedia の記法であるタグを取り除くため，Wikipedia Extractor を使用し，テキストのみを抽出した．全てのダンプデータを利用すると計算コストが膨大になるため，ファイルの先頭から 10 万行を抽出して使用した．

5.3 結果と考察

それぞれのコーパスから 500 文のテキストをランダムに選択した．テキストから特徴量を計算し，NHK NEWS WEB と NEWS WEB EASY で学習済のモデルで推定した結果を表 2 に示す．

こどもコーパス，青空文庫ではやさしい日本語と分類された文が 80%を超えていた．こどもコーパスは文が短いことや漢字が少ないことが分類に影響したと考えられる．青空文庫から選んだのは，宮沢賢治の「銀河鉄道の夜」だが，登場人物の会話で物語が展開するため，口語で短い文が連続する部分がある．特徴量のうち，単語数が少ないことが分類に影響したと考えられる．

ウェブ文書リードコーパスはやさしいと分類した文が比較的多かったが，500 文中 160 文は通常の日本語であると分類されている．複数の文書形態を含むコーパスのため，分析しにくいが，リード文は短く明確に伝えるために，1 文に多くの情報を入れなければならない．そのため，ひらがなより漢字を使って文の長さを短縮したことが，分類結果に影響したと考えられる．

Wikipedia コーパスは，通常の日本語と分類された文が最も多かった．文中の単語にリンクを設定でき，単語を補足する必要ないため文の難易度が高くなっていると考えられる．

表 2 コーパスを分類した結果

コーパス	やさしい日本語と分類した文の数	やさしい日本語率
(a)	424	84.8%
(b)	416	83.2%
(c)	340	68.0%
(d)	190	38.0%

6 おわりに

本研究では，やさしい日本語ニュースである NEWS WEB EASY と，そのニュースに対応する NHK NEWS WEB からテキストを抽出して，やさしい日本語の特徴量を分析した．ランダムフォレストで分類器を生成したところ，80%を超える正解率で学習できた．またランダムフォレストの木構造と Permutation Importance で特徴量の重要度を計り，「サ変接続名詞率」「単語数」「漢字率」「受身率」「出現頻度最大値」の 5 つが分類に影響を与えるということがわかった．

ランダムフォレストで生成した分類器で他のコーパスを評価したところ，こどもコーパスと青空文庫はやさしい日本語が 80%以上，ウェブ文書リードコーパスは約 70%がやさしい日本語であると分類した．Wikipedia コーパスは逆に 60%以上が通常の日本語であると分類した．得られた特徴量から難易度を判定できる可能性が示された．

謝辞 本研究は JSPS 科研費 19K12247 の助成を受けて行われました．

参考文献

[1] K. Papineni, “BLEU: a Method for Automatic Evaluation of Machine Translation,” In Proc. of ACL, pp. 311-318, 2002.

[2] W. Xu, “Optimizing Statistical Ma- chine Translation for Text Simplification.,” TACL, Vol. 4, pp. 401-415, 2016.

[3] 劉志宇ほか, “日本語を学習する外国人を対象とした日本語テキスト難易度推定手法,” 研究報告自然言語処理（NL）, Vol. 2012-NL-205, No. 11, pp. 1-5,, January 2012.

[4] 北村達也, “日本語読解学習支援システム「リーディング・チュウ太」,” 甲南大学紀要.知能情報学編, Vol. 6,No. 2, pp. 243-253, November 2013.

[5] 張萌ほか, “「やさしい日本語」作成支援のための日本語の難易度自動推定の検討.,” 研究報告自然言語処理（NL）, Vol. 2012-NL-206, No. 6, pp. 1-6, May 2012.

[6] 永田亮ら, “作 文履歴をトレース可能な子供コーパスの構築,” 自然言 語処理, Vol. 17, No. 2, pp. 2_51-2_65, 2010.