インドネシア語教育のための文章難易度判定モデルの構築

佐近優太[1]
[1]神田外語大学
ysakon322@gmail.com

概要

インドネシア語は市販の読解教材が少なく，大学等では書籍・雑誌・新聞記事など母語話者向けテキストを教材として用いることが多い．しかし，これらは難易度統制がなされておらず，学習者レベルに合わせた調整が難しい．インドネシア語の文章難易度自動判定には，教科書を対象に言語的特徴量とBERT 等の埋め込み表現を併用した先行研究があるが，難易度指標のないニュース記事への適用可能性は十分に検証されていない．本発表ではニュース記事を対象に難易度自動判定の可能性を検討し，ファインチューニング済み BERT モデルでも新聞記事では十分な精度が得られないことを示す．

1 はじめに

インドネシア語はオーストロネシア語族マレー・ポリネシア語派に属するマレー語を基盤とし，インドネシアの国語・公用語として広く用いられている．インドネシア語は学科や専攻における第一外国語科目の他，専攻語ではない言語科目としても，多くの大学でカリキュラムに加えられている[1][2]．そうした中で文法解説書は日本語で書かれているだけでも多くあり，また文法記述書の作成の試みも行われている[3]．一方で文法を一通り学んだ後に使用する読解教材についてはその数が乏しい．上述のように多くの大学でインドネシア語の講義が開講されているが，以前と比べれば教員・教育機関の数は減り，独自の教材を開発することが難しくなっている[3]．そうした状況の中で，多くの大学では読解教材としてインドネシア語で出版された書籍や新聞のニュース記事を採用している[4]．

しかし，母語話者向けに執筆された媒体を教育教材として採用する際には，難易度調整が困難であるという問題がある．学習者の習熟度と文章難易度の対応付けは，教員の主観的判断に依存して行われることが多く，その結果，教材の難易度が学習者にとって適切な水準から乖離し，十分な教育効果が得られない可能性がある．さらに，学習者側の「学びやすさ」の観点からは，難易度の不一致が学習意欲の低下につながる懸念もある．

インドネシア語においても言語学的特徴と BERT による文章埋め込みモデルを併用して文章難易度の客観的な提示を目指した研究は存在する[5]．しかしこれはインドネシアの小学校教育用にデザインされた教科書のデータを用いたものであり，教育への応用という前提がないニュース記事の難易度判定にそのまま応用できるかは不明である．

そこで本発表ではニュース記事に焦点を当て，インドネシア語文章難易度の自動判定モデルを構築することを目指す．

2 関連研究

文章難易度判定では，文法的特徴量や音節特徴量などの言語的特徴量，ならびに BERT などの事前学習モデルやファインチューニング済みモデルから得られる埋め込み表現特徴量を用いて予測を行うことが多い．文法的特徴量と音節特徴量を用いた研究としては Imperial[6]および Imperial et al.[7]が挙げられ，これらの研究では英語に加えてセブアノ語・フィリピノ語を対象に難易度判定の精度検証を行っている．具体的には，英語では 155 個[8]，セブアノ語・フィリピノ語では 54 個[9]の言語的特徴量を使用している．さらに，BERT による埋め込み表現抽出と組み合わせることで，英語では F1 スコア 0.893，セブアノ語でも 0.852 を達成したと報告されている．

一方，BERT 系の事前学習モデルをファインチューニングし，難易度予測を行う研究も多い．近年では，フランス語の CamemBERT をファインチューニングすることで，他手法より高い F1 スコアが得られたとの報告がある[10]．また東南アジア諸語に関しても，フィリピンで話される諸言語に加えベトナム語でも研究が進んでいる．例えば，ベトナム語の PhoBERT と文法的特徴量の組み合わせが提案され

ており[11], 英語のように大規模データベースが利用できない言語においても手法の有効性が示されている.

本研究の対象であるインドネシア語については, Yasin & Romadhony がインドネシア教育文化研究技術省の公開する小学校教科書を対象に文章分類を行っている[5]. 同研究では, 言語的特徴量に加え, BERT を基にしたインドネシア語事前学習モデルである IndoBERT[12]から抽出した埋め込み表現特徴量を用いてモデルを構築した. その結果, 埋め込み表現をサポートベクターマシーン(SVM)で分類した手法が最も高い精度を示し, F1 スコア 0.73 を得たと報告している. 一方, 文法的特徴量および音節特徴量による分類は, 特徴量数が少ないことも影響して, F1 スコアは 6 割前後にとどまるとしている.

以上を踏まえると, インドネシア語に関する先行研究の課題として, (1) ファインチューニング済みモデルを用いた検証が行われていない点, (2) 言語的特徴量の利用数が限定的である点, が挙げられる. そこで本研究では, 他言語で提案されている難易度判定手法がインドネシア語にも適用可能であるかを検証する. 加えて, 先行研究では難易度が事前に統制された教科書のみを対象としており, 難易度ラベルが厳密に設計されていない新聞記事のような文章への適用可能性は十分に検討されていない. したがって本研究では, 教科書とは性質の異なる新聞記事も対象とし, 既存手法の汎用性を検証する.

3 手法

3.1 データ

本発表はインドネシア語学習者向け教材に焦点を当てるため, インドネシアの教育・文化・研究・技術省の言語育成振興局がオンラインで提供する BIPA daring[i]のテキストを使用する(以下 BIPA). このうち一般学習者に向けに公開されているテキストは 7 段階に分かれるが, そのうち最新である 2019 年度版のレベル 2 から 7 のテキストの内, 読解教材部分を抽出した. レベル 1 は基本的な文法・表現の説明であるため除外している. データは一般向けと大学生向けの二種類があり, 扱う内容が異なるがレベルは同じように設定されている. それぞれの種類で各レベルにつき 10 編, 一部レベルのみ 10 編を越えて掲載されており, 全 123 編, 総語数は 28726 語である.

ニュース記事については, インドネシアで発行されている KOMPAS から 2020 年 1 月の記事を無作為に 121 編抽出した(以下 KOMPAS). 総語数は 79716 語である. インドネシア語を専攻している大学院生一人によって 5 段階で難易度を主観で判定してもらい, それを正解ラベルとしている.

3.2 手順

手順は BIPA, KOMPAS 共に同じである. 最初にデータを学習/訓練用とテスト用に 8:2 の割合で分けた.

それぞれのデータにおいて, 正解ラベルは先行研究[5]に合わせて三段階にまとめている. BIPA は二段階ごとに一つにまとめ(レベル 2, 3: A/レベル 4, 5: B / レベル 6, 7: C), ニュース記事はラベルの不均衡を考慮し, 1,2 に判定されたもの(A), 3 に判定されたもの(B), 4,5 に判定されたもの(C)の三段階とした. KOMPAS の記事の選出はランダムで行っているため, ラベルに偏りが生じている(表 1).

表 1 KOMPAS のレベル毎の記事数

A	B	C
44	55	22

最初に事前学習モデルである indolem/indobert-base-uncased を学習/訓練データによってファインチューニングをした. ファインチューニングはデータの少なさを考慮して 5 分割交差検証を行っている. その後言語的特徴を別途作成し, BERT のみ, 言語的特徴のみ, BERT+言語的特徴の 3 つのパターンでテストデータに対して精度の検証を行った. 予測は logistic regression(LR), 多層パーセプトロン(MLP), Random Forest(RF), SVM の 4 つの分類器を用いた.

言語的特徴量である文法的情報及び音節情報の特徴量は, Imperial & Ong[9]の中からインドネシア語でも利用できる特徴量を選択し, 加えて接辞付与率, 非インドネシア語基本語[13]を加え, 合わせて 30 個の言語学的特徴量を作成した(Appendix A). 実装はすべて Google Colaboratory 上で行っている[ii].

[i] https://bipa.kemendikdasmen.go.id/belajar_eng.php

[ii] 詳細は https://github.com/YutaSakon を参照

4 結果

4.1 BIPA

最も高い精度を示したのは，ファインチューニング済み BERT モデルから抽出した埋め込み表現特徴量を用いた手法である(表 2).

表 2 テストデータに対する予測精度(BIPA)

	BERT	言語的特徴	Combined
LR	0.703492	0.726902	0.703492
MLP	0.742680	0.687895	0.664680
RF	**0.790065**	**0.730509**	**0.742680**
SVM	0.709714	0.583429	0.742680

先行研究では事前学習モデルをそのまま用いていたのに対し，本研究ではファインチューニングを施すことで精度が向上することを確認した．データ数が限られる条件下でも性能改善が観察されたことは，他言語における知見と整合的であり，当該手法がインドネシア語においても一定の有効性を有することを示唆する．さらに言語的特徴量についても先行研究を上回る精度が得られた．先行研究では言語的特徴量のみを用いた場合に精度が低下することが報告されていたが，本研究では特徴量を拡充したことで精度の向上が確認された．図 1 に，比較的高い精度を示した RF における特徴量重要度を示す.

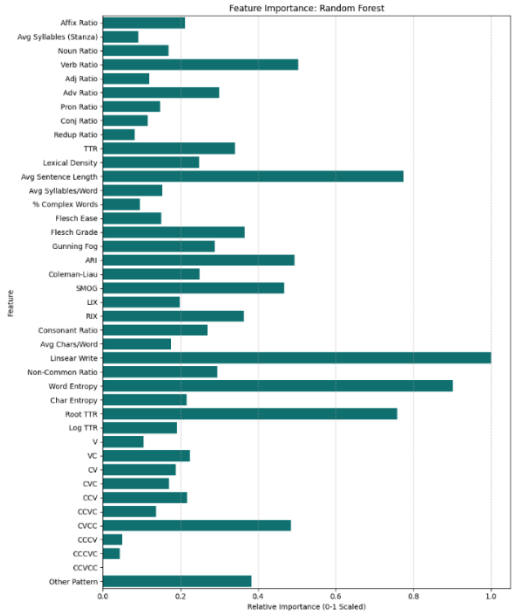

図 1　RF における特徴量の重要度 (BIPA)

図 1 より Linsear Write, Word entropy, Root TTR, Avg. sentence length の寄与が大きいことが分かる．Linsear Write は音節数に関連する可読性指標であり，3 音節以上の語が用いられている程度を反映する．インドネシア語では接辞により多様な語形成が可能であり，複数の接辞が組み合わさる場合には語が長くなり，音節数も増加しやすい．接辞自体はインドネシア語で頻繁に用いられるため，接辞が付与される割合（Affix ratio）そのものは難易度に大きく影響しない．一方で，接辞の複合によって語が複雑化することで，文章難易度が上昇すると考えられる．
Word entropy および Root TTR はいずれも type-token 比に基づく指標であり，語彙の多様性を表す．さらに Avg. sentence length（平均文長）と併せて，これらは言語を問わず文章難易度判定において重要な指標であると言える

図 2 に, BERT 特徴量と言語的特徴量を併用し, RF で予測した際の予測/正解ラベルの対応を示す.

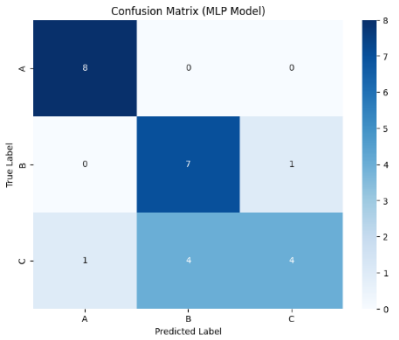

図 2　RF における予測/正解ラベル(BIPA)

ここからは難易度が高くなると予測が難しくなることがわかる. BIPA は大学生または社会人が対象となるプログラムであるため，中級レベル(レベル 4 以降)でも高度なトピックが扱われる．そうした内容の難しさが影響を与えていると考える.

4.2 新聞記事(KOMPAS)

次に KOMPAS の結果を示す.

表 3 テストデータに対する予測精度(KOMPAS)

	BERT	言語的特徴	Combined
LR	0.485333	0.486857	0.485333
MLP	0.485333	0.518454	0.485333
RF	**0.528448**	**0.541178**	**0.528448**
SVM	0.485333	0.452269	0.528448

全体として，KOMPAS は BIPA と比較して予測精度が低かった．この要因として，KOMPAS の記事は教育教材として作成されたものではなく，難易度統制や学習者向けの配慮が前提となっていない点が挙げられる．加えて，本研究の手法は BIPA においても難易度が高いクラスの分類では精度が低下することが確認されており，難易度の上昇に伴って誤分類が増加する傾向が示唆される．KOMPAS は母語話者を主対象として書かれているため，語彙や構文の複雑さが相対的に高く，これも精度低下の一因となったと考えられる．分類器間の比較では，BIPA と同様に KOMPAS においても Random Forest（RF）が最も高い精度を示した．一方で，BIPA とは異なり，KOMPAS では言語的特徴量による分類が最も高い精度となった．ただし，BERT 特徴量との差は大きくはなく，両者は概ね近い性能を示した．図 3 に RF における特徴量重要度を示す．

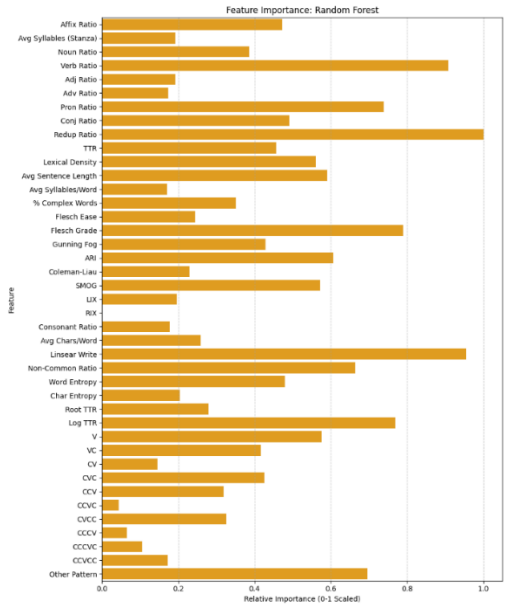

図 3　RF における特徴量の重要度（KOMPAS）

type-token 比などは依然として重要であるが，その他重複語の比率と動詞の比率が重要な特徴量として挙げられている．このうち重複語は rumah-rumah など同じ単語を二回繰り返すような形式を示す．この場合は rumah「家」という単語が続いて「家々」といった複数の意味を表すが，動詞や副詞を重複させることも可能であり，その場合は複数以外の様々な意味を表し得る．そのため学校文法でも比較的難易度の高い項目として扱われており，教育教材である BIPA よりも重要な指標として抽出されたと考える．一方で，相対的に平均文長の重要度は下がっている．これは教育教材では低難易度帯で文長が抑えられるものの，新聞記事ではそうした配慮が行われないためである．

次に，BERT 特徴量+言語特徴量を基に RF で予測を行った際の予測/正解ラベルの対応を示す．

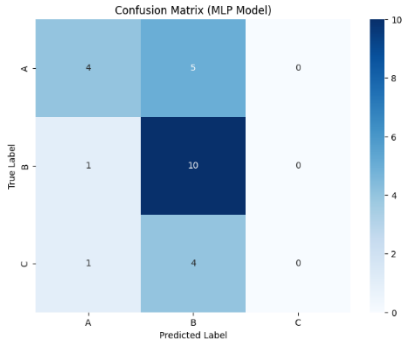

図 4　RF における予測/正解ラベル(KOMPAS)

この図からは，難易度にかかわらず予測がうまく行われていないことがわかる．また BIPA と異なりラベルが B(中難易度)に偏っている影響も見られる．

4 おわりに

本研究の要点は次の二点にまとめられる．

・インドネシア語においても，当該言語の事前学習モデルをファインチューニングすることで，より高い精度が得られる．
・難易度が高く，かつ事前に難易度統制がなされていない文章については，埋め込み表現特徴量・言語的特徴量のいずれを用いても予測が難しく，新聞記事に対する難易度判定には課題が残る．

高難易度の文章を対象とした先行研究として，ベトナム語の研究[11]では大学生向け教材という比較的難易度の高いテキストを扱っているにもかかわらず，比較的高い精度を得ている．本研究の結果との差異は，難易度ラベルの付与主体が専門家であるか学生であるかという違いに起因する可能性がある．本研究は学習者にとっての「学びやすさ」に焦点を当てているため，今後は難易度判断を行う学生の人数を増やし，アノテータ間の判断の一致度を検証する．さらに，ラベル付与基準の明確化と統一を進めることで，予測精度の向上を目指す．

謝辞

本研究は JSPS 科研費 JP23KJ0343 の助成を受けたものです.

参考文献

[1] 浦野崇央, シシリア・タントリ・スルヤワティ. 留学に際する教育カリキュラム統合の可能性: 摂南大学生によるストモ博士大学での学修を事例として. インドネシア言語と文化 23, pp. 31-42, 2017.

[2] Sri Budi Lestari. 外国語としてのインドネシア語―専攻語ではない言語科目としての教育実践からの提案―. 思言: 東京外国語大学記述言語学論集 20, pp. 125-142, 2025.

[3] 原真由子, 森山幹弘, 降幡正志. インドネシア語基本文法の記述：教材作成のための共同研究からの報告. インドネシア 言語と文化 23. pp. 7-30. 2017.

[4] 井口由布. マレー語・インドネシア語教育の実践－APU 方式の確立へむけて－, Ritsumeikan Center for Asia Pacific Studies, pp. 117–134. 2007.

[5] Sanding Adhieguna Rachmat Yasin and Ade Romadhony. Building an elementary Indonesian textbook readability baseline model. **10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)**, pp. 1–6, 2023.

[6] Joseph Marvin Imperial. BERT embeddings for Automatic Readability Assessment. **Proceedings of Recent Advances in Natural Language Processing**, pp. 611-618, 2021.

[7] Joseph Marvin Imperial, Lloyd Lois Antonie Reyes, Michael Antonio Ibanez, Ranz Sapinit, and Mohammed Hussien. 2022. A Baseline Readability Model for Cebuano. **Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)**, pp. 27–32, 2022.

[8] Sowmya Vajjala and Ivana Lucic. ́OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. **Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 297–304, 2018.

[9] Joseph Marvin Imperial and Ethel Ong. 2021a. Application of lexical features towards improvement of filipino readability identification of children's literature. arXiv preprint arXiv:2101.10537.

[10] Wafa Aissa, Thibault Bañeras-Roux, Elodie Vanzeveren, Lingyun Gao, Rodrigo Wilkens, and Thomas François. Assessing French Readability for Adults with Low Literacy: A Global and Local Perspective. **Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing**, pp. 20517–20539, 2025.

[11] Hung Tuan Le, Long Truong To, Manh Trong Nguyen, Quyen Nguyen, and Trong-Hop Do. A study of Vietnamese readability assessing through semantic and statistical features. **Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation**, pp. 71–81, 2024.

[12] Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-Trained Language Model for Indonesian NLP. **Proceedings of the 28th International Conference on Computational Linguistics**. pp.757–770. 2020.

[13] Ivan Lanin and Jim Geovedi and Wicak Soegijoko. Perbandingan distribusi frekuensi kata bahasa Indonesia di Kompas, Wikipedia, Twitter, dan Kaskus. **Proceedings of Konferensi Linguistik Tahunan Atma Jaya Kesebelas (KOLITA11)** pp. 249-252, 2013.