

CEFR 準拠 CDS 自動推定システムの精度改善を目的とした 推定手法改良の試み

里見 陸[†] 宮崎 佳典[‡] 谷 誠司^{†††} 安 志英^{††††} 元 裕環^{†††††}

静岡大学[†] 静岡大学[‡] 常葉大学^{†††} 群山大学^{††††} 高麗大学^{†††††}

1. はじめに

CEFR とはヨーロッパ言語共通参照枠の略で、欧州の言語教育・学習の場で共有する枠組みとして、2001年に欧州評議会が発表した[1]。欧州をはじめとする世界の各言語で実際に利用されており、現在、英語やヨーロッパ各国の言語を含め、多くの言語に対応している。CEFRには「読む、書く、話す、聞く」といった技能項目があるが、本研究では「読む」能力を対象とする。CEFR レベルは難易度毎に A1, A2, B1, B2, C1, 最高レベルの C2 に分けられているが、2017年にはCEFRを補完するものとして[2]が公開され、初学習者向けに、A1 レベルよりもさらに初級段階のレベルとして PreA1 レベルが追加された。

各 CEFR レベルは「その言語を使って何ができるか」という形で言語力を表す Can-Do Statement (以下 CDS) を複数用いて構成されており、PreA1 から C2 レベルまでの CDS は合計 38 個ある。ただし、CEFR レベルや CDS の中にはその言語を母語話者とする人にとっても難易度の高いものも存在し、本研究では B2 レベルに属する CDS の内の 1 つと、C1, C2 レベルに属する CDS を除いた計 34 個の CDS を研究対象とすることとしている。CDS は、CEFR レベルとは別に、それぞれが表す文章を読むのに必要な能力の種類によって 5 つの「下位尺度」のいずれかに属する。下位尺度が表す能力としては、「情報や要点を読み取る能力」、「手紙やメールを読む能力」等といったものがある。そのため、CDS は CEFR レベルと下位尺度の組み合わせで表現でき、文章の難易度と種類を内包した指標であると言える。

2. 本研究の目的

日本語教育へ CEFR を活用することに関して、日本語の CEFR 準拠テキストコーパスが作成されている例が管見の限りほとんど見受けられないことから、網羅的に研究されている事例は現時点において少ないと考えられる。これに対し、我々研究グループは日本語文章の CDS を自動推定するシステムの作成に取り組んでいる。特に本発表では、先行研究におけるシステムの推定手法に対して新たなアプローチを適用することでシステム推定精度の改善を目的としている。

3. 先行研究の推定手法

先行研究[3]では、PreA1 レベルの文章とそれ以外の文章には難易度や特徴に大きな差があると考えられたため、まず文章が PreA1 レベルかそれ以外の A1-B2 レベルの文章かを 2 値分類することをステップ 1 とし、ステップ 2 としてそれらに属する CDS のみを対象として CDS を推定するという手法を用いた。最新の先行研究[4]においても同様の手法を用いている。

この際、2 値分類と、A1-B2 と推定された文章に対する CDS 推定には同じ特徴量が用いられているが、PreA1 と推定された文章に対する CDS 推定にはそれらとは異なった特徴量が用いられている。

推定精度の算出は、CEFR の知識を有する、外国にて日本語教育の経験がある 10 名から収集した CDS 情報付きの 555 例文を用いた 10 分割交差検定によって行う。教師データを用いて 2 値分類のモデル、並びに A1-B2 と推定された文章の CDS 推定のモデル、PreA1 と推定された文章の CDS 推定のモデルを作成し、それにテストデータを入力して、本来のテストデータに付与された CDS と推定された CDS を比較して精度を算出する。

テストデータの正解 CDS と推定された CDS から適合率と再現率を算出する。さらに、適合率と再現率の調和平均である F 値を算出し、これを最終的な精度として考えることとする。すると、先行研究⁽³⁾の推定手法における各値は表 1 のようになった。

An Attempt to Infer CEFR Can-Do Statements Based on a Japanese Document Classification Method for Better Precision by Improving Estimation Method

[†]Riku Satomi, Department of Informatics, Graduate School of Integrated Science and Technology, Shizuoka University

[‡]Yoshinori Miyazaki, College of Informatics, Shizuoka University

^{†††}Seiji Tani, Tokoha University

^{††††}An Jiyoun, Kunsan University

^{†††††}Won Yukyoung, Korea University

表1 先行研究の推定手法の精度

適合率	再現率	F 値
44.5%	31.8%	37.1%

また、ステップ1の2値分類の精度も10分割交差検定により算出すると、値は表2のようになった。

表2 先行研究のステップ1の推定精度

適合率	再現率	F 値
97.9%	98.5%	98.2%

4. 本研究の提案手法

先行研究の手法では PreA1 とそれ以外では難易度や特徴に大きな差があると考えられたため2値分類をする手法を取っていたが、A1 から B2 レベルに関しても、各レベル間には難易度や特徴には差があると考えられる。そこで、ステップ1として文章の CEFR レベルを推定した後、ステップ2としてその CEFR レベルに属する CDS のみを対象とした推定をするという手法を提案する。また、本発表では先行研究に対して推定手法を変更することによる精度の変化を得ることを目的とするため、CEFR レベルや CDS の推定には先行研究の手法における2値分類等と同じ特徴量を用いている。

その結果、10分割交差検定による精度は表3のようになり、提案手法の精度は先行研究に比べて適合率、再現率が共に向上し、結果 F 値も向上した。

表3 提案手法の精度

適合率	再現率	F 値
56.7%	40.3%	47.1%

また、表2と同様にステップ1の2値分類の精度も10分割交差検定により算出すると、値は表2のようになった。この結果により、ステップ1の精度は先行研究の手法の方が高いことが示された。しかしながら最終的な CDS 推定の精度は提案手法の方が高いことから、対象とする分類の数が少ない方が推定の精度は向上する傾向にあり、提案手法の方が各推定における分類数のバランスが良かったため最終的な精度が高くなったという仮説が考えられる。

表4 提案手法のステップ1の推定精度

適合率	再現率	F 値
79.8%	82.8%	81.3%

ここまでは、ステップ1とステップ2の各推定において参照する特徴量は全て同じとしていたが、提案手法では推定モデルが増えたことによりそれぞれの推定により適した特徴量を選定しやすくなり、さらなる精度の向上にも期待できる。そこで、現在は各推定においてより精度が向上する特徴量を考察し、適用する実験を行っている。実験的に、文章に含まれる数字の数を特徴量として適用してみたところ、CEFR レベルの推定と B1, B2 レベル内の CDS 推定に有効であるという結果が得られ、CDS 推定の精度は1%程度上昇した。

5. まとめと今後の展望

本研究では、現在開発を行っている CEFR 準拠 CDS 自動推定システムについて、先行研究の推定手法に対して新たなアルゴリズムを適用することで精度の改善を行うことができた。加えて、推定における分類数と精度の関係性も示唆されたことから、推定の精度を向上させるのにより適した分類数、分類方法があるのかを調べるのが今後の展望として挙げられる。

また、各 CDS が対象とする文章の説明を読むなどすることでステップ1、ステップ2の各推定により適した特徴量を考察することも今後の展望となる。

参考文献

- [1] Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Cambridge University Press (2001)
- [2] Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment Companion Volume with New Descriptors (2017)
- [3] Vuong Hong Duc, Huynh Nguyen Tra My, Y. Miyazaki, S. Tani, A Pilot Study to Infer CEFR Can-Do Statements Based on a Japanese Document Classification Method Including the Pre-A1 Level, Proceedings of 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 294-297 (2020)
- [4] CAO HOAI GIANG, 宮崎佳典, 谷誠司, 安志英, 元裕環, CEFR に対応した日本語例文自動分類システムの BERT 適用による精度改善の試み, 韓国日本学会第102回国際学術大会 (2021)