CEFR準拠教科書における語彙・構文の特徴分析とレベル自動分類

水嶋海都 †, 荒瀬由紀 †, 内田諭 ‡
† 大阪大学大学院情報科学研究科, ‡ 九州大学大学院言語文化研究院
†{mizushima.kaito, arase}@ist.osaka-u.ac.jp, ‡uchida@flc.kyushu-u.ac.jp

1 はじめに

国際化やインターネットの普及に伴って世界共通言語のひとつである英語を使用する機会はますます増えており、ノンネイティブ話者の英語学習を補助することが重要となっている。言語教育において、学習者のレベルに応じた教材を用いて教育を進めることは不可欠である。しかし、言語教育者が学習者の外国語能力に適合した教材を作成することは多くの時間と労力を費やす作業であり、さらに言語能力レベルに沿った言語特性についての深い理解が求められる。

そこで本研究では、言語能力レベル別の語彙・構文的特徴（Text Profile）を分析し、その特性を明らかにする。さらに抽出した特徴を用いて英語エッセイの自動分類に取り組む。英文の分類を行うことで英語学習者の英語レベルを判定することができ、また、言語教育者が学習者の言語能力レベルに適した教材を準備する補助ができる。本研究では言語学習者の外国語能力レベルを表す国際基準である Common European Framework of Reference for Languages （CEFR）準拠の教科書のデータを用いて各レベル間の分類実験を行うとともに、各レベルにおいて特徴的な言語的特性の分析をする。

また、隣接する CEFR レベル間の 2 値分類を行う際の各特徴量の重要度を明らかにすることで個々の CEFR レベルにおける特有の言語特徴や難易度の違いによる文章構造の差異を詳細に分析する。

2 関連研究

ライティングの習熟度を推定する研究に小林ら [5] がある。大学生に制限時間 30 分で TOEFL 形式のライティング課題を与え収集した 69 本のエッセイをデータとして使用し、目的変数には e-rater(R)[*1] で採点した 6 段階の評価を用いる。考慮した言語的特徴は総語数や異語数、平均単語長などの 12 種類であり、分類手法にはランダムフォレスト [2] を採用している。全体の推定精度は 62.32%であり、特に予測に有効であった特徴量は総語数と異語数である。

また、小林のライティングの習熟度推定に関する研究 [4] では日本人英語学習者コーパス (CEEJUS) を使用した。目的変数には各作文に付与された TOEIC テスト (R) 型の模擬試験に基づく 4 段階の習熟度を用いる。用いた言語的特徴としては、総語数や異語数、異語率など 20 種類である。また、これらの素性には異語率をベースに総語数の平方根を分母とする指標である Guiraud Index やテキストの Readability を表す Flesch-Kincaid grade level[3] も含まれている。分類手法にはランダムフォレストを用い、回帰分析を行った結果、異語率や平均文長が予測に大きく寄与していることが分かる。全体的な分類の精度は 58.23%である。

以上の先行研究においては文の構造を特徴づける上で有効な構文解析木を用いた素性は導入されていない。そこで、本稿では構文情報に基づく素性や先行研究において分類に効果的であると示されている単語長や単語難易度といった素性と構文解析木を組み合わせた素性を導入し、CEFR レベルにおいて区分されてる英文エッセイの言語特性を分析する。

3 Text Profile として用いる素性

表 1 に今回エッセイ分類で用いた素性の一覧を示す。

単語難易度に基づく素性 難易度の低い文章ほど文中の単語難易度は易しく、反対に難易度の高い文章ほど単語は難しくなると考えられる。文中で使用されている単語の難易度は学習者の語彙力を反映しており、文章の難易度を推定するにあたり重要な要素になり得る。

[*1] https://www.ets.org/erater

表 1: 素性一覧

素性	素性の概要
num_words	単語総数
word_length_1_3	単語長が 1〜3 文字である単語の割合
word_length_4_6	単語長が 4〜6 文字である単語の割合
word_length_7_	単語長が 7 文字以上である単語の割合
avg_word_length	平均単語長
types	単語の種類数（異語数）
TTR	type/num_words（異語率）
MLS	平均文語数
avg_difficulty	単語難易度の平均
A1_per	単語難易度が {A1〜C2} レベルであるものの割合
A2_per	
B1_per	
B2_per	
C1_per	
C2_per	
avg_difficulty	単語難易度の平均
sum_D_score	D_score の合計
avg_D_score	文あたりの平均 D_score
sum_L_score	L_score の合計
avg_L_score	文あたりの平均 L_score
avg_(カテゴリ)	文法カテゴリの文あたりの平均句数（全 13 項）
(カテゴリ)_per	文法カテゴリの全句数に対する割合（全 13 項）
avg_max_depth	各文に対する構文木の最大の深さの平均

本稿では、CEFR-J Wordlist[*2] を文中の単語の難易度を設定するために使用した。CEFR-J Wordlist において、単語レベルは見出し語と品詞の組み合わせで決定される。単語レベルは易しい順から A1、A2、B1、B2 まで区分されている。今回、CEFR-J で規定されている 4 段階のレベルに加えて、English Profile[*3] により規定されているさらに上位レベルである C1、C2 レベルの単語であり CEFR-J Wordlist に含まれていないものを追加した。

単語難易度に関する指標として、エッセイ中に使われている単語の平均難易度を表す avg_difficulty とエッセイに含まれる A1〜C2 それぞれの各 CEFR-J Wordlist 単語レベルの割合を用いた。離散値で設定されている単語難易度から連続値である平均難易度を計算するにあたり、A1〜C2 のカテゴリにそれぞれ 1〜6 の整数値を割り当てることで平均値を計算した。

構文情報に基づく素性　構文解析木は文章の情報として各フレーズや単語の品詞情報や文章全体の構文構造を含み、文章の特徴を分析する際に非常に有効である。そこで、本稿では構文解析木から得られる情報を使った素性を用いる。まず、エッセイに含まれる各文を構文解析器 Enju[*4] を用いて解析する。解析結果には Enju により文法カテゴリがそれぞれのノードに付与されており、ラベリングされた文法カテゴリをカウントする。各カテゴリに加えて内容語の**文あたりの平均句数** [avg_(カテゴリ)] と**全句数に対する割合** [(カテゴリ)_per] をそれぞれ素性として用いる。

また、より難易度の高い文ほど複雑な構造の構文木を持つと考えられ、構文木の深さは難易度の高い文ほど大きくなる。そこで、各文の構文木の最大の深さに対するエッセイあたりの平均値も素性として利用した。

単語難易度・構文構造を組み合わせた素性　先行研究や予備実験において単語数、構文解析木の深さ、単語長、単語難易度は分類において有効な素性であることが示されている。高難易度の文章であればあるほど使われる単語の難易度は高くなり、また同時に文あたりの単語数も増加し、構文構造も複雑になる。そこで、これらの素性を包括的に表現する新たな指標である D_score、L_score を提案する。

文に対する構文解析木が与えられたときリーフノードは文中の各単語に該当する。それぞれのリーフノードに対し、前述の単語難易度リストにしたがって、個々の単語に対応する難易度をそれぞれ割り当てることができる。本研究では D_score を式 (1) のように定義し、リーフノードに当たる各単語の難易度を考慮する。l は構文解析木のリーフノードを、$d(l)$ はルートノードからリーフノードまでの深さを、$f(l)$ はリーフノードにあたる単語の難易度をそれぞれ表す。

$$D\_score = \log_{10} \sum_{l} d(l) f(l) \tag{1}$$

また、単語難易度の代わりに、先行研究で有効であった単語長を用いることで L_score を式 (2) のように定義する。$g(l)$ はリーフノードにあたる単語の長さを表す関数である。

$$L\_score = \log_{10} \sum_{l} d(l) g(l) \tag{2}$$

D_score、L_score を用いることで複合的な文章難易度を表すことができる。エッセイの素性としてはエッ

[*2]『CEFR-J Wordlist Version 1.0』(2013) 東京外国語大学投野由紀夫研究室.

[*3]http://vocabulary.englishprofile.org/

[*4]http://www.nactem.ac.uk/enju/index.ja.html

表 2: 分類結果

分類クラス	1to1	5to1	10to1
5 classes	44.6%	53.2%	58.2%
4 classes	50.5%	59.0%	65.1%

セイ中の各文ごとの D_score、L_score のそれぞれの合計値、そして文あたりの平均値を用いる。

4 分類実験

4.1 データセット

本稿の実験で使用するデータは、CEFR 準拠で編纂されたレベル別コーパスである（A1:164,585 語、A2：278,750 語、B1：486,787 語、B2：582,763 語、C1：272,678 語、C2：29,471 語）。このコーパスではテキスト内の英文をユニット内の題材ごとに Reading, Listening など技能別に分類している。本稿ではこのまとまりをエッセイと定義し、それらをデータ単位とする。ただし、語彙リスト（Vocabulary）のセクション（176,843 語）は単語が列挙されたもので、文を構成しないため、除外した。また、C2 クラスのデータ数は他クラスと比べ、非常に少ないため、本稿では C1 と C2 を統合し、合わせて C クラスとして扱う。

さらに、上記のエッセイの 1 エッセイあたりの単語総数は全平均で約 104 語であり、比較的短い英文となっている。そこで、エッセイを 5 つまとめてひとつのデータとして扱う場合と 10 個をまとめて扱う場合の実験も行う。

4.2 CEFR レベルによる分類

本稿では分類手法としてランダムフォレスト [2] を用いる。分類クラスは CEFR レベルに基づく A1, A2, B1, B2, C の 5 クラス分類である。また、C クラスは分類が難しく全体の精度を大きく下げているため、C クラスを除いた A1、A2、B1、B2 の 4 クラス分類実験も行った。パラメータに関しては、チューニングを行ったところ分類精度が各パラメータに対してセンシティブでなかったこともあり、サンプリングする素性数は推奨とされる全素性数の正の平方根とし、木の数は 1000 とした。

分類結果として、ランダムフォレストにおける OOB(out-of-bag) accuracy を表 2 に示す。1to1、5to1、10to1 はそれぞれ 1 エッセイ、5 エッセイ、10 エッセイをデータ最小単位とした場合を表す。

各データセットにおける分類精度を見ると、10 エッセイを最小単位とした 10to1 が最も良い精度を示した。これは、語数が多いほど文章情報量が大きくなることで、エッセイを特徴づける特徴量が獲得できたために、分類精度が向上したと考えられる。実際に 1to1 から 10to1 にかけて最小データ単位を大きくすればするほど精度が良くなっている。

C クラスを含めた 5 クラス分類の結果と C クラスを除いた 4 クラスの分類結果を比べてみるといずれのデータセットにおいても 5%〜7% 程度、4 クラス分類の方が良い結果となった。大きな精度向上が見られた要因は、どのデータセットにおいても C クラスのうち 6 割程度が B2 クラスに誤分類され、C クラスの分類のみ他クラスと比べて極端に難しかったからである。C クラスが B2 クラスに誤分類される原因として、構文構造に関する特徴量の分布を検証したところ、C と B2 のエッセイで非常に近い分布となったことから文章構造の難易度自体は B2 クラスで頭打ちであり、C クラスは文章として複雑性が増すというよりむしろ簡潔になるなどより Readability に優れ、洗練されている文章であるためと考えられる。

4.3 Gini 係数による各素性の重要度ランク

ランダムフォレストでは Gini 係数を用いて素性の重要度をランク付けする。表 3 は 1to1 と 10to1 の本実験における各素性の分類に対する寄与度のランキングを示したものである。

表 3: 分類に対する素性の寄与度のランキング（**太字**は共通して有効である主な素性）

rank	1to1	10to1
1	**avg_difficulty**	**avg_difficulty**
2	**A1_per**	**A1_per**
3	**MLS**	A2_per
4	NP_per	avg_CP
5	avg_word_length	B2_per
6	avg_D_score	avg_TRACE
7	TTR	**MLS**
8	avg_VP	**avg_max_depth**
9	**avg_max_depth**	NP_per
10	word_length_1_3	avg_content

1to1 と 10to1 における寄与度のランキングを見ると、いずれも上位は単語の平均難易度を表す avg_difficulty や A1 レベルの単語の割合を示す A1_per など単語の難易度に関する指標である。また、MLS や avg_max_depth といった指標も有効であると言える。

1to1 では avg_word_length や word_length_1_3 など個々の単語に対する指標や avg_D_score が上位に現れているのに対して、10to1 では補文素（CP）といった文法カテゴリに関する指標が有効であることが分かる。

4.4 2 値分類による各クラスの言語特徴の分析

表 4 に隣り合うレベル間の 2 値分類したときの素性の重要度の各ランキングと分類精度を示す。A1 と A2 を分類した場合の各素性の分類に対する寄与度のランキングを見ると、上位には A1_per や A2_per など難易度に関する指標が現れている。特に B1 クラスの単語難易度の割合を表す B1_per が全体の分類と比べて特徴的である。また、文法カテゴリとしては補文素（CP_per, avg_CP）や関係詞節（REL_per, avg_REL）が有効である。

次に、A2 と B1 の分類では、分類に有効な素性が全体の分類のときに比べて大きく異なる結果となった。ランキングのトップが内容語の平均句数（avg_content）となっており、その他の文法カテゴリとしては動詞句（VP）や副詞句（ADVP）も特徴的である。また、avg_max_depth や D_score、 L_score が上位に現れていること、単語の難易度に関する指標が下位に沈んでいることから A2 と B1 では単語の難易度よりむしろ構文構造が大きく変化していると考えられる。

B1 と B2 の分類の場合、特徴的な素性が比較的高難易度である B2 クラスの単語難易度の割合と関係詞節であるという結果になっている。

5 まとめ

本稿では学習者が書いた英文のレベル推定やレベル毎の言語特徴を明らかにすることを目的として、CEFR 準拠教科書データを用いて英文の分類実験を行った。全体の分類に大きく寄与する特徴量は単語の難易度に関する諸指標や平均文長などであった。また、隣り合うレベル間で 2 値分類を行うことで各レベルの言語特徴を分析した。A1 と A2 においては単語の難易度、文法カテゴリとしては補文素や関係詞節などが識別的である。A2 と B1 では、構文構造に大きな違いが見られ、B1 と B2 では B2 クラスの難易度の単語の割合などが特徴的であった。

表 4: 隣接レベル間の 2 値分類における素性の寄与度ランキングと分類精度（**太字**は全体の分類と比べ特徴的な素性）

rank	A1A2 間	A2B1 間	B1B2 間
1	A1_per	**avg_content**	**B2_per**
2	A2_per	avg_VP	avg_difficulty
3	avg_difficulty	avg_ADVP	A1_per
4	**B1_per**	**avg_max_depth**	**avg_REL**
5	**avg_CP**	avg_CP	A2_per
6	**CP_per**	MLS	NP_per
7	avg_REL	**avg_D_score**	B1_per
8	MLS	**avg_L_score**	**REL_per**
9	avg_word_length	CP_per	avg_TRACE
10	REL_per	NP_per	COOD_per
分類精度	73.1%	79.8%	79.0%

謝辞

本研究は、JSPS 科学研究費補助金基盤研究 A「学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究」（研究課題番号：24242017）の助成を受けたものである。

参考文献

[1] C. Bishop, “Pattern recognition and machine learning,” Springer, 2006.

[2] L. Breiman, “Random Forests,” Machine Learning, 45(1), pp.5-32, 2001.

[3] J.P. Kincaid, R. Fishburne, R. Rogers and B. Chissom, “Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. ,” Research Branch Report, pp.8-75, 1975.

[4] 小林雄一郎, “ランダムフォレストを用いた英語習熟度の自動推定,” 言語処理学会第 18 回年次大会, pp.979-982, 2012.

[5] 小林雄一郎, 金丸敏幸, “パターン認識を用いた課題英作文の自動評価の試み,” 信学技報, pp.37-42, 2012.