英語学習者の英作文からのCEFRレベル別基準特性の特定

林 正頼[1]　石井康毅[2]　高村大也[3]　奥村 学[3]　投野 由紀夫[4]

[1] 東京工業大学 総合理工学研究科 [2] 成城大学 社会イノベーション学部
[3] 東京工業大学 精密工学研究所 [4] 東京外国語大学 大学院総合国際学研究院

hayashi@lr.pi.titech.ac.jp, ishii@seijo.ac.jp,
{takamura, oku}@pi.titech.ac.jp, y.tono@tufs.ac.jp

1　はじめに

欧州の外国語学習者が参照する，コミュニケーション能力別のレベルを示す到達度指標として，ヨーロッパ言語共通参照枠 (Common European Framework of Reference for Languages： CEFR)[1] があり，欧州では広く普及している．近年，その日本語版であるCEFR-J[1] が投野らによって公開された．

現在公開されている CEFR-J 中の記述の一例を挙げると，“書くこと”の初-中級レベルに相当する英語能力として，“自分の経験について，辞書を用いて，短い文章を書くことが出来る．”とある．このように，到達度指標が ‘can do’ (何ができるか) の形で明示はされているが，一方で，CEFR-J が準拠する CEFR の ‘can do’ は言語中立の能力記述文であるため，定性的な英語能力についての言及がされていない．そのため，全体的に具体性に欠け，抽象的な記述が多く見られることから，各国語で各レベルの ‘can do’ に対応する言語特徴 (基準特性) を整備する作業が行われている．

つまり，この CEFR-J に基づいた英語教育を進めていくためには，具体性に欠け，抽象的な記述が多く見られる現在の状態から，具体的に，どのような語彙や表現が使えるべきか (語彙や表現のリスト)，どのような構文が使えるべきか (構文のリスト) といった情報を体系的に整備していくことが必要不可欠である．そして実際，このような取り組みが投野らのグループにより現在進められている[2]．

それを踏まえ，本稿では，英語学習者の英作文から CEFR レベル別基準特性を自動的に特定する．その際，作文中に生じる誤りに特に着目する．英作文を書く英語能力を判別する基準特性としては，どのような語彙が使えるか，どのような構造を持つ文が書けるかといった，文章中の正しく使われているケース (正用例) 以外に，文章中に存在する文法的な誤り (誤用例) が考えられる．英語学習者によって書かれた英作文には，誤りが含まれる可能性が高く，どのような誤りをしているかという情報は，英語学習者の英作文のレベルを判別する上で重要かつ不可欠な手がかりであると言える．しかし，これまで，英語教材からの CEFR レベル別基準特性の特定においては，教材の文章中に「正用例」しか出現しないので，正用例のみを用いるのが原則であった．

2 節で説明するように，日本人英語学習者によって書かれた英作文から，石井ら [2] によって整備されてきた文法項目に関する素性，および作文中に含まれる誤りに関する素性を抽出した上で，それらの素性を用い，CEFR レベルが付与された英作文集合を元に，機械学習技術を適用することで，英作文のレベル判別を行う分類器を自動的に学習する．そして，学習された分類器から，英作文のレベル判別に特に有効と考えられる素性を列挙することで，CEFR レベル別基準特性を自動的に特定する．

2　JEFLL コーパス

JEFLL(Japanese English as a Foreign Language Learner) コーパス[3] は，中学・高校の日本人英語学習者，約 1 万人分の自由英作文データをコーパス化したものである．このコーパスは，中学・高校の日本人英語学習者によって，ある論題について書かれた英作文 (原文) と，原文を英語教育者が文法的に訂正したもの (訂正文) のペアから構成される．原文と訂正文は，一文の原文に対して一文の訂正文が存在し，一対多の関係となるペアは存在しない．ただし，単語単位での対応付けはされていない．

コーパスは XML 形式で保存されており，各単語の原形，品詞情報などに加えて，使用されている文法項

[1] http://www.tufs.ac.jp/ts/personal/tonolab/cefr-j/index.html
[2] https://kaken.nii.ac.jp/d/p/24242017.ja.html
[3] http://jefll.corpuscobo.net/index.htm

目を示すタグが単語 (またはフレーズ) に付与されている．文法項目リストは，石井らによって整備されたもので [2]，今回は，2015 年 10 月 8 日時点の 295 種類の文法項目を用いている．表 1 はその一部である．さらに，各英作文に対しては，英語教育者により CEFR-J に基づき，A1，A2，B1，B2 の 4 段階のレベル分類も行われている．

表 1: 文法項目の一例

ID	文法項目
1	人称代名詞主格 (I am)
3	人称代名詞主格 (he/she is)
11	指示形容詞 (this/that+名詞)
137	助動詞類 (should)
253	wish+仮定法過去

3 提案手法

本稿では，英語学習者の英作文から CEFR レベル別基準特性を自動的に特定する際，JEFLL コーパスから基準特性となりうる素性集合をまず抽出する．得られた素性集合から，隣接する 2 つのレベル間での 2 値分類を行う分類器を用いることで学習し，最後に，分類に有効であった素性を列挙し，レベル判別に有効であると考えられる素性集合を基準特性として特定する．

3.1 素性

提案する素性集合は，次の 3 種類の素性に分けることができる．なお，素性値には，文章情報を除き，頻度情報を用いている．

3.1.1 基本素性

ベースラインとなる素性として，品詞情報および文章情報を使用した．品詞情報を表す素性としては，英作文中に出現する品詞タグの 1-gram, 2-gram, 3-gram の頻度を用いる．しかし，原文には，スペルミスや，学習者が書こうとした英単語の日本語が含まれていたりして，そのままでは品詞タグ付けが正しく行えない．そこで，今回は訂正文に対して品詞タグ付けを用い，品詞タグの 1-gram, 2-gram, 3-gram を作成し，これらを品詞情報を表す素性として用いた．

なお，単語の表層形は，JEFLL コーパス内の論題に偏りが見られ，また，単語表層形の情報は本稿の目的とは合致しないため，使用していない．また，文章情報としては，その文章の総単語数，総文数，一文あたりの平均単語数をそれぞれ用いている．

3.1.2 誤りに関する素性

JEFLL コーパスは原文と訂正文のペアで構成されるが，単語同士の対応はとれておらず，原文-訂正文間でのアライメントを行う必要がある．そこで，今回は望月ら [3] による編集距離を用いた英文自動エラータグ付与ツールで，動的計画法に基づく単語間のアライメントを行った．アライメントの出力結果としては，以下の 4 通りが考えられる．

- 原文と訂正文の単語が同一の単語である
- 単語が原文と訂正文で異なる（添削者が単語を書き換えた） (置換誤り)
- 単語が原文には存在しないが，訂正文には存在する（添削者が単語を追加した） (脱落誤り)
- 単語が原文には存在するが，訂正文には存在しない（添削者が単語を削除した） (余剰誤り)

今回は，置換，脱落，余剰の誤りがあった単語の頻度を，その品詞ごとに集計し，素性として用いた．たとえば，添削者が形容詞を追加しているような場合，「形容詞 (脱落)」という素性が用いられる．ただし，品詞が，前置詞や限定詞といった，機能語とみなせる品詞の場合は，単語の表層形をそのまま素性として用いた．すなわち，たとえば，添削者が前置詞 ‘in’ を ‘at’ に置き換えているような場合,「in(置換)」という素性が用いられる．

3.1.3 文法項目に関する素性

2 節で述べたように，JEFLL コーパスには，どのような文法項目が使われているかのタグが付与されている．したがって，英作文中にどのような文法項目が何回出現しているかを素性として用いることは容易である．しかし，ある文法項目が原文内で使われているが，その箇所が誤り箇所と重なっている場合，それは正しく使用されているとは言えないことになる．そこで，前節で述べた，誤りに関する素性を作成する際に用いた，原文-訂正文間の単語のアライメント結果を利用し，原文中で文法項目が出現している単語に関して，

- 置換，脱落，余剰の誤りがある単語が含まれる (誤用例)
- 上の誤りがある単語を含まない (正用例)

の 2 つに分け，295 種類の文法項目の出現それぞれに対し，正用例，誤用例の頻度を素性として利用した．

4 実験

今回は，一つの英作文を一事例として実験データを作成した．JEFLL コーパスのレベル別の事例数は表 4 の通りである．なお，使用した JEFLL コーパスには B2 レベルと分類された英作文も存在するが，事例数が少なく，今回の実験設定では正しく学習ができなかったため，A1，A2，B1 の 3 つのレベルの事例のみを使用した．よって，隣接するレベル間での 2 値分類であることから，A1-A2 レベル間，A2-B1 レベル間での実験を行った．

表 2: JEFLL コーパスの事例数

レベル	事例数
A1	3360
A2	4900
B1	1520
B2	45
合計	9825

品詞 n-gram を作成するために，Stanford POS Tagger[4] を使用した．また，機械学習手法としては，Support Vector Machines(SVM) を用いた．なお，SVM の実装としては，LIBSVM[5] を用い，線形カーネルを採用した．パラメータ C と γ は，グリッドサーチによって決定した．素性の組み合わせは以下の 4 通りで，5 分割交差検定を用いた実験を行った．

(1) 基本素性のみ

(2) 誤りに関する素性のみ

(3) 文法項目に関する素性のみ

(4) すべて

その結果，分類正解率は表 3 のようになり，すべての素性を組み合わせた (4) が最良の結果を得ることを確認した．

表 3: 分類正解率

素性	Acc(A1-A2 間)	Acc(A2-B1 間)
(1) 基本素性	.830	.969
(2) 誤り素性	.667	.832
(3) 文法項目素性	.792	.935
(4) すべて	**.840**	**.971**

[4]http://nlp.stanford.edu/software/tagger.shtml
[5]http://www.csie.ntu.edu.tw/~cjlin/libsvm/

5 基準特性となる素性の列挙

素性を，誤りに関する素性のみ，文法項目に関する素性のみにそれぞれ限定し，分類器を学習させた際，重みの絶対値の大きい素性を上位 10 個ずつ列挙した結果が表 4 から表 7 となる．なお，表 6，7 において “(誤)”が付いている素性は，文法項目の誤用例を表している．このように，分類に貢献していると考えられる素性集合を，誤りに関する素性，文法項目に関する素性から抽出することにより，CEFR レベル別の基準特性となる素性集合を特定できるものと考えている．

6 おわりに

本稿では，CEFR-J に基づく，隣接する 2 つのレベル間の英作文データに対し教師あり学習による 2 値分類を行い，誤りに着目した素性が有効であることを確認した．また，学習された分類器から，CEFR-J のレベル判別において有効である，重みの絶対値が大きい素性を，CEFR レベル別基準特性候補として列挙した．今回の JEFFL コーパスでは訂正文が存在し誤り箇所が特定できていたが，通常は訂正文が存在するとは限らない．今後の課題として，誤り個所が特定されていない場合にその個所を特定する高精度な手法の構築が考えられる．

謝辞

本研究は科学研究費基盤研究 (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」(課題番号：24242017，代表：投野由紀夫) の助成を実施した．

参考文献

[1]Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schne, Barbora Tindlov, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the 9th International Conference on Language Resources and Evaluation.* pp. 1281–1288.
[2] 投野由紀夫，石井康毅. 2015. 英語 CEFR レベルを規定する基準特性としての文法項目の抽出とその評価. 言語処理学会第 21 回年次大会, pp. 884–887.
[3] 投野由紀夫，望月源. 2012. 編集距離を用いた英文自動エラータグ付与ツールの開発と評価.『コーパスに基づく言語学教育研究報告』 No.9. pp. 71–92.

表 4: A1-A2 間で有効な誤りに関する素性

A1 側		A2 側	
誤りタイプ (A1 側)	重み	誤りタイプ (A2 側)	重み
固有名詞単数 (余剰)	245.0	the (脱落)	-75.4
単数名詞 (脱落)	215.4	複数名詞 (置換)	-51.2
動詞 (過去) (脱落)	137.8	I (余剰)	-38.8
副詞 (脱落)	109.7	動詞 (原形) (余剰)	-37.4
動詞 (原形) (脱落)	69.0	副詞 (置換)	-33.8
形容詞 (脱落)	63.8	形容詞 (余剰)	-27.0
動詞 (過去) (置換)	44.3	the (置換)	-25.1
複数名師　(脱落)	40.3	形容詞 (置換)	-23.5
動詞 (分詞) (脱落)	35.0	動詞 (過去分詞) (置換)	-23.3
動詞 (現在) (脱落)	32.7	of (置換)	-22.9

表 5: A2-B1 間で有効な誤りに関する素性

A2 側		B1 側	
誤りタイプ	重み	誤りタイプ	重み
and (置換)	305.7	動詞 (過去分詞)(余剰)	-291.9
out (脱落)	255.2	動詞 (過去) (置換)	-231.6
so (置換)	246.2	形容詞 (置換)	-217.7
my (脱落)	192.7	副詞 (比較) (置換)	-213.1
if (脱落)	188.3	副詞 (比較) (余剰)	-205.1
could (置換)	186.1	動詞 (過去)(余剰)	-203.0
it (置換)	185.8	動詞 (現在) (置換)	-193.1
them (置換)	185.7	動詞 (現在三単現) (置換)	-190.7
at (脱落)	163.7	形容詞 (比較) (余剰)	-186.3
of (余剰)	150.5	動詞 (原形) (置換)	-184.4

表 6: A1-A2 間で有効な文法項目に関する素性

A1 側		A2 側	
文法項目	重み	文法項目	重み
時制・相 (過去)	158.3	that 節 (目的語)	-23.9
前置詞	95.1	人称代名詞目的格	-23.1
等位接続詞	75.7	to 不定詞	-19.0
従属節	72.4	時制・相 (過去完了)	-18.2
時制・相 (現在)(do)	67.9	助動詞類 (would)	-18.1
等位接続詞 (誤)	57.8	複合関係代名詞 (what)	-17.3
関係代名詞 (目的格) の省略 (誤)	51.4	時制・相 (過去完了)(誤)	-16.8
人称代名詞所有格	41.0	wh-ever	-16.8
関係代名詞 (目的格) の省略	39.2	人称代名詞目的格 (誤)	-16.6
時制・相 (過去)(誤)	31.8	関係代名詞 (主格)	-16.4

表 7: A2-B1 間で有効な文法項目に関する素性

A2 側		B1 側	
文法項目	重み	文法項目	重み
助動詞類 (have to)	434.4	間接話法 (say・explain・report)(誤)	-512.0
助動詞類 (be going to)	338.0	助動詞類 (should)	-497.1
句動詞 (動詞+名詞句・代名詞+パーティクル)(誤)	256.0	比較級 (誤)	-495.0
助動詞類 (need)(誤)	254.5	副詞節 (when)(誤)	-493.3
助動詞類 (need)	250.4	時制・相 (現在)(do)	-485.5
助動詞類 (be going to)(誤)	245.3	助動詞類 (will)(誤)	-479.6
疑似関係代名詞 (as)	241.1	間接話法 (ask)	-462.0
ask・tell+目的語+to+do	237.2	等位接続詞 (誤)	-451.9
時制・相 (過去完了)(誤)	228.3	存在の There(There+be)(誤)	450.7
時制・相 (過去進行)	224.1	関係代名詞 (主格)(誤)	-436.2