# 英語学習者の作文における誤りタグの自動付与

林 正頼[1]　高村大也[2]　奥村 学[2]　投野 由紀夫[3]
[1] 東京工業大学 総合理工学研究科 [2] 東京工業大学 精密工学研究所
[3] 東京外国語大学 大学院総合国際学研究院
hayashi@lr.pi.titech.ac.jp, {takamura, oku}@pi.titech.ac.jp
y.tono@tufs.ac.jp

## 1 はじめに

日本の英語教育の抜本的な改革のために,「ことばを使って何をするか」という言語機能主体の到達度指標を新たに設定する動きが広がっている[1]．また，ヨーロッパ言語共通参照枠 (Common European Framework of Reference for Languages : CEFR) [1] という，欧州の外国語学習者のための，コミュニケーション能力別のレベルを示す到達度指標が世界的に広く普及している．その中で，2012 年に，CEFR の日本語版である CEFR-J[2] が投野らにより公開された．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．”のようである．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を整備する作業（参照レベル記述と呼ぶ）が行われている．

CEFR-J に基づいた英語教育を推進していくためには，CEFR-J によって定められた各レベルに対応して，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[3]．英語の文章を書く際に，英語能力を判別する要素としては，上に挙げたような，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「正用例」（正しく使われているケース）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．本稿ではこの英語学習者の書いた文章中の文法的な誤りに着目する．

学習者の作文中の誤りが自動的に検出できれば，その誤りの種類と，種類ごとの頻度の情報が得られるので，これを元に，その学習者の英語能力を判別できる可能性がある．この考えを元に，我々は現在，1. 統計的機械翻訳を用いたアプローチにより，英語学習者の作文中の誤りを自動的に検出し，2. 1. で得られた誤りの種類と頻度の情報を素性に加え，統計的機械学習手法を用いて，学習者の英語能力のレベルを自動的に判別する研究を行っている．このアプローチを用いる際，1. で誤りのタグが付与された訓練データ（誤っている文とそれを正しく添削された文のペア）が大量に必要となるが，現状では誤りのタグの種類が統一されておらず，異なる誤りのタグセットが付与されたコーパスや，誤りのタグが付与されていないコーパスなど，多種多様なデータが存在しており，限られた量のデータセットのみしか利用できていない．

そこで，本稿では，誤りのタグが付与されていないが，誤りが含まれており，かつその誤りが正しく添削されている英作文のデータが大量に存在することから，その誤り個所に誤りのタグを自動的に付与するシステムの構築を目指す．最終的には 1. における訓練データを大量に入手することを本稿の目的とする．

## 2 提案手法

本稿では，まず，誤りタグが付与された訓練データから多値分類器を作成し，同様のタグが付与されたテストデータで分類精度を確認する．そして，誤りタグが付与されていないデータセットから評価用データを作成し，学習させた多値分類器で分類を行い，評価データが正しく分類されているかを評価する．

[1] http://www.mext.go.jp/b_menu/houdou/23/07/1308888.htm
[2] http://www.tufs.ac.jp/ts/personal/tonolab/cefr-j/index.html
[3] https://kaken.nii.ac.jp/d/p/24242017.ja.html

表 1: NUCLE に含まれる誤りタグ一覧

| タグ | 詳細 | タグ | 詳細 | タグ | 詳細 |
|---|---|---|---|---|---|
| Vt | 動詞の時制 | Prep | 前置詞 | Woinc | 語順 |
| Vm | 助動詞 | Wci | イディオム | WOadv | 形容詞/副詞の位置 |
| V0 | 動詞の不足 | Wa | 頭字語 | Trans | 単語/句の連結 |
| Vform | 動詞の形態 | Wform | 語形 | Mec | スペル/句読点/大文字 |
| SVA | 主語-動詞の整合性 | Wtone | 砕けた表現 | Rloc- | 冗長な表現 |
| ArtOrDet | 冠詞/決定詞 | Srun | 一文中に複数文存在 | Cit | 引用 |
| NN | 名詞の単数/複数 | Smod | 懸垂修飾語 | Others | その他 |
| Npos | 所有代名詞 | Spar | 単語の並列 | Um | 意味が不明 |
| Pform | 代名詞の形態 | Sfrag | 不完全な文 | | |
| Pref | 代名詞の使用 | Ssub | 従属節 | | |

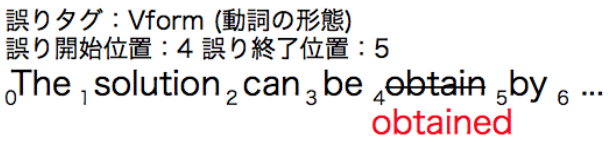


図 1: NUCLE で付与されている情報

## 2.1 NUS Corpus of Learner English

NUS Corpus of Learner English (NUCLE) [2] は，シンガポール国立大学の学生によって書かれた，さまざまなトピックから成る約 1400 の英文エッセイから構成され，その英文の中に含まれている誤りの箇所にラベルが付与されたコーパスである．図 1 にラベルの概要を示す．

ラベルには，誤りが開始する位置と終了する位置，訂正後の単語集合，誤りタグが付与されている．誤りタグは表 1 で示すとおり，28 種類から構成される．本稿では誤りがある箇所を誤り箇所，訂正後の箇所を訂正箇所と定義する．また，この NUCLE に付与されている誤りタグを本稿の基準とした．なお，NUCLE は近年の CoNLL の shared task [3] [4] である誤り訂正のためのデータセットとして用いられており，提案手法によって，NUCLE の誤りタグに準じた訓練データを増加させることが出来るなら，誤り訂正のタスクにも有用な手法と期待できる．

## 2.2 実験設定

上で述べたように，誤りタグが付与された訓練データから，誤りタグ総数 28 種類のラベルを分類する多値分類器を構築する必要がある．多値分類器として，Support Vector Machine (SVM) を用いた．なお，SVM の実装として LIBSVM[4] を使用した．カーネルは線形カーネルを用い，正則化パラメータ $C$ は，0.1 から 1 までは 0.1 刻み，1 から 10 の 1 刻みで変化させ，最も高い正解率が得られる値 $C = 0.2$ を採用した．マルチクラスの分類器を作成する際には，one-versus-rest 法を用いた．

訓練データは，NUCLE 内に存在する誤りが含まれる文のみから作成した．なお，NUCLE のデータには 1 文につき複数の誤りが含まれている場合も存在する．現在，英語学習者の作文中の誤りを自動的に検出する手法として，誤りの種類別に統計的機械翻訳を用いたモジュールを作成することを検討している．その際の用いる訓練データとしては，1 つの文に対してタグが 1 つのみ付与されている訓練データの方が今後の方針と合致している．このため，1 訓練事例につき 1 つの誤りのみを含むように，誤りの数だけ異なる訓練事例として処理を行った．その結果，訓練事例は 44314 件となった．

## 2.3 素性

今回の実験設定において，誤り箇所に関する素性と，誤り箇所の周辺に関する素性の 2 つが有効な素性として考えられる．以下に詳細を述べる．なお，スペルミスや，文頭が小文字であるなど，従来の研究では前処理で除かれる可能性がある語も，今回は誤りとして分類される対象であることから，前処理などは行わなかった．

### 2.3.1 誤り箇所に関する素性

誤り箇所と訂正箇所の単語表層形のペアと，品詞タグのペアを使用した．冠詞が不足しているなど，誤り箇所に単語が存在しない場合は擬似的に {NONE} を誤り箇所の単語とした．また，誤り箇所-訂正箇所間の単語数の差とともに，誤り箇所-訂正箇所間の単語にレーベンシュタイン距離を用いた．図 1 の例では，

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

表 2: 分類結果

| タグ | 精度 | 再現率 | F1 | タグ | 精度 | 再現率 | F1 | タグ | 精度 | 再現率 | F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Vt | 0.34 | 0.58 | 0.43 | Prep | 0.74 | 0.70 | 0.72 | Woinc | 0.44 | 0.22 | 0.29 |
| Vm | 1.00 | 0.37 | 0.54 | Wci | 0.51 | 0.57 | 0.54 | WOadv | 0.00 | 0.00 | 0.00 |
| V0 | 0.42 | 0.29 | 0.34 | Wa | 0.00 | 0.00 | 0.00 | Trans | 0.54 | 0.24 | 0.33 |
| Vform | 0.19 | 0.14 | 0.16 | Wform | 0.18 | 0.09 | 0.12 | Mec | 0.46 | 0.43 | 0.44 |
| SVA | 0.66 | 0.60 | 0.63 | Wtone | 0.40 | 0.15 | 0.22 | Rloc- | 0.26 | 0.74 | 0.39 |
| ArtOrDet | 0.85 | 0.88 | 0.86 | Srun | 0.48 | 0.24 | 0.32 | Cit | 0.00 | 0.00 | 0.00 |
| NN | 0.73 | 0.85 | 0.79 | Smod | 0.00 | 0.00 | 0.00 | Others | 0.29 | 0.71 | 0.42 |
| Npos | 0.00 | 0.00 | 0.00 | Spar | 0.67 | 0.06 | 0.11 | Um | 0.55 | 0.31 | 0.40 |
| Pform | 0.00 | 0.00 | 0.00 | Sfrag | 0.00 | 0.00 | 0.00 | | | | |
| Pref | 0.30 | 0.25 | 0.27 | Ssub | 1.00 | 0.01 | 0.03 | | | | |

誤り箇所と訂正箇所のペアが (obtain, obtained) である．この場合，“ e ”と“ d ”の挿入が必要となるため，レーベンシュタイン距離は 2 となる．その値を素性として使用した．なお，(was, have been) など，誤り箇所と訂正箇所の単語数が異なる場合でも同様の処理を行った．さらに，誤りの箇所が文中のどこに生じているのか，誤り箇所の開始位置の情報も追加した．

#### 2.3.2 誤り箇所の周辺に関する素性

誤り箇所の前後 3 単語の表層形と品詞タグのユニグラム，バイグラム，トライグラムをそれぞれ使用した．誤り箇所が文頭付近，または文末付近に存在する場合，バイグラムやトライグラムが抽出できない可能性がある．そこで，表層形，品詞タグのどちらの場合も，まず文頭と文末を示す記号<s>と</s>を挿入し，それでもなお素性が抽出できない場合は，ダミー語（<dummy>）を挿入することで対処した．

### 2.4 実験結果

CoNLL 2014 の誤り訂正タスク [3] にも用いられている，NUCLE と同じタグ付けがされているデータセットから，訓練事例の際と同様に，1 訓練事例につき 1 つの誤りが含まれる文のみから構成されるテストデータを作成し，評価実験を行った．テスト事例は合計 3462 件となった．各エラータグの分類結果を表 2 に示す．なお，全体の正解率は 58.0%（3462 事例中 2008 事例正解）であった．訓練データが多い，冠詞または決定詞誤り（ArtOrDet：6612 事例）や，名詞の単複誤り（NN：3743 事例）はスコアが高くなる傾向にあり，一方で懸垂修飾語を含む誤り（Smod：47 事例），頭字語に関する誤り（Wa：48 事例）など，訓練データが少ない誤りタグの場合はスコアが低くなる結果となった．

## 3 Lang-8 データへの適用

### 3.1 Lang-8 Learner Corpora

前節で構築した多値分類器が，Lang-8[5] データに対して有効であるかの評価を行う．Lang-8 とは，投稿者が学習したい言語で文章（投稿文）を投稿すると，投稿文が母語である添削者が，投稿文を添削して添削文を返信する，言語学習者向けの相互添削型 SNS である．それらの投稿文と添削文を収集したデータセットが Lang-8 Learner Corpora[6] である．本稿ではこのコーパスを評価セットとして用いる．このデータセットには，記事の id，文の id，現在学習中の言語，母国語の情報と，投稿文と添削文のペアが含まれており，誤りのラベルは付与されていない．本稿では，Lang-8 Learner Corpora から，投稿文と添削文のペアを無作為に 100 ペア選び出し，その中に含まれる誤りから評価データを作成した．ただし，Lang-8 の添削方法は，NUCLE のデータセットの誤りタグに準拠しておらず，そのまま評価データとして用いることができない．また，添削文の訂正方法が，添削者の意向に依存するため，目的のデータセットを自動的に作成することが現状では困難である．

そこで，無作為に選び出された 100 ペアから，NUCLE の誤りタグに対応する箇所のみを人手で選び出すことで，評価データを作成した．前節の訓練データ，テストデータと同様に，1 事例につき 1 つの誤りが含まれるようにデータを作成した．その結果，投稿文と添削文の 100 ペア中に 163 の誤りが含まれていることから，評価データは 163 件となった．

[5] http://lang-8.com/
[6] http://cl.naist.jp/nldata/lang-8/

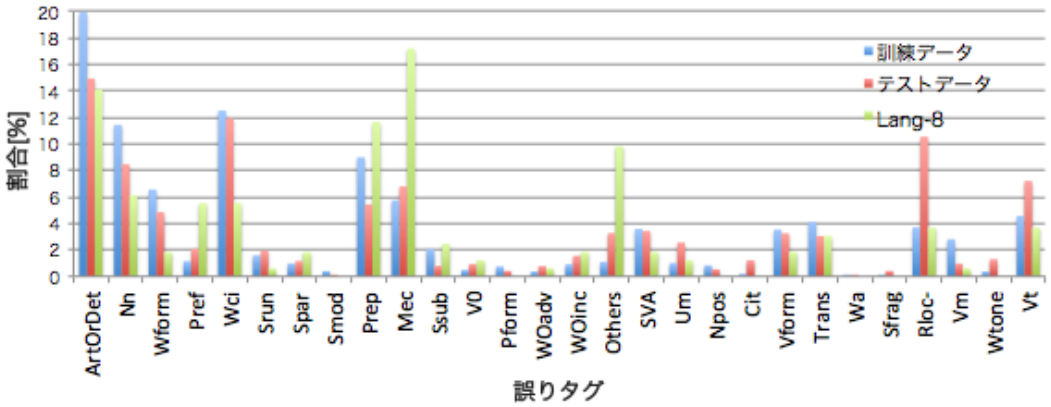


図 2: 各データのラベルの分布

### 3.2 Lang-8 データを用いた多値分類結果

全体の正解率は 40.5%（163 事例中 66 事例正解）となった．前節で用いたテストデータより正解率が大幅に低下することが確認された．原因として，訓練データと評価データ間の誤りタグの分布が異なることが考えられる．今回の実験で用いた訓練データ，テストデータ，評価データの各誤りタグの分布は図 2 のようになる．

この結果から，Lang-8 のデータの割合は，訓練データの割合に比べ，スペル，句読点，大文字などの誤り（Mec）は 3 倍程度，誤りタグがいずれにも該当しないその他の誤り（Others）が 10 倍ほど高いことがわかる．前節で述べたように，訓練事例の少ないタグは，正解率が低下する傾向があるため，この影響を受けたと考えられる．

また，これら結果から，今後検討すべき事項が 2 点考えられる．まず，訓練データの誤りタグの分布の偏りを無くすための手法，もしくはデータセットの偏りを踏まえた手法について検討する必要がある．そのデータセットが，どのような英語学習者によって書かれた英作文であるかを考慮しなければならない．例えば，母国語が異なると，母語干渉が起こり，英語の誤りの傾向が異なることが示唆されている．つまり，母国語を考慮した分類器を構築することによってより良い分類器が構築できると考えられる．

2 つ目として，今回基準とした誤りタグは適当であるかという点である．Lang-8 Learner Corpora からデータセットを作成するにあたって，誤りタグがいずれにも該当しないその他の誤り（Others）が散見された．これらの誤りの中で，更に細分化できる誤りが含まれている可能性もある．このことから，より適切な粒度のタグの基準について，議論をする必要が将来的にあると考えられる．

## 4 おわりに

本稿では，英語学習者の作文中に含まれる誤りが，どのような種類の誤りであるかを分類する多値分類器を作成し，その分類器が実際に誤りの分類に有効であるかの評価を行った．実験の結果，今回の提案手法は，訓練事例が多い誤りラベルの場合には特に有効であることが確認された．この手法によって，新たなデータセットの作成の可能性が示唆され，本来の目標である，CEFR-J に対応した英語能力レベル判別の具体的な特徴の抽出や，統計的機械学習手法を用いた誤り訂正タスクに寄与することが期待できる．

## 5 謝辞

本稿は JSPS 挑戦的萌芽研究 25540097 の助成を受け実施した．

## 参考文献

[1] Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, andrea Abel, Karin Schne, Barbora tindlov, and Chiara Vettori. The merlin corpus: Learner language and the cefr. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[2] Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 22–31, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[3] David S. Hdez. and Hiram Calvo. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, chapter CoNLL 2014 Shared Task: Grammatical Error Correction with a Syntactic N-gram Language Model from a Big Corpora, pp. 53–59. Association for Computational Linguistics, 2014.

[4] Tou Hwee Ng, Mei Siew Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, chapter The CoNLL-2013 Shared Task on Grammatical Error Correction, pp. 1–12. Association for Computational Linguistics, 2013.