文法誤りにおける一般誤りの分離可能性と解説文生成への応用

永田亮 [1]　木村学 [2]
[1] 甲南大学知能情報学部 [2] GRAS グループ株式会社
nagata-nlp2023 @ ml.hyogo-u.ac.jp.　manabu.kimura@gras-group.co.jp

概要

本稿では，ある種の誤り（一般誤りと呼ぶ）とその他の誤りでは，誤り検出器における検出規則の獲得過程が異なり，その性質により一般誤りのみを分離できるという新たな知見を報告する．また，その知見を利用して一般誤りの詳細なサブタイプを発見することについても述べる．更に，発見したサブタイプを利用して，解説文生成を分類問題として解くことを提案する．このアプローチには，訓練データ作成コストと生成結果の信頼性という面で有利な点があることを示す．

1　はじめに

本稿では，一般誤りの分離可能性という仮説を導入し，その仮説を文法誤り解説に応用する手法を提案する．仮説の概要は，一般誤りとその他の誤りでは，誤り検出器における検出規則の獲得過程が異なり，その性質を利用することで一般誤りのみを分離できるというものである．ここで，一般誤りとは，特定の内容語に依存しない規則で同定される文法誤りと定義する（便宜的に，その他の誤りを単語固有誤りと呼ぶ）．例えば，**In this café serves good coffee.* は，規則「主語は前置詞を伴わない」で誤りと同定できるが，この規則はどのような主語や前置詞の組み合わせにも適用できる．一方，単語固有誤り（例：**They protested on the situation.*）では，特定単語（この例では *protest* と *on*）により誤りと同定される．本稿では，上述の仮説を利用して，前置詞の用法に関する文法誤りについて，与えられた学習者コーパスから一般誤りのみを分離できることを示す（§3）．

本仮説は，BERT ベースの誤り検出器の性能に関する報告 [1] に着想を得ている．同報告では，この検出器の性能曲線が図 1 の実線のようになることを示している．図 1 から，数百文の訓練データで性能が劇的に改善し，それ以降性能向上は緩やかになることがわかる．この理由を，我々は次のように予想する．BERT ベースの検出器ではごく少量の訓練データから一般誤りの検出規則が獲得され，一方で，単語固有誤りについてはより多くの訓練データが必要となる．一般誤りの同定規則は汎用的であるため，異なる単語からなる複数の訓練事例から獲得可能である．特に，BERT のように文法知識を有するモデル [1, 2, 3] では，一般誤りの検出規則の獲得が容易になる．そのため，一般誤りに関する性能曲線は図 1 の赤い曲線（破線）のような形になる．一方，単語固有誤りについては，特定の単語を含んだ訓練事例が必要となり，相対的に訓練事例が少ない．また，汎化がより難しい．結果，図 1 のように直線的に性能が改善する．結果として観測される性能曲線は，二種類の破線の重ね合わせとなる．

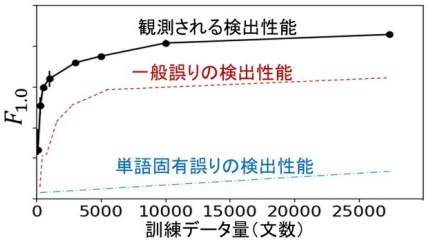

図 1　誤り検出性能曲線と仮説的な説明．

もしこの仮説が完全に成り立つのであれば，BERT ベースの検出器を少量のデータで訓練し，誤り検出を行えば一般誤りのみが得られるはずである．本稿では，このアイデアを拡張し一般誤りの事例を得る手法を提案する．また，獲得された一般誤りにクラスタリングを適用することで，詳細なサブタイプを効率よく発見できることも示す（§4）．更に，得られたサブタイプにより解説文生成 [4] を分類問題として解くことが可能になることも示す（§5）．これは，解説文生成における大きな二つの課題（1）訓練データが大量に必要となる，（2）フェイク解説文が生成されてしまう，の解決に繋がる．

2　関連研究

本稿の仮説は文献 [1] に着想を得ている．同文献は，与えられた文中の各トークンの正誤を推定する

誤り検出を対象にして性能を調査している．様々なコーパスに対して，エンコーダを BERT，出力層を softmax 層とした誤り検出器が，図 1 のような性能曲線を示すと報告している．以降の実験では，ハイパーパラメータの設定も含めて同じ検出器を用いる．

解説文生成 [4] も本研究に関係が深い．解説文生成とは，語学学習のための解説文章を生成するタスクである．Hanawa ら [5] は深層学習に基いた手法が同タスクに有効であることを示している．同時に，深層学習ベースの生成は，生成能力が非常に高いためフェイク解説文を生成してしまうことを指摘している．フェイク解説文とは，存在しない規則を説明する解説文のことである（例：***considerate* は自動詞なので目的語の前に前置詞は必要ありません。**[1]）．フェイク解説文は，誤った知識を獲得させてしまう可能性があるため極力避けるべきである．

フェイク解説文を避ける有効な手段は，解説文生成を分類問題として解くという方法である．すなわち，文法誤りを詳細なサブタイプに分類し分類結果に紐づいた解説文を出力する．この方法であれば，事前に解説文を用意しておけるため，フェイク解説文が出力されることはない． 問題は，生成タスクでは分類カテゴリが自明でないということである．たとえ分類カテゴリが与えられたとしても，誤りを詳細なカテゴリに人手で分類して訓練データを作成することは容易ではない．本稿では，仮説を利用してこれらの問題を効率よく解決することを試みる．

3 一般誤りの分離可能性

3.1 手法

ここでのタスクは，入力として与えられた学習者コーパスから，一般誤りのみを抽出するというものである．ただし，入力コーパス中の誤り位置は与えられているとする．すなわち，誤り（もしくは訂正）情報が付与されたコーパスから一般誤りのみを抽出するという問題設定を考える．なお，2 節で述べたように，以降では，誤り検出器として文献 [1] の BERT ベースの検出器を用いる．

基本的な処理の流れは，1 節の仮説に基づき，(1) 入力コーパスを少量データに分割，(2) その少量データで誤り検出器を訓練，(3) 訓練済み検出器を残りのコーパスに適用し，検出された事例を一般誤りとして出力，となる．この単純な方法では，単語固有誤りが混入してしまう．なぜなら，少量の訓練コーパスでも，学習者やライティングトピックに共通してみられる単語固有誤りが，ある程度の頻度になり検出規則が獲得されてしまうからである．

そこで，提案手法では，コーパスの多重分割とトピック交差訓練という二つの工夫を行う．以下，図 2 に基いて提案手法を説明する．

ステップ（1）で，コーパスの多重分割を行う．具体的には，入力コーパスから N 文を M セット抽出する．ただし，N は小さな値とし，各セットに重複はないとする．残りは，一般誤り抽出用として，ステップ（3）で使用する．

ステップ（2）で，上述 M セットを用いて誤り検出器の訓練を独立に行う．ここでは，各トークンの正誤を推定する二値分類問題を解くことに注意が必要である．したがって，新たに一般／単語固有誤りの情報をアノテーションする必要はない．

最終ステップ（3）で，M 個の訓練済み検出器を残りのコーパスに適用して誤りを検出する．仮説に従い，M のうち大部分（θ_M 個以上の検出器）で検出された誤りを一般誤りとして出力する．

オプションとして，トピック交差訓練を行うことができる．ステップ（1）でコーパスの多重分割を行う際に，訓練用サブコーパスと一般誤り抽出用コーパスでライティングトピックが被らないように分割する．そうすることで，ライティングトピックに共通する単語固有誤りが検出されることを抑制する．

3.2 評価と分析

対象データとして，学習者コーパス ICNALE [6] をベースにした前置詞誤り解説文データセット [7] を用いる．ICNALE では，アルバイトと喫煙に関する二種類のトピックが用意されている（以降，それぞれ，PTJ と SMK と表記する）．当該データセットでは，各文書に対して，前置詞に関連した誤りの位置と解説文の情報が付与されている．ただし，通常の前置詞誤りより広い範囲の誤りを対象にしている．例えば，主語として使用された動詞（例：*Lean

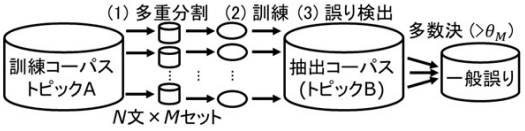

図 2 一般誤り抽出の流れ．

1） 自動詞は目的語の前に前置詞を必要とする．そもそも，*considerate* は形容詞である．

English is difficult.），句と節の混同（例：**because of I like it*）などを含む（詳細は文献 [7] を参照のこと）．

このデータセットの訓練データで訓練を行い，評価データを対象にして抽出を行った（分割は文献 [5] に従う）．付録 A に同データの統計を示す．各種パラメータは次のように設定した：$M = 10$ (多重分割の数)；$N = 800$（サブコーパス中の文数）：$\theta_M = 1, \ldots, 10$（一般誤りと認める検出数）．

図 3 に結果を示す[2)]．同図より，高い精度で一般誤りのみを抽出できていることが分かる．特に，トピック交差訓練により recall，precision 共に高まり，$\theta_M = 9, 10$ では単語固有誤りが混入していないことが分かる．全体的に，recall はそれほど高くはないが，ある程度のサイズの学習者コーパスを入力とすれば，一定量の一般誤りの事例を精度高く収集できるともいえる．次節以降で，このように収集された一般誤りの事例が，詳細な誤りサブタイプの発見と解説文生成に有益であることを示す．

4　詳細なサブタイプの発見

前節の結果にクラスタリングを適用することで，一般誤りの詳細なサブタイプを発見することを試みる．ただし，一定量の一般誤りの事例を対象とするため，一方のトピックの訓練データで誤り検出器の訓練を行い，もう一方のトピックの訓練データを抽出対象とする．一度に大量の事例を吟味することは困難であることを考慮して，少量の事例から始め，段階的に事例の量を増しながらクラスタリングする．幸い，パラメータ N と θ_M により，一般誤り抽出量を調整できる．具体的には，$N = 200$ から始め，θ_M を 10 から 1 づつ減らして行き，事例数が 30 を超えたところで抽出を終了する．所定の事例数が得られない場合は，$N = 400, 800, 1600$ と順に増加させ，同じ手順で抽出を繰り返す．

このようにして得た一般誤りの事例（候補）に対して，Ward 法を利用した階層型クラスタリングを適用し，詳細な誤りタイプの発見を行う．距離として，ベクトル間の L_2 ノルムを用いる（ノルムが 1 となるように正規化することで余弦類似度によるクラスタリングと等価とした）．ベクトルとして誤り検出器内の BERT の出力を用る（誤り区間が複数トークンに渡る場合，先頭のトークンに対応するベクトルを用いる）．クラスタの解釈性を考慮して，いずれかのクラスタ内の事例数が 10 を超えたところで

2)　図の左から右へ進むにつれて θ_M が小さくなっている．

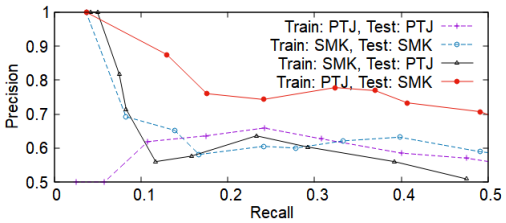

図 3　一般誤り分離の性能．

クラスタリングを終了する（クラスタ間距離の最大値が 1.5 を超えた場合も終了する）．このクラスタリング結果を人手で吟味し，各事例に詳細なサブタイプのラベルを付与する．この過程を三回繰り返し最終的な結果とする．二巡目以降では，それ以前のクラスタリング結果に，新たに抽出した一般誤り事例を加えてクラスタリングを行う．

以上の手順で一般誤りにおける 21 種類のサブタイプを発見した．紙面の関係から詳細は付録 C に譲るが，前置詞を伴った主語，他動詞に前置詞を付けた誤り，主語として使用された動詞などを発見した．事前に一般誤りの抽出を行うことで，似通った事例がクラスタリング対象となる傾向がみられた（少量データで訓練されたどの検出器でも検出される誤りであるため）．通常，サブタイプ数が増えるにつれ，適切なラベルを選択するのが難しくなるが，クラスタリングによりサブタイプごとに事例がまとまっているため作業が容易になる．特に，二巡目以降のクラスタリングでは，それ以前にラベル付けされた事例が一緒に提示されるため，その効果は高くなる．比較のために，抽出対象のコーパスからランダムに誤り事例を 800 選択し，同様にクラスタリングを適用したが，事例の約 4 割が単語固有誤りとなり，ラベル付けがより困難であった（クラスタ終了条件により実際にクラスタリングされたのは，PTJ で 109，SMK で 104 の事例であった）．実際，付録 C の表 6 に示したクラスタリングの良さを表す各種統計量にもそのことが表れている．

5　サブタイプに基づく解説文生成

前節で得られたサブタイプを利用して，解説文生成を分類問題として解くことを考える．入力中の誤り箇所を特定し，詳細なサブタイプを推定し，そのサブタイプに対応した解説文を生成するというタスクを想定する（ただし，後述するように，評価実験では推定したサブタイプと人手で付けたサブタイプの一致で性能評価を行う）．したがって，誤りトークンでなく，誤りを含む文が入力の単位となる．誤り箇所とサブタイプの推定には，BERT ベースの誤

り検出器を流用する．すなわち，出力 softmax 層をサブタイプ（に加えて単語固有誤りと正しい語）とする．

実験手順は次の通りである．まず，サブタイプごとに事例の増加を行った．サブタイプの事例と訓練データ中の事例との余弦類似度を計算し，類似度が高いものから順に各サブタイプの事例数の合計が最大 40 となるまで追加した．その結果を人手で吟味し，必要に応じてサブタイプのラベルを変更した．その結果，事例数が 10 以上となったサブタイプを生成対象とした（PTJ と SMK でそれぞれ 9 タイプと 13 タイプが対象となった）．対象となったサブタイプは付録 C の表 5 に記されている．なお，この過程およびクラスタリングの過程で得られた単語個別誤りも訓練データに含めた．

この結果を訓練事例（うち各サブタイプ 2 割を開発データ）としてサブタイプを推定した．検出性能向上のため，オリジナル（二値分類）の BERT ベース検出器も併用した．オリジナルの検出器を 10 個の異なるシードを用いて全訓練データで訓練した．その結果を，評価データに適用し，多数決をとり誤り箇所を決定した．誤り箇所に，サブタイプ用の BERT ベース検出器を適用しサブタイプを推定した．こちらも異なるシードで 10 回訓練し多数決をとった．なお，単語固有誤り／正しい単語と推定した場合は解説文を生成しないと取り扱った．また，訓練と評価では異なるトピックのデータを用いた．

表 1 の「分類」行に生成性能を示す[3]．「対象のみ」行は，上述のサブタイプだけを対象にした性能を示す．比較用に，「生成」行に，文献 [5] で最高性能を達成した深層学習生成手法（pointer generator）による生成性能も記した．また，表 2 に，PTJ を対象にしたサブタイプごとの性能も示す（紙面の関係から，SMK については付録 B に示す）．なお，評価データ中に出現しなかったサブタイプは省略した．

表 1 より，生成問題として解いたほうが性能は高いことがわかる．分類問題として解いた場合，解説対象となるのは一般誤りの限られたサブタイプのみであることが大きな理由である．ただし，訓練データの作成にかかるコストを考慮する必要がある．本実験では，PTJ で 261 事例（29.0/サブタイプ），SMK で 275 事例（21.2/サブタイプ）にサブタイプのラベルを付与しただけである．これに対して，pointer generator の訓練には約 10 倍の訓練データを用いている．しかも，単なるラベル付けではなく，解説文の記述が訓練データ作成に必要となる．加えて，分類問題の場合，フェイク解説文が生成されないことが保証されており生成結果の信頼性も高い．出力されるのは表 2 に示されるようなサブタイプに紐づいた解説文だけである．一方で，生成問題として解いた場合，何が出力されるかを事前に把握することは困難であり，実用上の大きな課題となる．実際，人手で分析したところ，pointer generator の生成結果のうち，PTJ で 8.3%，SMK で 6.6% がフェイク解説文に該当した．以上の通り，分類問題として解く本稿のアプローチは，訓練データ作成コストと生成結果の信頼性という面で大きな利点がある．

3） 評価データ中の全誤りに対して，人手でサブタイプのラベルを付与し，推定されたラベルと一致した場合，生成に成功したとみなした．

表 1 生成性能評価結果.

	PTJ		SMK	
	Recall	Precision	Recall	Precision
生成	0.268	0.469	0.355	0.559
分類	0.134	0.652	0.117	0.379
対象のみ	0.462	0.652	0.357	0.379

表 2 サブタイプごとの生成性能（PTJ）.

サブタイプ	頻度	Recall	Precision	$F_{1.0}$
iV+dobj	25	0.360	0.900	0.514
IN+inf	11	0.818	0.818	0.818
tv+IN	11	0.364	0.333	0.348
Q+*of*	9	0.556	0.833	0.667
tV/iV	4	0.250	0.500	0.333
MD+to-inf	3	0.667	0.500	0.571
cV+to-inf	1	0	0	0

6 おわりに

本稿では，一般誤りの分離可能性という仮説を導入し，その仮説を解説文生成に応用することについて述べた．コーパス中の一部の事例ではあるが，一般／単語固有誤りの識別を直接訓練せずとも 高精度で一般誤りのみを分離できることを示した．また，その結果を用いて一般誤りの詳細なサブタイプを発見した．更に，その結果を利用すると，解説文生成が分類問題として解けることも示した．生成性能は生成問題として解いた場合に劣るものの，分類問題として解くことで，訓練データ作成コストと生成結果の信頼性という面で有利な点があることを示した．今後は，前置詞誤り以外でも仮説が成り立つかを調査する予定である．

謝辞

本研究は JSPS 科研費 JP22K12326 の助成を受けたものです.

参考文献

[1] Ryo Nagata, Manabu Kimura, and Kazuaki Hanawa. Exploring the capacity of a large-scale masked language model to recognize grammatical errors. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 4107–4118, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. Frequency effects on syntactic rule learning in transformers. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 932–948, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Ryo Nagata. Toward a task of feedback comment generation for writing learning. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3206–3215, Hong Kong, China, November 2019. Association for Computational Linguistics.

[5] Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. Exploring methods for generating feedback comments for writing learning. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 9719–9730, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Shinichiro Ishikawa. **A new horizon in learner corpus studies: The aim of the ICNALE project**, pp. 3–11. University of Strathclyde Publishing, Glasgow, 2011.

[7] Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. Creating corpora for research in feedback comment generation. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 340–345, Marseille, France, May 2020. European Language Resources Association.

付録

A　実験に用いたデータの統計量

表 3 に，実験に用いたデータ [7] の統計量を示す．評価行のカッコ内の数は一般誤りの数を表す．一般誤りと単語固有誤りの判断は，付与された解説文の内容を参照し人手で行った．なお，データの分割は文献 [5] に従う．

表 3　対象データの統計値．

	PTJ		SMK	
	文数	誤り数	文数	誤り数
訓練	12,163	2,439	12,312	2,341
開発	1,129	245	1,160	230
評価	1,042	224 (120)	1,023	214 (108)

B　誤りタイプごとの生成性能

表 4 に SMK におけるサブタイプごとの生成性能を示す．評価データに出現しなかったサブタイプは省略した．

表 4　サブタイプごとの生成性能（SMK）．

サブタイプ	頻度	Recall	Precision	$F_{1.0}$
iV+dobj	11	0.364	0.800	0.500
tV+IN	10	0.100	0.125	0.111
Q+*of*	10	1	0.500	0.667
IN+CLAUSE	10	0	0	0
BE+to/for	8	0.250	0.500	0.333
IN+inf	7	0.571	0.333	0.421
IN+subj	5	0.600	0.500	0.545
cV+to-inf	4	0.250	1	0.400
IN+RB	2	0	0	0
MD+to-inf	1	0	0	0
tV/iV	1	0	0	0
tV+TO+dobj	1	0	0	0

C　サブタイプ発見の詳細

4 節で発見された一般誤りの詳細なサブタイプを表 5 に示す．全部で 21 種類のサブタイプが確認されたが，紙面の関係で解説文生成の対象となったサブタイプのみを掲載した．

表 6 に，サブタイプ発見時に利用したクラスタリングに関する統計量を示す．各種統計値は，PTJ，SMK でそれぞれ三回クラスタリング（計六回）したときの平均値である．また，「ランダム」とは，抽出対象のデータから誤り 800 個をランダムに選択し，クラスタリングした結果に対応する（ただし，クラスタに課された諸条件により実際にクラスタリングされたのは平均で 106.5 事例である）．クラスタ併合率とは，二つ以上のクラスタに同一のラベルが付与されたためにクラスタの併合が起こった割合である．具体的には，クラスタ内の事例数に対する併合が起こったクラスタ内の事例数の割合である．同様に，クラスタの分割回数に対して分割率も計算した．

表 6 より，提案手法と「ランダム」では一般誤りの割合が大きく異なることがわかる．その影響で，「ランダム」では単語固有誤りを人手で除去するという作業が多くなる（また，単語固有誤りはクラスタリングでうまくグルーピングされにくい傾向がみられた）．更に，クラスタの併合率の値により，提案手法のほうが同じサブタイプの事例が二つ以上のクラスタに入ることが少ないことがわかる．これは，一般誤りの抽出の効果により，似通った事例がクラスタリング対象となるからである（すなわち少量データで訓練されたどの検出器でも検出される誤りであるため）．

表 5　獲得された一般誤りの詳細なサブタイプ．

サブタイプラベル	詳細と例
iV+dobj	自動詞と直接目的語の組み合わせ **I agree the idea.*
IN+inf	前置詞を用いた *to*-不定詞 **a book for read*
tV+IN	他動詞と前置詞の組み合わせ **We thanked for him.*
Q+*of*	数量詞+*of* と数量詞（形容詞）との混同 **Most of the students in US work part-time.*
tV/iV	他動詞／自動詞の使い分けと前置詞 **He thought it.*
MD+to-inf	助動詞+*to*-不定詞 **We can to go.*
IN+CLAUSE	前置詞+節 **This is because of the book is interesting.*
BE+JJ+to/for	*be* 動詞+評価の形容詞+*to*/*for* の使い分け **It is good to me.*
IN+subj	前置詞を伴う主語 **In this café serves good coffee.*
cV+to-inf	使役動詞＋目的語＋ *to*-不定詞 **He made her to go.*
tV+TO+dobj	*to*-不定詞をとる他動詞+*to*+名詞 **I want to the book.*
IN+RB	前置詞＋副詞 **at everywhere*

表 6　クラスタリング結果に関する各種統計量

	提案手法	ランダム
クラスタリングされた事例数	48.3	106.5
一般誤りの割合	0.87	0.41
発見されたサブタイプ数	16	16
サブタイプ当たりの事例数	6.9	6.7
クラスタの併合率	0.14	0.27
クラスタの分割率	0.13	0.10