文法誤り検出/訂正における訓練データと評価データのドメイン不一致による性能向上現象

木村 学[1] 永田 亮[1,2]
[1] 理化学研究所 [2] 甲南大学
manabu.kimura @ a.riken.jp nagata-nlp2025 @ ml.hyogo-u.ac.jp.

概要

本稿では，文法誤り検出/訂正においては，訓練データと評価データのドメインが同じときよりも異なるときのほうが性能が高くなるという直感に合わない現象が起こることを報告する．特に，書き手の習熟度が低いデータを訓練に用いると同現象が生じやすいことを示す．この現象は，実質の訓練データ量という観点から自然に説明できることも明らかにする．更に，このことが外国語習得研究および文法誤り検出/訂正研究に与える示唆についても述べる．

1 はじめに

本稿では，文法誤り検出/訂正では，モデルの訓練に関して直感に反した現象が生じることを報告する．従来の知見では，機械学習モデルの性能を高めるためには，モデルの適用先（評価データ）のドメインと訓練データのドメインを一致させる**ドメイン内訓練**が重要であるとされてきた．実際，様々なタスクで性能が高くなることが示されている（例：画像認識 [1]，機械翻訳 [2]，機械読解 [3]）．しかしながら，文法誤り検出/訂正では，両者のドメインが異なる**ドメイン外訓練**がドメイン内訓練と同等以上の性能を達成することを示す．具体的には，書き手の習熟度が低い文章を訓練データとして用いると，検出/訂正対象の文章の書き手の習熟度にかかわらず，性能が最も良くなる傾向にあることを示す．特に，訓練データの量が少ない際にこの傾向は顕著となる．一から大量の訓練データを作成するのではなく，既存のモデルを fine-tune することが現在の主流であることを考慮すると，この知見は文法誤り検出/訂正システムを開発する上で重要である．

4 節では，文法誤り検出を対象にして，ドメイン外訓練がドメイン内訓練の性能を上回ることを示す．英語学習者コーパス 5 種類を使用し，訓練データ，評価データ 25 通りの組み合わせについて，ドメイン内外訓練の効果を検証する．

5 節では，文法誤り訂正を対象にして同様の検証を行う．ただし，文法誤り検出で顕著な性能の逆転現象が見られた 3 種類の学習者コーパスを対象とする．

6 節で実験結果を考察し，一見不可解な性能の逆転現象が，実質的な訓練データの量という観点から，シンプルに説明できることを明らかにする．更に，その考察から得られる外国語習得研究と誤り検出/訂正研究に関する示唆についても述べる．

2 関連研究

文献 [4, 5] によると，文法誤り検出では，Transformer エンコーダに基づくトークン分類手法が最高性能を達成する．このことを踏まえ，本稿では，BERT [6] に基づいた手法 [7] を検出器として用いる．

文法誤り訂正手法の性能については文献 [8] が詳しい．同文献によると，様々なベンチマークデータで，Transformer デコーダに基づいた手法が高い性能を達成する．本稿では Transformer デコーダの中で幅広く利用されている GPT-4 を訂正器として用いる．

我々が知る限り，文法誤り検出/訂正でドメイン外訓練がドメイン内訓練を上回ることを報告する研究は過去に存在しない．Chollampatt ら [9] は，英文誤り訂正で，訓練データにおける母語の影響を調査しているが，ドメイン内外訓練の優劣を調査するものではない．特に，ドメイン内外の訓練データ量が 7 倍以上異なり，ドメインの影響は定かでない．

3 実験データ

本稿では，書き手の習熟度をドメインと捉え，習熟度が異なる様々な英語学習者コーパスでドメイン内外訓練の性能差を明らかにする．具体的に

は，BEA-2019 Shared Task [10] で配布された誤り情報付き英語学習者コーパス 5 種類（BEA-A，BEA-B，BEA-C，FCE，NUCLE）を用いる．各コーパスの習熟度の関係は，BEA-A < BEA-B ≒ FCE < BEA-C < NUCLE である．各コーパスの詳細は付録 A に示す．

文法誤り検出の実験では，上述 5 種類から得られる 25 通りの訓練データと評価データの組み合わせについて性能を評価する．更に，訓練データ量と検出性能の関係を調査するために，訓練データのランダムサンプリングを行う．まず，各コーパスを，訓練，開発，評価の三つに分割する（分割の比率などの詳細は，付録 A を参照のこと）．その後，訓練データから，100，300，500，1,000，3,000，5,000，9,000，全文とサンプリングし訓練に用いる．ただし，NUCLE は文量が多いので，さらに，30,000 文についても調査する．サンプリングは異なるランダムシードで 4 回行い各文量について 4 種類のセットを作成する．それぞれで，検出器の訓練を行う．

文法誤り訂正の実験では，OpenAI では 1 日に実行できるジョブの数に限界があり，また，訓練に多くの時間を要するためサンプル数を少なくする．具体的には，100，500，1,000，5,000，全文とサンプリングする．それぞれの文量で用意するデータセットも，1,000 文以下のとき 3 セット，5,000 文以上では 1 セットと誤り検出のときより少なくする．

4 文法誤り検出

4.1 文法誤り検出手法

本稿で扱う文法誤り検出の問題は，従来研究 [4, 11] と同様に，与えられた文中の各単語の正誤を判定するトークン二値分類問題とする．文法誤り検出手法は，文献 [7] の手法とする．具体的には，BERT の最終層の上に softmax 層をつけて，各単語の二値分類を行う．訓練は，事前学習済みの BERT のモデル bert-base-uncased[1] を fine-tune することで行う．

モデルの訓練は異なるランダムシードで 4 回行う．開発データの $F_{1.0}$ が最も高かったエポックのモデルを評価に用いる．性能値として，$F_{1.0}$ の 4 回の平均値を報告する．なお，最大エポック数などの主要なハイパーパラメタは文献 [7] と同じ値を使用する．

4.2 実験結果

図 1 に結果を示す．図 1 中の小図は，5 種類の学習者コーパス（評価データ）に対応し，小図上のラベルが評価データ名を表す．また，括弧内は，CEFR の習熟度である（文献 [12] に記載された情報に基づいている；なお，CEFR では A，B，C，C1 の順に習熟度が高くなる）．横軸と縦軸は，それぞれ訓練データ量（文数）と $F_{1.0}$ に対応する．ただし，横軸は対数スケールである．また，ドメイン内訓練は太線，ドメイン外訓練は細線で示している．各プロットは検出性能の平均値であり，エラーバーは標準偏差の大きさを表す．

図 1 より，評価データ 5 種類のうち 4 種類の BEA-C，BEA-B，NUCLE，FCE についてドメイン外訓練がドメイン内訓練を上回るという逆転現象が見られる．習熟度が最も高い BEA-C が評価データであるときに，その傾向が最も顕著である．また，見方を変えると，習熟度が最も低い BEA-A を訓練データとすると，ドメイン内外訓練にかかわらず検出性能が最もよくなるともいえる．BEA-B，NUCLE，FCE では訓練データの量が増えるとドメイン内訓練は最高性能に近づくが，それでもドメイン内訓練がドメイン外訓練の性能を上回ることはない．これらの結果は，機械読解などの他のタスクで知られる，ドメイン内訓練がより高い性能を達成するという知見に反する．

5 文法誤り訂正

5.1 文法誤り訂正手法

文法誤り検出と同様の直感に異なる現象が起きるか確認するため文法誤り訂正でも実験を行う．評価データとして，文法誤り検出で逆転現象が顕著であった BEA-C を用いる．訓練データは BEA-A，BEA-B，BEC-C とする．訂正モデルには OpenAI が提供する GPT-4[2] を用いる．

学習者が記述した原文と訂正文を入力と出力のペアとして fine-tune を行う．具体的には，原文と訂正文をプロンプトに組み込み，訂正モデルを fine-tune する[3]．プロンプトは文献 [13] と同一のものを用いる．ハイパーパラメタも同文献と同一とする．ただし，OpenAI のサービスの制限の関係で，エポック

1） https://huggingface.co/google-bert/bert-base-uncased

2） gpt-4o-2024-08-06.

3） https://platform.openai.com/docs/guides/fine-tuning

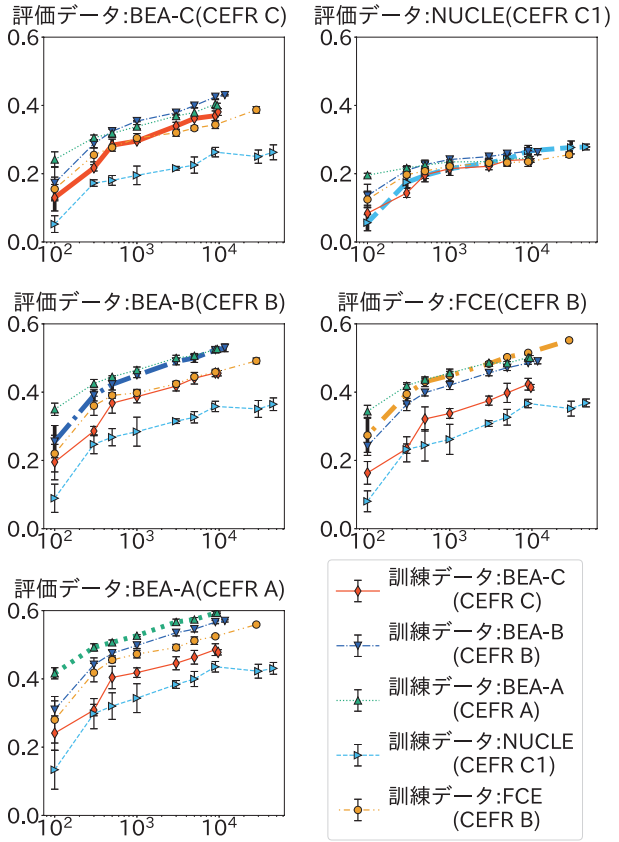

図 1　ドメイン内外訓練における訓練データ量（文数）と誤り検出性能（$F_{1.0}$）の関係．太線がドメイン内訓練に対応する．

数は 3 に固定する．

評価尺度は $F_{1.0}$ とする（スコアラ ERRANT [14] を用いる）．文法誤り検出の実験と同様に異なるランダムシードで訓練を行い，開発データについての $F_{1.0}$ が一番高いエポックのモデルを用いる．ただし，誤り検出に比べ訓練に多くの時間を要するため，1,000 文以下のときは 3 回，5,000 文以上のときは 1 回の訓練とする．

5.2　実験結果

図 2 に結果を示す．図の見方は図 1 と同様である．

図 2 より，誤り検出ほど顕著ではないが，ドメイン外訓練がドメイン内訓練と同等以上の性能を達成することがわかる．訓練データが 5,000 文以上の場合には，ドメイン内訓練がドメイン外訓練を若干上回るが，最高性能を達成しているわけでない．この結果は，一定量以上の訓練データを用いて大規模なモデルを fine-tune することの難しさを表している．fine-tuning の方法を工夫することで異なる結果が得られる可能性があるが，その方法は自明でない．

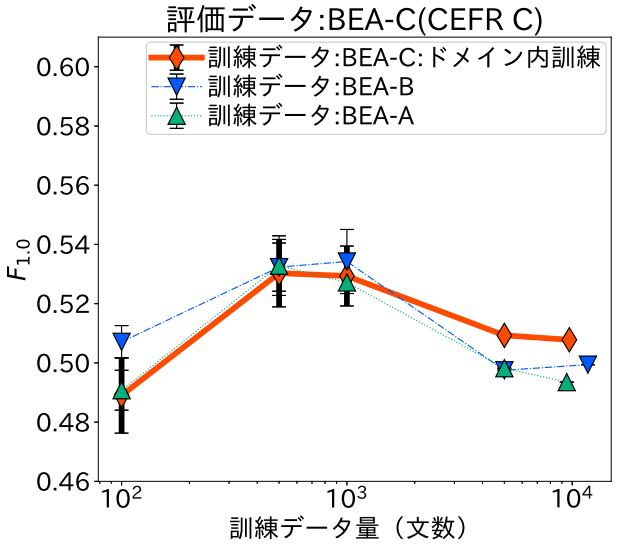

図 2　ドメイン内外訓練における訓練データ量（文数）と誤り訂正性能（$F_{1.0}$）の関係．

6　考察

4 節と 5 節で見た性能の逆転現象は一見すると直感に合わず，不思議に思える．しかしながら実質的な訓練データの量という観点から自然に説明可能である．

誤り検出/訂正における正例は誤りと訂正のペアである．従って，実質的な訓練データの量は，訓練データ中の文数ではなく誤りの数とするのが自然である．前節までに，書き手の習熟度が低いコーパスを訓練データに用いると性能の逆転現象が顕著となることを見た．この理由は，習熟度が低いと誤りが頻出し，実質的な訓練データの量が多くなるためと予想される．

このことを踏まえて，評価結果を誤り数と検出性能の関係として改めて示す．図 3 と図 4 が，それぞれ誤り検出と訂正の結果に対応する．図の見方は，図 1 と図 2 と同様である．ただし，横軸は文数でなく誤り数に対応する．各プロット点の横軸の値は，シードが異なる複数のデータセットの誤り数の平均値であり，エラーバーは標準偏差の大きさを表す．

予想通り，図 3 と図 4 では，ドメイン内訓練（太線）がどの図でも（ほぼ）最上位に位置している．また，各線はより密集し，全体的に性能差が小さくなっていることがわかる．

このように，実質的な訓練データの量という観点から分析すると，一見不可解に見えるドメイン内外訓練の性能逆転現象もシンプルに説明できる．逆転現象も，ある意味当然のことに思える．

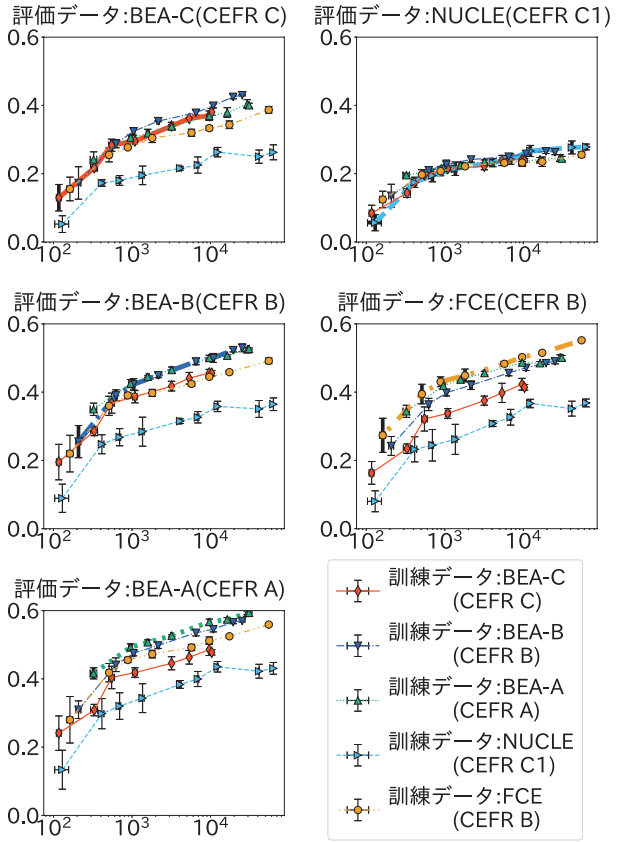

図 3　ドメイン内外訓練における実質の訓練データ量（誤り数）と検出性能（$F_{1.0}$）の関係.

しかしながら，依然，大きなが疑問が残る．なぜ，図 3 と図 4 では，ドメイン内訓練に匹敵する性能を達するドメイン外訓練が存在するのであろう？実質の訓練データ量が同じであれば，同種の誤りをより多く含むであろうドメイン内訓練のほうが性能がよくなると予想できる．両図の結果は，この予想に従わない．

以上の疑問は，次に述べるように，外国語習得研究と誤り検出訂正の研究に重要な示唆を与える．図 3 と図 4 は，習熟度にかかわらず共通した誤りが多く存在することを示唆する．言い換えれば，上級者（例えば，CEFR C）になっても，初学者（CEFR A）や中級者（CEFR B）と同種の誤りをするということである[4)]．そうでなければ，BEA-A（CEFR A）や BEA-B（CEFR B）の訓練データで，BEA-C（CEFR C）の誤りをドメイン内訓練と同等以上に検出/訂正できないはずである．もう一つの示唆として現状の誤り検出/訂正手法の性能限界がある．習熟度固有の誤りが存在したとしても，それが検出/訂正できなければ性能差としては現れない．特に，習熟度が高

4)　ただし，付録 A の表 1 に示すように誤りの頻度自体は低くなる.

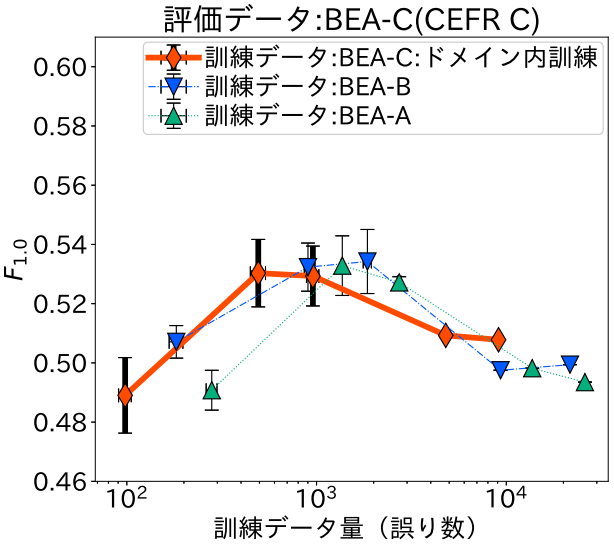

図 4　ドメイン内外訓練における実質の訓練データ量（誤り数）と訂正性能（$F_{1.0}$）の関係.

いコーパスでは，検出/訂正に書き手の意図やテキスト外の知識が必要となる誤りが出現する可能性が高い．仮にそのような誤りが訓練データと評価データの両方に含まれたとしても現状の手法では検出/訂正することは困難である．その結果，ドメイン内外訓練に差がつきにくい傾向となる．

以上の考察は誤り検出/訂正システムを開発する上でも重要である．通常，訓練データの作成は，文や文書単位で行われる（誤り単位で訓練データの作成を行うことは困難である）．したがって，できるだけ誤りを多く含みそうな習熟度の低い（具体的には，CEFR A などの）書き手の文章を収集することが効率的なシステム開発に繋がる．今回の実験結果によると，CEFR A のデータで，（性能限界の範囲内で）CEFR B，C にも対応できる．既に，英語については比較的豊富な訓練データが利用可能であるため，この知見はそれほど有益でないかもしれない．しかしながら，このことは恐らく英語以外についても成り立つと予想されるため，適用範囲は広い．

7　おわりに

本稿では文法誤り検出/訂正において，ドメイン外訓練の性能がドメイン内訓練を上回る現象を報告した．この現象は，習熟度の高い評価データに対して，習熟度の低い訓練データを使用した場合に顕著になることを明らかにした．また，この現象を実質的な訓練データの量という観点から説明した．更に，このことが外国語習得研究と誤り検出/訂正研究に与える重要な示唆を報告した.

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] Judy Hoffman, Eric Tzeng, Jeff Donahue, Yangqing Jia, Kate Saenko, and Trevor Darrell. One-shot adaptation of supervised deep convolutional models. In **2nd International Conference on Learning Representations, Workshop Track Proceedings**, 2014.

[2] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, pp. 385–391, 2017.

[3] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In **Proceedings of the 2nd Workshop on Machine Reading for Question Answering**, pp. 1–13, 2019.

[4] Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. Multi-class grammatical error detection for correction: A tale of two systems. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 8722–8736, 2021.

[5] Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. In **Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning**, pp. 1–16, 2023.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proc. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 4171–4186, 2019.

[7] 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, 2022.

[8] Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In **Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 17–33, 2024.

[9] Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. Adapting grammatical error correction based on the native language of writers with neural network joint models. In **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 1901–1911, 2016.

[10] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 52–75, 2019.

[11] Marek Rei and Helen Yannakoudakis. Compositional sequence labeling models for error detection in learner writing. In **Proc. of 54th Annual Meeting of the Association for Computational Linguistics**, pp. 1181–1191, 2016.

[12] Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. Grammatical Error Correction: A Survey of the State of the Art. **Computational Linguistics**, Vol. 49, No. 3, pp. 643–701, 2023.

[13] Anisia Katinskaia and Roman Yangarber. GPT-3.5 for grammatical error correction. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, pp. 7831–7843, 2024.

[14] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, pp. 793–805, 2017.

[15] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 180–189, 2011.

[16] 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, 2013.

A　実験データの詳細

実験データの要約を表 1 に示す．表 1 には，文数及び一文あたりの誤り数を統計情報として記載している．表中の CEFR 習熟度のレベル，分野の情報は文献 [12] に記載された情報に基づいている．各コーパスの更なる詳細は文献 [15, 16, 10] を参考されたい．

各コーパスには ERRANT [14] で定義された誤りラベルが付与されている．文法誤り検出（4 節）では，誤りラベルがある単語を**誤**，それ以外の単語を**正**として利用している．

実験では，コーパスを訓練，開発，評価の三つに分割して利用した．FCE では，あらかじめ提供されている分割に従った．BEA-A，BEA-B，及び BEA-C については，訓練，開発データのみ配布されている．配布された開発データを評価データとして使用した．表 1 に示す割合で，配布された訓練データを訓練データと開発データに分割した．NUCLE についてはあらかじめ分割されていないため，表 1 に示す割合で，訓練，開発，評価データに分割した．

本稿の実験では，訓練データの文量と性能との関係を調べるために，文法誤り検出では訓練データから 100，300，500，1,000，3,000，5,000，9,000[5)]，全文と抽出したデータセットを用意する[6)]．具体的には，まず，全文をランダムに並べ替える．先頭から 100 文，先頭から 300 文，先頭から 500 文，. . . と切り取り，それぞれの文量に対応するデータセットを用意する．文法誤り訂正では，100，500，1,000，5,000，全文を用意する．文法誤り検出と異なり 300，3,000，9,000 文が無いのは，OpenAI では 1 日実行できる fine-tuning のジョブの数に限界があり，訓練に多くの時間を要するためである．

上記の方法で選んだデータセットには偏りがある可能性がある．特に 100 文のときはその偏りが性能に大きな影響を与える．そこで，ランダムシードを変え前述の手順で複数回サンプリングを行い[7)]，各訓練データ量について複数の訓練データを作成した．従って，各文量の訓練データについて複数の検出/訂正結果を得ることができる．4 節と 5 節では，報告する図はこのようにして得た結果に対する性能値の平均と標準偏差を計算している．加えて，6 節では，誤り数の平均と標準偏差を計算している．

表 1　実験で用いたコーパスの統計量．

コーパス	文数	誤り数/文	分割比 [%] 訓練/開発/評価	CEFR	分野
BEA-C	11,852	1.08	82.0/9.0/9.0	C	exam
BEA-B	14,322	2.16	82.0/9.0/9.0	B	exam
BEA-A	11,530	3.26	82.0/9.0/9.0	A	exam
NUCLE	57,151	1.26	80.0/10.0/10.0	C1	essay
FCE	33,236	1.97	85.3/6.6/8.1	B	exam

5)　全文が 10,000 を満たないコーパスがあるため，9,000 とした．
6)　NUCLE は文量が多いので，さらに，30,000 文についても用意した．
7)　サンプリングの具体的な回数は本文に示した．