大規模マスク言語モデルの文法誤り認識能力

木村 学 [1]　永田 亮 [2]　塙 一晃 [3,4]
[1]GRAS グループ株式会社　[2] 甲南大学　[3] 理研 AIP　[4] 東北大学
manabu.kimura @ gras-group.co.jp　nagata-nlp2022 @ ml.hyogo-u.ac.jp.
kazuaki.hanawa @ riken.jp

概要

本稿では，マスク言語モデルベースの文法誤り検出手法についてその誤り検出能力を調査する．実データ及び設計された人工データの実験から，**仮説 (i)**「言語モデルは，母語話者コーパスからの事前学習の際に，ある種類の文法誤りを認識するために必要な文法の知識を得ている．」，**仮説 (ii)**「言語モデルは，ごく少数の訓練データで finetune するだけで，事前学習時に獲得する文法知識を誤り検出ルールに変換し，文法誤り検出に対する高い汎化能力を獲得している．」が示唆されることを示す．

1 はじめに

近年，BERT [1] のような，大規模なコーパスで事前学習されたマスク言語モデルが，自然言語処理の幅広いタスクにおいて，性能を飛躍的に向上させることが示されてきた．これらの結果は「文法誤り検出」でも効果的であることを示唆している．しかしながら，文法誤り検出の見地での研究は他のタスクよりも圧倒的に少ない．実際，言語モデルは母語話者によって生成された言語データ（本稿は英語に特化する）から学習されているので，言語学習者によって生成された言語データには能力を発揮しないのではないか，文法誤りについての知識が存在しないのではないかという疑問が生じる．

実際のところは，文法誤りの検出と訂正のタスクで言語モデルの性能は，少数ではあるが報告されている [2, 3, 4, 5]．これらの研究は文法誤りの検出と訂正のタスクで言語モデルの効果を検証しているが，言語モデルが誤り検出/訂正に何故有効であるのかという疑問は解決されないままである．

本稿では，リサーチクエスチョン「母語話者コーパスから事前学習された言語モデルは fintune によって誤り検出タスクに何故効果を発揮するのか？」への答えを計算機実験や結果の詳細な分析から経験的に探る．具体的には，図 1 で示すように，finetune の訓練データ量の変化と検出性能 $F_{1.0}$ の関係を非言語モデルベースの手法と比較することで調査する．同図では，非言語モデルベースの手法が全ての訓練データで達成する $F_{1.0}$ を，5-10%の訓練データ量で BERT ベースの手法は達成する結果を得ている（詳細は 5 節に記述する）．さらに，BERT ベースの手法は 500 文程度で $F_{1.0}$ の上昇が飽和し始めている．これは，BERT ベースの手法に，**概要** で述べた**仮説 (i) (ii)** が成り立つことを示唆する．実際の学習者の誤りを含む実データ，設計された誤りを含む人工データの両者で実験を行い，この仮説を支持する結果や分析を報告する．

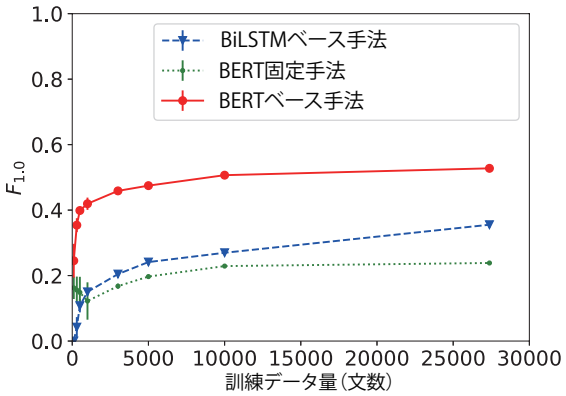

図 1　訓練データ量と性能の関係（訓練/評価データ: FCE）

2 関連研究

マスク言語モデルは文法誤り検出/訂正に適用されてきた．文献 [2, 3] は誤り検出の手法が言語モデルの利用で再現率と適合率が高まることを示している．文献 [6] は文法誤り検出に BERT ベースの文脈埋め込みを使用し，他の埋め込み手法と比較している．彼らは，BERT ベースの文脈埋め込みが効果的であることを示しているが，本研究のように BERT の finetune は行っていない．文献 [4, 5] は文法誤り

訂正で性能を同様に改善している．本稿はこれまでの研究で報告されてきた誤り検出の能力が，何故そしてどういう場合に獲得するのかを，実験や結果の詳細な分析から経験的に探る．

言語モデルの言語知識を調査する研究は様々なものが存在する [7, 8, 9]．一般的なアプローチは，文脈が与えられたときに，言語モデルが不適切な単語よりも適切な単語に高い確率を与えるかを評価するものである．調査対象の言語知識は，構文的/意味的知識から一般常識に及ぶ．本稿はこれら従来研究とは異なり，学習者コーパスに実際に出現する誤りを対象にし，マスク言語モデルの汎化性能を調査する（例えば，主語に前置詞が付随する文法誤りや自動詞/他動詞の使用に関する文法誤りなどを扱う）．

3 文法誤りデータ

本稿では，検証に実データと人工データを使用する．

実データは，英語学習コーパス，FCE [11] と解説文付き ICNALE [10] を使用する．FCE は，学習者の文とその誤りを訂正した文が含まれている．誤りタグについては，ERRANT [13] を用いて取得した．解説文付き ICNALE は，トピック「アルバイト」及び「喫煙」についての英語学習者のエッセイ文が集められた ICNALE [12] に，誤りの解説文が付与されたものである[1)]．ICNALE では，解説文付き ICNALE に収録されている前置詞誤りを検出対象とした．なお，これらのコーパスは誤りがない文も含んでいることに注意されたい．

人工データは AQUAINT Corpus に含まれる 1998-2000 年の New York Times [14] を使用し，次の誤りを加えることで生成する．

前置詞置換不定詞　*to* 不定詞の *to* が *for* に置き換わっている誤り (例：*to read* → **for read*)

動詞句主語　主語として使用された動詞句 (例：*To learn English is easy.* → **Lean English is easy.*)

前置詞＋主語　前置詞を伴う主語 (例：*The cafe serves good coffee.* → **In the cafe serves good coffe.*)

他動詞＋前置詞　前置詞を伴う他動詞 (例：*We discussed it.* → **We discussed about it.*)

自動詞＋目的語　目的語を伴う自動詞 (例：*We agree with it.* → **We agree it.*)

以上は構文解析の結果から自動生成できる．一文あたりは一つの誤りとし，前処理として spaCy[2)] をかけることで，構文解析の結果を得る[3)]．ただし，誤り文を生成する際には次のように制限を設けている．ランダムに加える前置詞は，at, about, to, in, with の 5 種類とした．ただし，**前置詞置換不定詞**のみ，常に to を for に置き換えた．**他動詞＋前置詞**で誤りを加える他動詞は，訓練/開発データでは answer, attend, discuss, inhabit, mention, oppose, resemble のみを，評価データは approach, consider, enter, marry, obey, reach, visit のみとした．同様に，**自動詞＋目的語**で誤りを加える自動詞は，訓練データと開発データでは agree, belong, disagree, relate のみを，評価データでは apply, graduate, listen, specialize, worry のみとした．したがって，対象とする自動詞と他動詞は，訓練/開発データと評価データとの間で重複しない．

実験では，訓練データの文数と検出性能との関係を調べるためにランダムに文をサンプリングする．実データの FCE と ICNALE については，その数を 100，300，500，1,000，3,000，5,000，10,000，全文と変化させている．開発データと評価データは，訓練データの変化に関わらず別の固定した文を用いる（開発データと評価データの具体的な統計量は付録 B に示す）．なお，訓練，開発，評価データには，学習者の誤りがない文も含まれている．

人工データの訓練データは，誤りの種類毎に $2^k (1 \leq k \leq 10)$ 文をランダムにサンプリングすることで，10 個の訓練データのセットを用意する．開発データについては誤りの種類毎に 200 文をランダムに選択する．評価データについては同様に誤りの種類毎に 200 文をランダムに選択し，さらに，誤りを加えていない他の 200 文を追加した．実データと同様に，開発データと評価データは訓練データの変化に関わらず固定されている．

4 文法誤り検出手法

本稿で扱う文法誤り検出の問題を次のように定式化する．単語の系列と長さを $\{w_i\}_{i=1}^N$ および N とそれぞれ表記する．さらに対応するラベルの系列を $\{l_i\}_{i=1}^N$ と定義する．ここで，l_i は w_i に対応するラベルである．ラベルは，データによって次の (i)，(ii) のいずれかを仮定する．(i) 実データの場合：文法誤りが有るか無いかのみを扱い，ラベルをそれぞれ E 及び C と定義する．(ii) 人工データの場合：文法誤

1) 説明の簡略化のため，解説文付き ICNALE も以後 ICNALE と呼ぶ．

2) https://spacy.io/

3) 信頼できる構文解析結果を得るため，文は長さが 3 トークン以上かつ 26 トークン以下のものだけ選ぶ．

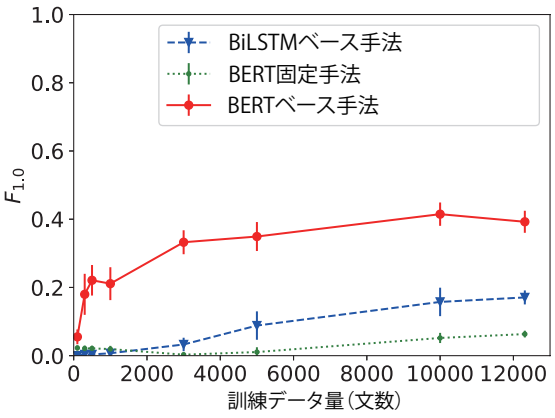

図 2　訓練データ量と $F_{1.0}$ の関係（訓練データ：ICNALE 喫煙，評価データ：ICNALE アルバイト）

りについての 5 種類と誤りの無いことを扱い，6 種類のラベルを定義する．このとき，文法誤り検出を $\{w_i\}_{i=1}^N$ が与えられた時に最適なラベルの系列 $\{l_i\}_{i=1}^N$ を予測する問題とする．

本稿で調査する BERT ベースの手法は，以下の流れで処理する．

(1) サブワード化:　全ての w_i をサブワード $\{s_j\}_{j=1}^M$ に変換する．一般にサブワードの総数 M は，入力の単語の系列の総数 N と異なる．

(2) エンコード:　$\{s_j\}_{j=1}^M$ を BERT により，埋め込みベクトル $\{\boldsymbol{b}_j\}_{j=1}^M$ にエンコードする．

(3) トークン分類:　最適なラベルを以下で計算する．

$$l_i = \arg\max \mathrm{softmax}(\boldsymbol{W}\boldsymbol{b}_j) \tag{1}$$

ここで，$\boldsymbol{W}$ は $k \times h$ の行列で，k は 2 か 6（5 種類の誤りと正しいラベルの数）のどちらかである．入力の単語の系列と対応するサブワードの系列の長さは異なる．各予測 l_i に対し，対応する単語 w_i の先頭のサブワードをエンコードした $\boldsymbol{b}_j$ のみ考慮する[4]．

比較手法の一つとして，BiLSTM ベースの誤り検出手法を選ぶ．処理は BERT ベース手法と基本的に同じであるが，次の点が異なる．**ステップ (2)** で BERT の代わりのエンコーダとして BiLSTM を使用する．**ステップ (1)** で $\{w_i\}_{i=1}^N$ を，各単語 w_i に対応する埋め込みベクトルの系列 $\{\boldsymbol{e}_i\}_{i=1}^N$ に変換する．ここで，各埋め込みベクトル $\boldsymbol{e}_i$ は通常の単語埋め込みと文字ベースの埋め込みを結合したベクトルである．文字ベースの埋め込みは [15] に従い，別の BiLSTM を用意し各単語の文字を変換することで行う．以後，この手法を **BiLSTM ベース手法**と呼ぶ．

また，BERT の finetune の効果がどの程度あるかも調査する．すなわち，BERT ベースの手法の BERT 部分を finetune 時に固定し，出力層のみを訓練データで調整するといったことも比較のため行う．以上の手続きの手法を **BERT 固定手法**と呼ぶ．

各モデルの訓練は異なるランダムのシードによって 5 回行う．次節で報告する性能の値 ($F_{1.0}$，再現率) は 5 回の結果の平均値である．評価に使用するモデルは，開発データで $F_{1.0}$ が一番高いエポックのモデルを採用する．最大エポック数などの主要なハイパーパラメタは付録 A.2 に掲載する．

5　性能評価

5.1　実データの場合

まず，実データ FCE についての訓練データ量と $F_{1.0}$ との間の関係を本稿 1 ページ目の図 1 に示す[5]．BERT ベース手法は 500 文程度の訓練データ量で $F_{1.0}$ の上昇が飽和し始めている．一方，BiLSTM ベース手法は，全訓練データでも BERT ベース手法が 300-500 文で達成できる $F_{1.0}$ の値に留まっている．以上の結果は，BERT ベース手法の高い汎化能力を示している．また，BERT 固定手法の性能は高くない．これは深層学習モデルで自由に学習できるパラメータ数が制限され，データから誤り検出するルールを上手く獲得できていないためと考えられる．言い換えると，BERT は全体のパラメータを少量の訓練データで fintune することで，誤りの検出ルールの多くを獲得できている可能性がある．

同様のことが ICNALE でも確認できる．なお，図 2 では訓練データが「喫煙」，評価データ「アルバイト」とアウトドメイン設定にも関わらず BERT ベース手法は高い汎化能力を示していることは注目に値する（他の ICNALE の結果は付録 B に示す．）．

さらに，図 1 の検出結果を誤りの種類毎に分けて，再現率についてプロットし直した[6]．その結果を図 3 に示し考察する．SPELL 以外の全ての誤りにおいて，BiLSTM ベース手法はほぼ線形に改善している一方で BERT ベース手法は早い段階で性能が

4)　他に末尾を考慮するなどできるが，素朴に先頭を選ぶ．

5)　ここで，図を描くための詳細な設定は付録 A に示す．各プロット点の訓練データ数は 3 節で述べたとおりである．

6)　誤りの種類毎の再現率は，モデルの検出結果を ERRANT を使って分類し計算している．誤りの種類のうち PUNCT 及び OTHER，そして発生頻度が 150 以下の誤りの種類は，紙面の関係で取り除いている．

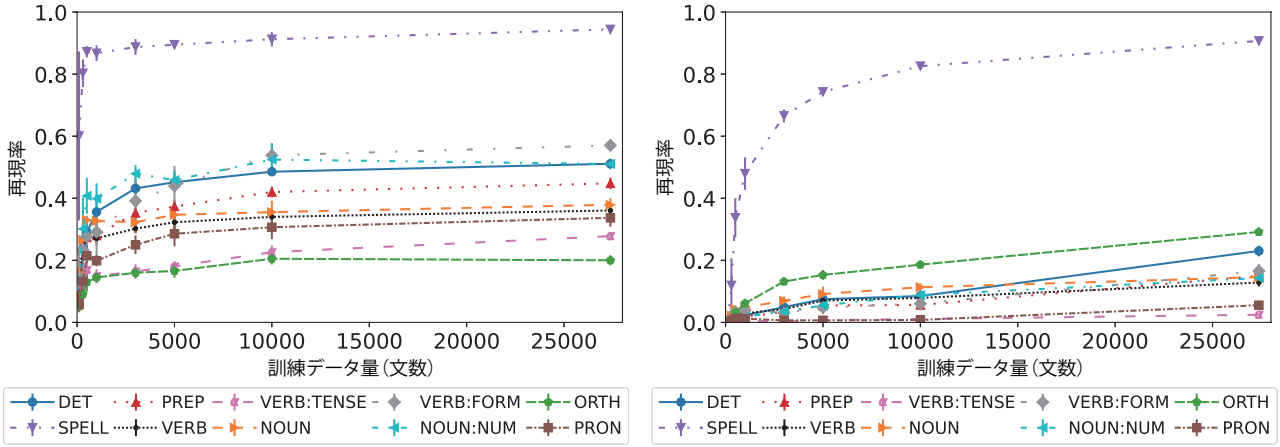

図 3　誤り種別の訓練データ量と再現率の関係（左：BERT ベース手法，右：BiLSTM ベース手法，訓練/評価データ: FCE）

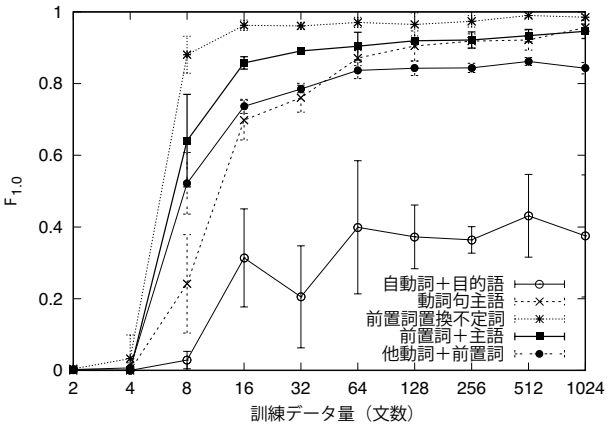

図 4　誤りの種類毎の訓練データ量と $F_{1.0}$ の関係（BERT ベース手法，訓練&評価データ：人工データ）

限界に近づいている．さらに，BERT ベース手法の 500 文のときの再現値は，BiLSTM ベース手法が全訓練データで達成する値をほとんどの誤りの種類で上回っている．この結果は，BERT に様々な文法知識が内在することを示唆する．

5.2　人工データの場合

実データで示した BERT ベースの誤り検出手法の汎化能力を人工データで精査する．図 4 に，人工データでの訓練データ量と $F_{1.0}$ との関係を誤りの種類毎に示す．**自動詞＋目的語**を除いて，16 文程度の少数のデータで性能が十分に高くなっている．例えば，**前置詞＋主語**は 16 文の訓練データだけで $F_{1.0}$ が 0.8 を超えている．一般的に，任意の名詞が主語になりえ，その出現位置も多岐にわたるため，動詞のような品詞，主語のような文の構造の概念が無いと検出が難しいタスクである．従って，BERT が品詞や文の構造といった言語知識を保持していること，それを少数の訓練データの fintune によって文法誤りの検出ルールに変換していることを示唆している．**他動詞＋前置詞**でも 16 文の訓練データだけで $F_{1.0}$ が 0.8 を超えており，同様の議論が繰り返される．今回の問題設定では，訓練データに現れない動詞の誤りを検出しなければならない（対象の自動詞/他動詞について訓練データと評価データの間で重複がない）ため，前述の言語知識に加え，各動詞の属性（自動詞，他動詞）を保持していないとこの種類の誤りを検出することは難しい．それにも関わらず BERT ベース手法は，**他動詞＋前置詞**の誤り（例：**mention in*）を訓練データで見ただけで，同じ種類の誤り（例：**visited in Atlanta*）を認識できている．以上の観測から，BERT が言語的知識を持つこと，finteune によって誤り検出のルールに変換できるという**概要で述べた仮説**が示唆される．

6　おわりに

本稿では，文法誤りを認識する大規模マスク言語モデルの能力を調査した．本稿の知見は次の 2 点に要約される．(1) BERT ベースの誤り検出手法は汎化性能が非言語モデルよりも高いことを実データでの実験で示し，500 文程度の訓練データで性能が飽和し始めることを示した．(2) 人工データと実データの実験を通じて，**仮説 (i)**「言語モデルは，母語話者コーパスからの事前学習の際に，ある種類の文法誤りを認識するために必要な文法の知識を得ている．」，**仮説 (ii)**「言語モデルは，ごく少数の訓練データで finetune するだけで，事前学習時に獲得する文法知識を大量の誤り検出ルールに変換し，文法誤り検出に対する高い汎化能力を獲得している．」を支持する結果を示した．

謝辞

本研究の一部は，JST，さきがけ，JPMJPR1758 の支援を受けたものである．

参考文献

[1] 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, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[2] Yong Cheng and Mofan Duan. Chinese grammatical error detection based on BERT model. In **Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications**, pp. 108–113, Suzhou, China, December 2020. Association for Computational Linguistics.

[3] Masahiro Kaneko and Mamoru Komachi. Multi-head multi-layer attention to deep language representations for grammatical error detection, 2019.

[4] Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4248–4254, Online, July 2020. Association for Computational Linguistics.

[5] Bohdan Didenko and Julia Shaptala. Multi-headed architecture based on BERT for grammatical errors correction. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 246–251, Florence, Italy, August 2019. Association for Computational Linguistics.

[6] Samuel Bell, Helen Yannakoudakis, and Marek Rei. Context is key: Grammatical error detection with contextual word representations. In **Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 103–115, Florence, Italy, August 2019. Association for Computational Linguistics.

[7] Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. How is BERT surprised? layerwise detection of linguistic anomalies. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 4215–4228, Online, August 2021. Association for Computational Linguistics.

[8] Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 34–48, 2020.

[9] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[10] Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. Creating Corpora for Research in Feedback Comment Generation. In **Proc. of the 12th Language Resources and Evaluation Conference**, pp. 340–345, 2020.

[11] 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, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

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

[13] Christopher Bryant, Mariano Felice, and Ted Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. In **Proceedings of 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 793–805, 2017.

[14] David Graff. The aquaint corpus of english news text, 2002.

[15] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 1638–1649, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[16] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 54–59, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

表 2　実データセットの統計

コーパス	FCE			ICNALE(アルバイト)			ICNALE(喫煙)		
分割	訓練	開発	評価	訓練	開発	評価	訓練	開発	評価
英文数	27,380	2,129	2,581	12,163	1,129	1,042	12,312	1,160	1,023
トークン数	435,768	33,720	40,498	205,355	18,276	17,192	201,304	18,242	17,318
誤り数	41,277	3,335	4,374	2,439	244	222	2,342	230	212

A　他の実験設定

A.1　実験で使用するモデル/ツール

BERT ベース手法で使用する BERT モデルは huggingface が提供する bert-base-uncased を用いる．BiLSTM ベース手法で使用する BiLSTM の実装及び埋め込みベクトルは，FLAIR [16] が提供するものを使用する．埋め込みベクトルのうち，単語の埋め込みはコード上で “WordEmbeddings(‘en’)” と実装する学習済みベクトル，文字列埋め込みは “FlairEmbeddings(‘news-forward’)” 及び “FlairEmbeddings(‘news-backward’)” を使用する．

A.2　ハイパーパラメタ設定

表 1　主要なハイパーパラメタ

バッチサイズ	32(実データ), 5(人工データ)
最適化手法	Adam with decoupled weight decay regularization
学習率	$5e^{-5}$, (0.9, 0.999)
ε	$1e^{-8}$
Weight decay	$1e^{-2}$
エポック数	50(実データ), 10(人工データ)

表 1 に実験の主要なハイパーパラメタを示す．留意すべき点は，訓練に人工データを使用するとき，文の数は最小で 10 文程度なのでバッチサイズはさらに小さい 5 を使用することである．

B　実データや実験結果の補足

表 2 に，実験で使用した実データのコーパスの統計量を示す．

本文で報告できなかった，ICNALE についての実験結果を示す．トピック「アルバイト」，「喫煙」についての訓練データ量と $F_{1.0}$ の関係をそれぞれ図 5，6 に示す．トピック「アルバイト」で訓練し「喫煙」で評価したアウトドメイン設定の結果を図 7 に示す．全ての図が 5.1 節で実験/考察した FCE と同様の傾向を示している．

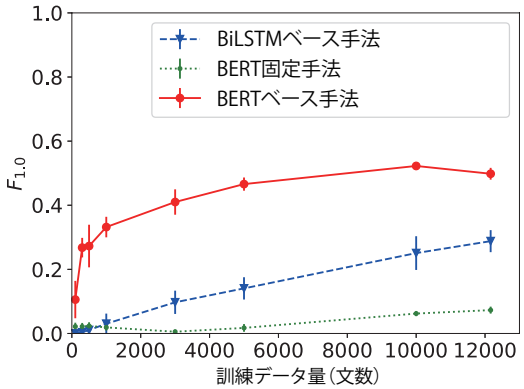

図 5　訓練データ量と性能の関係（訓練/評価データ：ICNALE アルバイト）

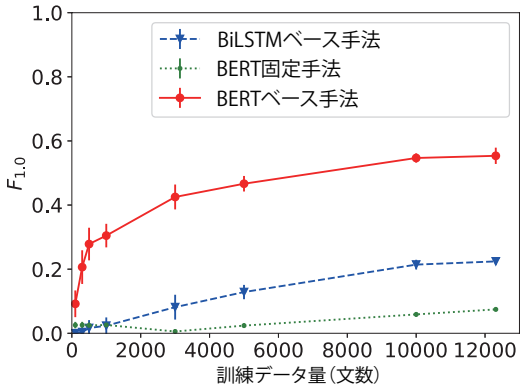

図 6　訓練データ量と $F_{1.0}$ の関係（訓練/評価データ：ICNALE 喫煙）

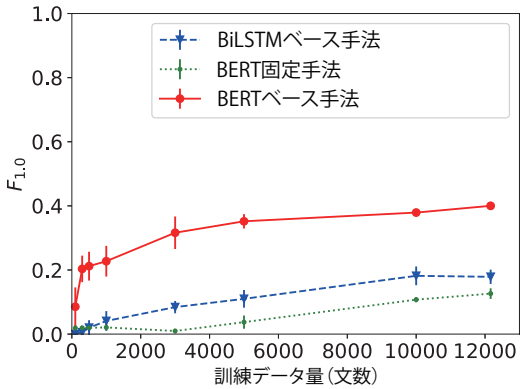

図 7　訓練データ量と性能の関係（訓練データ：ICNALE アルバイト，評価データ：ICNALE 喫煙）