事前学習モデルを用いた中国語文法誤り訂正

王 鴻飛　　黒澤 道希　　勝又 智　　小町 守
首都大学東京
{wang-hongfei, kurosawa-michiki, katsumata-satoru}@ed.tmu.ac.jp,
komachi@tmu.ac.jp

1 はじめに

文法誤り訂正とは，言語学習者が書いた誤りの含まれている文を入力とし，誤りを直した文を出力とするタスクである．ニューラルネットワークを用いた中国語文法誤り訂正の研究は word2vec などの単語分散表現によって訂正モデルを初期化することが多いが [11, 13]，単語分散表現で初期化する手法では埋め込み層しか初期化されず，訂正モデルの訓練に事前学習モデルが学習した言語の浅い情報だけが転移されている．また，中国語文法誤り訂正には以下の特徴がある：文字種が 10,000 種以上あり，日常的によく使われているものに限定しても約 3,000 種存在しており，発音もしくは字形が似ている漢字も多く存在している．そのため，学習者が書いた漢字の誤りが訓練データに含まれていない可能性があり，訂正候補を文脈から判断する必要がある．

一方，自然言語処理において BERT に基づく事前学習モデルが盛んに研究され，さまざまなタスクにおいて性能が向上している [4, 9]．BERT に基づく事前学習モデルは，大規模なコーパスで Transformer を用いたニューラルネットワークモデルを学習し，言語についての事前知識を学習する．その後，ダウンストリームのタスクでは事前学習モデルが学習した重みで事前学習モデルと同じ構造のニューラルネットワークモデルを初期化し，ダウンストリームタスクの訓練データで微調整を行う．このような 2 段階の手法により，大規模なコーパスから学習した事前知識をダウンストリームタスクに転移することで各ダウンストリームタスクにおいて性能が向上することが期待されている．

そこで，本研究では Transformer で中国語文法誤り訂正モデルを作成し，BERT に基づく中国語事前学習モデル [3] で Transformer の Encoder 部分を初期化し，中国語文法誤り訂正コーパスで微調整する方法を提案する．その結果，我々の手法は 1 つのモジュールからなるシングルモデルで NLPCC 2018 文法誤り訂正タスク [12] のテストセットで $F_{0.5}$ が 29.63 ポイントでトップチームのアンサンブルモデルとほぼ同等の結果を，さらに 4-ensemble で $F_{0.5}$ が 35.51 ポイントと同データで最高精度の結果を達成した．

2 先行研究

2.1 中国語文法誤り訂正

NLPCC 2018 で中国語文法誤り訂正タスクが開催され，言語学習ウェブサイトの Lang-8 から収集された約 100 万文の訓練データと PKU Chinese Learner Corpus から抽出された 2,000 文のテストデータが公開された．このタスクには 23 チームが参加し，6 チームが実験結果を報告した．

その中，1 位の Fu ら [5] は: 5-gram 言語モデルでスペリング誤り訂正器を作り，Transformer でサブワードレベルと文字レベルのモデルを訓練した．これらのモデルを組み合わせて 5 つの訂正結果を出力し，最終的に言語モデルを用いてこの出力結果に対してリランキングを行う．この手法は高い精度を報告しているが，使用したモデルの数が多く，組み合わせの手法も複雑である．

2 位の Zhou ら [13] は複数の訂正モジュールを段階的に組み合わせる手法を提案した．訂正モジュールはルールベースモデル，文字レベルと単語レベルの統計的機械翻訳モデル（SMT），と 4 つのニューラル機械翻訳モデル（NMT）より構成されている. これらのモデルは低レベルと高レベルモジュールより構成され，低レベルモジュールが 2 つの SMT モデルを 1 つに，4 つの NMT を 1 つにマージし，高レベルモジュールで低レベルモジュールのマージ結果とルールベースモデルの結果をマージし最終的な結果を出す．この手法は合計 7 つのモデルを訓練しており，コストが高い．

3 位の Ren ら [11] は文法を考慮できる wang2vec [8] で初期化した CNN seq2seq モデルを用いて訂正を行っ

た．しかし，CNN は BERT とのネットワーク構造が違うため，BERT を事前学習ニューラルモデルとして利用することには向いていない．

2.2 BERT を利用した英語文法誤り訂正

BEA 2019 [2] では，複数のチームが BERT [4] を要素技術として英語文法誤り訂正を行った．Asano ら [1] は BERT を文レベルの誤り検出モデルとして用いた．Kaneko ら [6] は学習者コーパスで BERT を微調整し BERT の出力スコアを reranking の素性として利用した．BERT を素性に入れることによって $F_{0.5}$ が約 0.7 ポイント向上した．Kantor ら [7] は BERT を入力単語の正しさを調べるための言語モデルとして用いた．この手法では入力文の単語をマスクし，BERT でマスクされた単語を予測させ，予測確率が設定された閾値を超えた場合，訂正対象として使っている．BERT を使うことで，$F_{0.5}$ が 0.27 ポイント向上した．これらの研究によって，文法誤り訂正タスクにおいて BERT に基づく事前学習モデルの有用性が示されている．

3 事前学習モデルを用いた中国語文法誤り訂正

3.1 中国語事前学習モデル

本研究で使用した事前学習モデルは BERT に基づいたモデルである．また，英語と異なり，中国語の事前学習データは文字単位で分割されたものである．Meng ら [10] は，中国語でのニューラルモデルへの入力単位として文字単位が優れていることを報告した．そのため，中国語の入力単位は文字単位であることが多い．

BERT は Transformer Encoder を用い，文脈の知識を把握するために Masked Language Model と Next Sentence Prediction のタスクで学習を行うモデルである．Masked Language Model とは，文の一部のトークンをマスクトークン（[MASK]）に置換し，置換されたトークンをモデルに予測させる仕組みである．Next Sentence Prediction とは，モデルの入力が文 A と文 B の 2 文であり，モデルに文 B が文 A の後文であるかどうかを予測させる仕組みである．

Cui ら [3] は BERT に基づいて，複数の中国語事前学習モデルを訓練し公開している[1]．本研究ではその中の Chinese-RoBERTa-wwm-ext を使用した．

表 1: Chinese-RoBERTa の whole word masking の例

[Original Sentence] 然后 准备 别 的 材料 。
[Original BERT] 然 后 准 [MASK] 别 的 [MASK] 料 。
[Whole Word Masking] 然 后 [MASK] [MASK] 别 的 [MASK] [MASK] 。

Chinese-RoBERTa-wwm-ext と BERT の主な違いは以下の通りである：

wwm (whole word masking)　BERT を提案した Delvin らはサブワード化された単語に対してサブワードごとにマスクするのではなく，単語全体をマスクする wwm という新たなマスク手法を提案した [2]．中国語を処理対象とする場合，トークンをマスクする際に，文字ごとにマスクを行うのではなく，単語全体でマスクを行う．表 1 は例を示している．これによって，モデルが単語単位の知識を学習できることが期待されている．Cui ら [3] はこの手法を用い，中国語事前学習モデルを訓練した．

学習手法　Liu ら [9] の手法に従って学習を行った．Liu らの研究では，Next Sentence Prediction を行わないことでモデルの性能が向上することを報告した．

訓練データ　Chinese Wikipedia（0.4B tokens）に加え，extended data（5.0B tokens）も使用した．extended data は baike（中国語百科事典）や QA（中国語版の知恵袋）より構成されるが，著作権上公開はされていない．

3.2 訂正モデル

本研究では Transformer を訂正モデルとする．Transformer は機械翻訳のような Sequence-to-Sequence タスクにおいて優れた性能を示しており，最近の研究に多く採用されている．

しかし，BERT 系の事前学習モデルは Transformer の Encoder 側だけ使用しており，文法誤り訂正をはじめとした Encoder と Decoder 両方用意する必要がある Sequence-to-Sequence タスクには直接適用できない．そのため，本研究では Chinese-RoBERTa-wwm-ext で学習されたパラメータで Transformer の Encoder を初期化し，Decoder をランダムに初期化した上で，中

[1] https://github.com/ymcui/Chinese-BERT-wwm

[2] https://github.com/google-research/bert

国語文法誤り訂正データで微調整を行うことで中国語文法誤り訂正器を構築した．

4 実験

4.1 実験設定

データ設定 本論文では NLPCC 2018 文法誤り訂正タスクのデータを使用した．データを文字単位で分割すると，文長が長くなり計算資源をより多く占有する．また，文法誤り訂正のタスクでは一般的に誤り文と訂正文が大きく変わらないという特徴がある．そのため，学習データに対して以下のデータフィルタリングを行った：①誤り側と訂正側の文が同じ文．②誤り側と訂正側の文の編集距離が 15 文字以上．③誤り側または訂正側の文の文字数が 64 以上．これらの条件を満たす文対は学習データから除外し，最終的に利用した文対は 971,318 文対である．

テストデータは PKU Chinese Learner Corpus から抽出された 2,000 文である．

モデル設定 本研究では fairseq 0.8.0[3]の Transformer モデルを用いて実験を行った．事前学習モデルの読み込みには pytorch_transformer 2.2.0[4]を使った．事前学習モデルで Encoder を初期化するコードは torshie の実装[5]を元に自ら実装した．単語分割には単語分割ツール Jieba[6]を使用している．

Transformer をベースラインモデルとし，以下のモデルを学習した．

BPE (random init.)：データは単語単位で分割した上で BPE をかけ，モデルのパラメータはランダムに初期化した．

BPE (pre-trained wang2vec)：データは単語単位で分割した上で BPE をかけ，モデルの Encoder と Decoder は wang2vec [8] で初期化した．

char (random init.)：データは文字単位で分割し，モデルのパラメータはランダムに初期化した．

char (pre-trained RoBERTa)：データは文字単位で分割し，モデルの Encoder は Chinese-RoBERTa-wwm-ext で初期化し，Decoder はランダムに初期化した．

パラメータに関しては，学習率：0.00003（pre-trained，4-ensemble），0.00001（その他），最大エポック数：20，最適化手法：Adam，layer 数：Encoder（12-layer），Decoder（12-layer），バッチサイズ：32，語彙サイズ：37,000（BPE），21,168（pre-trained），制限なし（random）と設定した．

[3]https://github.com/pytorch/fairseq
[4]https://github.com/huggingface/transformers
[5]https://github.com/torshie/bert-nmt
[6]https://github.com/fxsjy/jieba

表 2: 中国語文法誤り訂正の実験結果

[Our Model]	適合率	再現率	$\mathbf{F_{0.5}}$
BPE (random init.)	21.57	12.18	18.69
BPE (pre-trained wang2vec)	20.19	14.91	18.85
char (random init.)	25.14	14.34	21.85
char (pre-trained RoBERTa)	32.88	21.24	29.63
4-ensemble	41.94	**22.02**	**35.51**
[NLPCC 2018]			
Fu et al. [5]	35.24	18.64	29.91
Zhou et al. [13]	41.00	13.75	29.36
Ren et al. [11]	41.73	13.08	29.02
Ren et al. (4-ensemble) [11]	**47.63**	12.56	30.57

評価 評価は単語単位で行われるので，NLPCC 2018 文法誤り訂正タスクの設定に従って，システム出力文を区切りなしの 1 文とし，pkunlp[7]で改めて単語区切りを行い評価した．

評価指標は *MaxMatch* (M2) [8] を使用した．誤り文に対し，フレーズレベルのシステムの編集と正解編集を比べ，一致する編集の数で適合率と再現率を計算し，$F_{0.5}$ を最終的なスコアとして評価する．

4.2 実験結果

実験結果を表 2 に示す．Chinese-RoBERTa-wwm-ext で初期化するシングルモデル（char (pre-trained RoBERTa)）の精度はベースライン（BPE (pre-trained wang2vec)，char (random init.)）と比較して飛躍的に向上しており，事前学習モデルの有効性を示している．NLPCC 2018 文法誤り訂正タスクの参加チームと比較しても，同等の結果を示している．

4-ensemble モデルでは、シングルモデルと比較して $F_{0.5}$ が 6 ポイント近く向上し NLPCC2018 のベストチームと比較しても 5 ポイント近く向上した．

また，本手法では適合率が低下する傾向にあるが，再現率が他のモデルと比べ明らかに向上しており，4-ensemble を取ることにより適合率も他のモデルと同等のレベルを示している．

[7]http://59.108.48.12/lcwm/pkunlp/downloads/libgrass-ui.tar.gz
[8]https://github.com/nusnlp/m2scorer

5 分析

モデルの出力例文を表 3 に示す．RoBERTa を用いた事前学習モデルは大規模なコーパスで学習し豊富な文脈知識を持っているため，それを利用した提案モデルの出力はベースラインモデルと比べ流暢になった．

1 番目の例では，書き間違いである「持別」を正しく「特別」に訂正し流暢な文にした．その理由は単語単位でマスクすることで周辺文脈からの予測により補完されているためだと考えられる．

さらに，提案モデルの訂正箇所は参照文と違っているが，誤り文を大胆に訂正することで，出力がより流暢な文になる場合もある．2 番目の例では，「经验」（経験）を「经历」（経歴）に訂正するだけでなく，元の文が「人々は一生にさまざまなことを経験できる」と書かれており，「できる」の意味に相当する「会」を加えることより，ニュアンスとしてより流暢になる．これは事前学習モデルが言語情報を利用することにより流暢な出力が可能になったためだと考えられる．

表 3: モデルの出力例文

モデル	例文
src	持 别 是 北京 ， 没有 “ 自然 ” 的 感觉 。
gold	特别 是 北京 ， 没有 “ 自然 ” 的 感觉 。
BPE (pre)	持 别 是 北京 ， 没有 “ 自然 ” 的 感觉 。
char (rand)	持 别 是 北京 ， 没有 “ 自然 ” 的 感觉 。
char (pre)	特别 是 北京 ， 没有 “ 自然 ” 的 感觉。
4-ensemble	特别 是 北京 ， 没有 “ 自然 ” 的 感觉 。
和訳	特に北京では，「自然」の感覚がない．
src	人们 在 一 辈子 经验 很多 事情 。
gold	人们 在 一 辈子 经历 很多 事情 。
BPE (pre)	人们 在 一生 中 经历 了 很多 事情 。
char (rand)	人们 在 一辈子 经历 了 很多 事情 。
char (pre)	人们 一辈子 会 经历 很多 事情 。
4-ensemble	人们 一辈子 会 经历 很多 事情 。
和訳	人々は一生にさまざまなことを経験できる．

6 おわりに

本研究では，中国語事前学習モデルで訂正モデルの Encoder を初期化し，中国語文法誤り訂正コーパスで微調整した訂正モデルで中国語文法誤り訂正を行い，BERT に基づく事前学習モデルの有用性を示した．また，NLPCC 2018 文法誤り訂正タスクにおいてシンプルなモデルで最高精度の実験結果を示した．NLPCC 2018 文法誤り訂正タスクの手法は基本的に英語文法誤り訂正の研究手法に基づいたものが多いが，中国語の漢字誤りは主に漢字の形と発音の類似性より生じるものなので，今後の研究としてそれらの特徴に着目した中国語文法誤り訂正システムを研究したい．

参考文献

[1] Hiroki Asano, Masato Mita, Tomoya Mizumoto, and Jun Suzuki. The AIP-Tohoku system at the BEA-2019 shared task. In *BEA@ACL*, 2019.

[2] Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. The BEA-2019 shared task on grammatical error correction. In *BEA@ACL*, 2019.

[3] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. Pre-training with whole word masking for Chinese BERT. *ArXiv*, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[5] Kai Fu, Jun Huang, and Yitao Duan. Youdao's winning solution to the NLPCC-2018 task 2 challenge: A neural machine translation approach to Chinese grammatical error correction. In *NLPCC*, 2018.

[6] Masahiro Kaneko, Kengo Hotate, Satoru Katsumata, and Mamoru Komachi. TMU transformer system using BERT for re-ranking at BEA 2019 grammatical error correction on restricted track. In *BEA@ACL*, 2019.

[7] Yoav Kantor, Yoav Katz, Leshem Choshen, Edo Cohen-Karlik, Naftali Liberman, Assaf Toledo, Amir Menczel, and Noam Slonim. Learning to combine grammatical error corrections. In *BEA@ACL*, 2019.

[8] Wang Ling, Chris Dyer, Alan W. Black, and Isabel Trancoso. Two/too simple adaptations of word2vec for syntax problems. In *NAACL-HLT*, 2015.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, 2019.

[10] Yuxian Meng, Xiaoya Li, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. Is word segmentation necessary for deep learning of Chinese representations? In *ACL*, 2019.

[11] Hongkai Ren, Liner Yang, and Endong Xun. A sequence to sequence learning for Chinese grammatical error correction. In *NLPCC*, 2018.

[12] Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *NLPCC*, 2018.

[13] Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. Chinese grammatical error correction using statistical and neural models. In *NLPCC*, 2018.