26. 言語モデルと択一問題による英文前置詞の推定

［機械・電機システム工学専攻］ 宮下 塁
［指 導 教 員］ 小田 幹雄 教授

1. 緒言

第二言語の語学学習において、しばしば文法上の間違いや単語の間違いが見受けられる。文中における間違いを人手ではなくコンピュータを用いることで誤り訂正することができれば、英語学習の向上を図ることができる。

本研究では、英文における前置詞を誤り訂正の対象とし、文中の正しい前置詞を正確に推定することを目的とする。言語モデルには汎化能力の高い BERT 及びその改良モデルである XLNet、RoBerta を用いる。また、Swag タスクの四肢択一問題形式で前置詞全てを用いた場合と使用頻度が高い前置詞のみを用いた場合における精度の比較検討を行う。モデルの評価指標には、正答率(accuracy)を用いる。

2. 従来手法

2-1. Transformer

Transformer は双方向的モデルであり、主に入力用データを読み込む Encoder と出力用データを推定する Decoder、及びこれら二つの重みづけ学習に用いる Attention の 3 つから構成されている。Encoder と Decoder は 6 個の同一層で構成されており、各層は 512 次元の出力層から成っている。また、どちらも Feed Forward Network と Multi-Head Attention を用いている。位置別の Feed Forward Network(FFN)は(1)式で表すことができる。

$$FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2 \qquad (1)$$

それぞれ x は単語の位置,W は重みづけ行列,b はバイアスを示している。また、Transformer は文中の単語の語順情報を必要とするため、入力の埋め込み行列に、位置エンコーディング行列 PE を要素ごとに加算する。位置エンコーディングの行列 PE の各成分は(2)式で表される。

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right) \qquad (2)$$

ここで pos は単語の位置を示しており、d_{model} = 512次元の出力層とする[1]

2-2.Bert

Bert は Transformer の Encoder 層を多段接続した構造になっている。また、Bert は 2 段階学習を行っており、事前学習(Pre-training)でラベル付けをしていない膨大なデータを学習しておき、事後学習(Fine-Tuning)で各タスクに合わせた転移学習を行うことでモデルの精度向上を実現している[2]事前学習では、文章中のランダムに 15%の単語を[MASK]トークンとし、単語を推定する穴埋め問題(Masked Language Model):MLM や 2 つの文章の関連性を推定する問題(Next Sentence Prediction):NSP について学習を行う。また、大きな特徴として事前学習を用いるため、事後学習用の用意するデータセットは小さくても精度が出ることやデータの前処理が不要であることが挙げられる。Bert には事前学習におけるパラメータチューニングを改良した RoBerta や XLNet、AlBert といった改良モデルが多く存在し、これらのモデルはモデル評価タスクである GLUE タスクで高精度な結果を示している。

2-3. Swag

動画キャプションから取得した連続している 2 文のうち、前文を入力データとして、後文を 4 つの選択肢から推定する NSP 問題の一つが Swag タスクである。一般的には選択肢の候補は言語モデルで作成し、人間の常識によるデータの偏りを防止している[3]

3. 提案手法

従来の Swag タスクでは、事前学習として Bert や XLNet の Pre-trained Model を用いる。事後学習には、言語モデルで作成した全く内容の異なる動詞句以降の 4 つの選択肢から[MASK]トークンの推定を行う。表 1 に従来手法の学習データの例を示す。

Sentence 1	Sentence2
He holds up his arms in a specific manner.	He then [MASK]
[MASK]： 選択肢 4 つ	
0: walks outside and starts talking.	
1: backs out and adjusts various items to the case.	
2: sits on the ground and lifts his legs.	
3: talks and returns to the footage of him preparing.	

表 1 従来手法の Swag タスク

本研究では、誤り訂正の対象を前置詞として、文中における前置詞がより正確かどうかを推定することを目的としている。従って、従来手法同様に事前学習では既存の Pre-trained Model を用いるが、事後学習には前置詞のみが異なる文章を事前に用意し、選択肢として用いる。また、前置詞の推定を行う場合、前後の文の文脈は大きな関係性を持たないため Sentence1 と Sentence2 には同じ文章を用いる。表 2 に提案手法の学習データの例を示す。

Sentence 1	Sentence2
we stayed in zacatecas [MASK] a week we really had a good time .	we stayed in Zacatecas [MASK]
[MASK]： 選択肢 4 つ	
0: about a week we really had a good time .	
1: for a week we really had a good time .	
2: over a week we really had a good time .	
3: in a week we really had a good time .	

表 2 提案手法の Swag タスク

4. 実験方法

4-1. 実験手法

実験は Bert、XLNet、RoBerta の 3 つの事前学習モデルを用いて行う。また、事前に Bert の CoLA タスクで各文における各前置詞の出現確率を推定したうえで、正解の前置詞を含む選択肢 4 つの構成から成る基礎実験とする。

実験 1 では、出現確率を推定した前置詞 72 個を用い、

特に文に最も当てはまる可能性が高い前置詞 4 つを用いる。正解である前置詞がそのうちの 4 つに含まれてある場合はそのまま用い、含まれていない場合は当てはまる可能性がもっとも低い 1 つを正解の前置詞と入れ替えることで、正解の前置詞 1 つと不正解の前置詞 3 つの選択肢を構成する。

実験 2 では、ANC[4]を元に、前置詞の中で使用頻度が高い of, to, in, for の 4 つのみを用いる。実験 1 同様に、正解が上記 4 つに含まれていない場合は、正解の前置詞と for を入れ替えることで正解の前置詞 1 つと不正解の前置詞 3 つを構成する。

なお、実験 1,2 では 3717 個の文章から成る同じデータセットを用いており、train データを 3041 個とし、test データを 676 個としている。また、train、test どちらのデータにも正解の選択肢には 0~4 のラベル付けを行っている。

4-2. 学習モデル

学習に用いた事前学習モデルのパラメータを表 3 に示す。また、事後学習におけるハイパーパラメータは表 4 のように定めた。

表 3: 事前学習モデルとパラメータ

	Bert-base-cased	XLNet-Base-cased	RoBerta.base
Layer	12	12	12
Hidden	768	768	768
Heads	12	12	12
Parameters	110M	110M	125M

表 4: ハイパーパラメータ

	Bert	XLNet	RoBerta
系列長	128	128	128
学習率	5×10^{-5}	5×10^{-5}	5×10^{-5}
バッチサイズ	32	12	32
エポック数	10	10	10

表 4 での系列長とは 1 入力文に対する最大の長さを制限するパラメータであり、学習率とは重みパラメータをどれくらい更新するかの幅を表す。学習率は値が小さいほど性格に学習することができるが、それに応じて計算に時間がかかる。また、学習時にはデータをいくつかのサブセットに分けて用いるが、サブセットの大きさをバッチサイズパラメータで指定する。エポック数には、訓練データを学習させる回数を指定する。

5. 実験結果

上記の事前学習モデル及び事後学習におけるハイパーパラメータの設定で実験 1 及び 2 を実施したときの正答率の結果を表 5 に示す。

表 5: 実験結果

Accuracy(%)	Bert	XLNet	RoBerta
実験 1	68.343	69.763	71.450
実験 2	89.941	90.237	89.941

表 5 より、どのモデルを使用した場合でも実験 2 が実験 1 より高い値を示している。これは、学習に用いる前置詞は使用頻度が多いものほど重きを置いたほうが精度が向上すると捉えることができる。また、実験 2 ではどのモデルにも大きな差が見られなかったため、事前学習モデルの差による影響があまりないといえる。実験 1 の結果から Bert と RoBerta では約 3%もの差が出ている。これらから前置詞の候補がより多くなる場合は、事前学習による汎化能力の差が影響してくるといえる。そのため、Albert や T5 といった汎化能力の高い事前学習モデルを用いることで、より正確に前置詞を推定することが可能であると判断できる。

6. 結言

本研究では、Bert 及びその改良モデルの XLNet と RoBerta を用いて、Swag タスクによる四肢択一問題形式で全ての前置詞を用いた場合と使用頻度の高い前置詞のみを用いた場合における前置詞の推定を行い比較検討した。

今後の課題として、正答率を向上させるために、事前学習モデルの汎化能力が高い large-cased に変えることや、Albert や T5 モデルを使うことを視野に入れる必要がある。また、比較検討を行うために本実験では 3 つのモデルのハイパーパラメータは同じようなチューニングにしているが、各モデルで最適な値は異なるため、ハイパーパラメータのチューニングを変えることで、精度の向上を試みる必要がある。

参考文献

[1] “Attention Is All You Need” Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin
https://arxiv.org/abs/1706.03762

[2]” BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
https://arxiv.org/abs/1810.04805

[3] “The BEA-2019Shared Task on Grammatical Error Correction” Christopher Bryant, Mariano Felice, Oistein E.Andersen, Ted Briscoe
https://www.aclweb.org/anthology/W19-4406/

[4]”ANC(American National Corpus) Word frequency data corpus of contemporary American English”
http://www.anc.org/