E-020

ニューラル文法誤り訂正モデルにおける低頻度語処理法の提案

Low frequency word processing method in neural grammatical error correction model

町田 翔 † 延澤 志保 ‡ 荒井 秀一 ‡
Sho Machida Shiho Nobesawa Shuichi Arai

1. まえがき

文法誤り訂正 (Grammatical Error Correction, GEC) は，英語学習者による誤った英作文を自動で訂正するタスクである．近年，Recurrent Neural Network(RNN) を用いた分類器ベースによる GEC タスクの精度向上が報告されている [1, 2]．しかし，これらの手法は decoder の学習が困難になるため，ボキャブラリサイズを制限し，学習者の多くが誤る低頻度語を *unk* として一律に扱ってしまっている．GEC において低頻度語は誤っている語が多く含まれ，英語学習者の誤る可能性が高い語と捉えられるため，訂正すべき対象である．そこで本稿では，RNN を用いた分類器ベースモデルにおける低頻度語学習手法を提案することで，GEC タスクの精度向上を図る．低頻度語の出現頻度を底上げするため，データ拡張と単語を部分文字列に分割する低頻度語処理を行う．

2. 先行研究

Neural Machine Translation(NMT) は近年，RNN を用いた分類器ベースモデルによって精度向上が報告されている [3]．そこで，文法誤り訂正を誤った文から正しい文への翻訳として捉える NMT ベースの手法が提案されている [1]．このモデルを GEC に用いる場合，ボキャブラリサイズが大きいと decoder の学習が困難になるため，ボキャブラリサイズを制限する必要があり，低頻度語を校正することが困難であった．このように，ボキャブラリサイズの制限と training 中に出現していない単語を test 時に予測できないという問題点を解決するために，ボキャブラリをアルファベットで扱うこの方法では RNN を用いた分類器ベースモデルが提案された [4]．しかし，RNN のユニット数が増えてしまい，メモリ数との兼ね合いにより，長い文の学習と予測が困難であった．

3. データ拡張

training コーパスは第二言語学習を支援するサービスである「語学学習 SNS lang-8」の Lang-8 Corpora[1]を用いる．Lang-8 corpara は，英語学習者が作成した文 (source) と英語熟練者が校正した文 (target) で構成されるパラレルコーパスである．ニューラルネットワークモデルは高い表現力の代わりに過学習の恐れがあり，一般的に高い精度を得るためには大量の training データが必要になることが知られている．また我々は，語の出現頻度を底上げすることによって，RNN による学習が困難である低頻度語の異なり語数の削減を図る．NMT

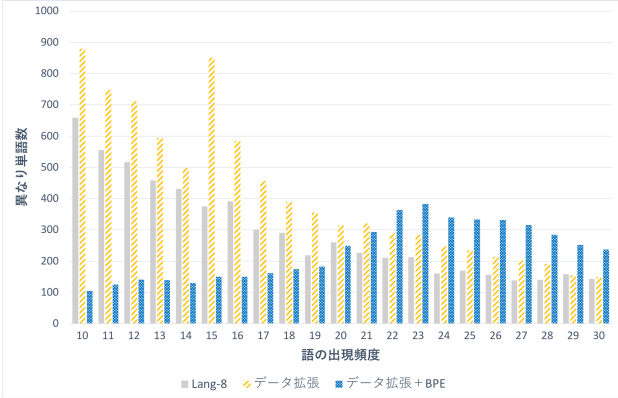

図 1: Lang-8 Corpora とデータ拡張したコーパス，BPE を用いた低頻度語処理後の低頻度語の異なり単語数の比較

タスクでは，training データに約 2M 文のパラレルコーパス用いて学習を行っているが，Lang-8 Corpora では約 0.4M 文しか存在しない．また，GEC において target は，正しい文法で記述されている必要がある．そのため，関沢らの手法 [5] を参考に品詞情報を用いたデータ拡張を行う．本研究では形容詞に着目し，Python の NLTK パッケージ[2]を用いて英文に品詞タグを付与し，形容詞を削除した文を生成した．形容詞は，名詞を修飾する言葉であり，欠けていたとしても文法を崩すことなく意味の通る文になる．Lang-8 Corpora に対してデータ拡張したことによって，生成した文の例を表 1 に示す．すべての形容詞についてこのデータ拡張処理を行った結果，441,187 文から 673,444 文まで training データを拡張することができた．Lang-8 Corpora をデータ拡張したコーパスの低頻度語の変化を図 1 に示す．図 1 から，Lang-8 Corpora の語の出現頻度に比べ，データ拡張を使用した方が全体的に語の出現頻度が底上げされていることがわかる．

表 1: データ拡張の例

org	Sumo is one of the *Japanese traditional* sports .
DA1	Sumo is one of the *traditional* sports .
DA2	Sumo is one of the *Japanese* sports .
DA3	Sumo is one of the sports .

†東京都市大学大学院　工学研究科　情報工学専攻
‡東京都市大学　知識工学部
[1]言語学習 SNS lang-8：http://lang-8.com

[2]NLTK パッケージ：https://www.nltk.org/

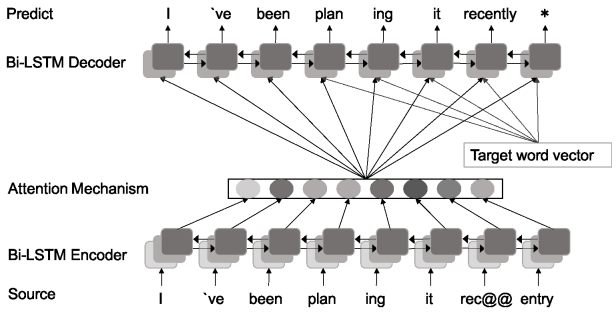

図 2: 提案手法のモデルのアーキテクチャ

4. 単語から部分文字列への変換

低頻度語を部分文字列として扱い，文字列としての出現頻度を上げるため，Byte Pair Encoding(BPE)[6] を用いる．BPE は，例えば low，lower，love，loved の 4 語が低頻度語の場合，共通している「lo」とその他に分割し，分割した文字列の出現頻度が高くなるようにする効果がある．Lang-8 Corpora をデータ拡張したコーパスとデータ拡張したコーパスに対して BPE を使用した場合の異なり単語数の比較を図 1 で示す．Lang-8 Corpora の場合，出現頻度が 1 の単語の割合が 53%を占めていたが，BPE を用いることにより 0.18%にまで抑えることが可能となった．また，BPE によって低頻度語を分割し文字列としての出現頻度を大幅に上げることができた．低頻度語の異なり単語数を減らすことによって，単語ベースでは *unk* とされてしまう低頻度語を学習に含ませることが可能になった．

5. 実験と評価

GEC のコンペティションである CoNLL-2014 の指標には，再現率よりも適合率を重視するため $F_{0.5}$ 値が用いられている [7]．BPE の単語を部分文字列に分割するための学習は，Lang-8 Corpora の source を用いた．ボキャブラリサイズは 1 万，2 万，3 万の 3 つで実験し，$F_{0.5}$ が最も高くなった 2 万を選択した．図 2 のモデルは，source を文字列ごとに encode し，attention mechanism を介して decoder で正解文を文字列ごと予測するアーキテクチャである．モデルは Ziang らのモデル [4] を参考にした．target word vector は予測する文字列の教師であり，attention で得たベクトルと共に decoder に入力し学習する．@タグは BPE による分割を表し，「recentry」が低頻度語であったことを示す．encoder を Bidirectional Long Short Term Memory(Bi-LSTM)[8] の 3 層とし，decoder を 2 層に設定した．CoNLL-2014 shard Task[7] に従って test した．提案手法と従来手法との評価比較の結果を表 2 に示す．低頻度語学習手法を導入前と導入後の評価比較も行った．表 2 に示す通り，従来手法と比較した結果．データ拡張と単語を部分文字列として扱い，低頻度語を学習に含ませることにより，GCE タスクにおいて精度向上したことから，提案手法の有効性が確認できた．．NUCLE[3] コーパスは，5.4K 文のパラレルコーパスである．NUCLE を用いることなく精度向上したため，我々が提案するデータ拡張が効果的だったことがわかった．また，ボキャブラリの制限がない文字ベースの手法 [4] と比べても，精度が良いことがわかった．しかし，現在 $F_{0.5}$ が 42.82point であり，RNN を用いた手法の state-of-the-art である単語ベースと文字ベースが組み合わされた Jianshu ら [2] の結果と 2.33point 離れているため，さらにモデルの調整が必要である．

[3]NUCLE : http://www.comp.nus.edu.sg/ nlp/corpora.html

表 2: 従来手法と提案手法の評価比較

モデル	training データ	$F_{0.5}$
単語ベース (Shamil ら)	Lang-8 + NUCLE	40.56
文字ベース (Ziang ら)	Lang-8	40.58
Hybrid モデル (Jianshu ら)	Lang-8 + NUCLE	45.15
提案手法 (単語ベース)	Lang-8	40.44
提案手法 (BPE)	Lang-8	41.53
提案手法 (BPE + DA)	Lang-8	**42.82**

6. 結論

本稿で提案した低頻度語学習手法により，GEC タスクにおける精度向上を達成した．単語の出現回数を底上げし，文字列として扱うことにより，誤った語が多く含まれる低頻度語を学習に含ませることを可能とした．英語学習者の誤りに形容詞誤りが少ないため，データ拡張に形容詞削除を用いることが効果的であったと考えられる．より精度向上を目指すため，ボキャブラリサイズの検討や，更なるハイパーパラメータの調整が必要である．

参考文献

[1] Chollampatt, S., Taghipour, K. and Ng, H. T.: Neural Network Translation Models for Grammatical Error Correction, *Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 2768–2774 (2016).

[2] Ji, J., Wang, Q., Toutanova, K., Gong, Y., Truong, S. and Gao, J.: A Nested Attention Neural Hybrid Model for Grammatical Error Correction, *CoRR*, Vol. abs/1707.02026, (2017).

[3] Sutskever, I., Vinyals, O. and Le, Q. V.: Sequence to Sequence Learning with Neural Networks, *CoRR*, Vol. abs/1409.3215, (2014).

[4] Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D. and Ng, A. Y.: Neural Language Correction with Character-Based Attention, *CoRR*, Vol. abs/1603.09727, (2016).

[5] 関沢祐樹, 梶原智之, 小町守：目的言語の低頻度語の高頻度語への言い換えによるニューラル機械翻訳の改善, 言語処理学会 第 23 回年次大会 発表論文集, pp. 982–985 (2017).

[6] Gage, P.: A New Algorithm for Data Compression, *C Users J.*, Vol. 12, No. 2, pp. 23–38 (1994).

[7] Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H. and Bryant, C.: The CoNLL-2014 Shared Task on Grammatical Error Correction, *Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14 (2014).

[8] Schuster, M. and Paliwal, K.: Bidirectional Recurrent Neural Networks, *Trans. Sig. Proc.*, Vol. 45, No. 11, pp. 2673–2681 (1997).