日本語文法誤り訂正における事前学習済みモデルを用いたデータ増強

- 概要
  - GECにおけるデータ増強
  - 関連研究
    - Direct Noise
      - 文法的に正しい文に含まれる各トークンに対して，置換(replace)・削除(delete)・挿入(add)・並べ替え(shuffle)の4つの操作をランダムに行うことで文法的な誤りを含む文を生成する手法
    - BackTranslation
      - 文法的に正しい文から文法的な誤りを含む文に変換するためのモデルを訓練し，このモデルに文法的に正しい文を入力することで文法的な誤りを含む文を得る手法
  - 提案手法
    - 日本語で学習されたBERTとDirect Noiseを用いたデータ増強を行うことで日本語GECの性能向上を目指す
    - 既存のDAの研究では、文法的に正しいデータがたくさんある状態を仮定している
      - 少数データに対応したDAを提案する

Algorithm 1 BERT-DA による擬似データ生成アルゴリズム

Input: $\mathcal{Y}^a = \{\boldsymbol{y}_j^a | j = 1, 2, ..., m\}$, S, L

Output: $\mathcal{D}^{a'} = \{(\boldsymbol{y}_k^{a'}, \boldsymbol{x}_k^{a'}) | k = 1, 2, ..., S \times L\}$

$\mathcal{Y}^a$ から S 対サンプリングしたものを $\mathcal{Y}^{a'} = \{\boldsymbol{y}_k^a | k = 1, 2, ..., S\}$ とする

for $k = 1, ..., S$ **do**

generate $\mathcal{Y}^{a'}$

$\boldsymbol{y}_k^a$ の中からランダムにマスクするトークン Z を指定

指定した Z を BERT の Masked LM で L 番目までのトークンに置換したものをそれぞれ $\boldsymbol{y}_k^{a'(1)}, \boldsymbol{y}_k^{a'(2)}, \ldots, \boldsymbol{y}_k^{a'(L)}$ とする

generate $\mathcal{X}^{a'}$

生成された $\boldsymbol{y}_k^{a'(1)}, \boldsymbol{y}_k^{a'(2)}, \ldots, \boldsymbol{y}_k^{a'(L)}$ それぞれに対して Direct Noise を適用し $\boldsymbol{x}_k^{a'(1)}, \boldsymbol{x}_k^{a'(2)}, \ldots, \boldsymbol{x}_k^{a'(L)}$ とする

end for

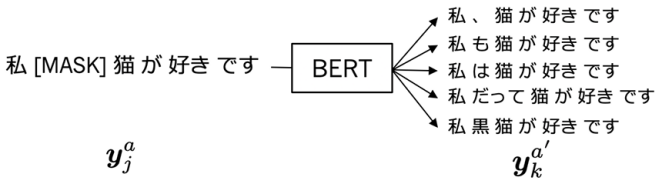

図 4 BERT を用いた $\mathcal{Y}^{a'}$ の生成

(1) 豊富な語彙，多様な言い回しによりドメインや生成元データが限られた状況での性能向上が期待できる．

(2) BERT は左右両方向の文脈を考慮できるため，生成されたテキストは文法的・意味的な整合性をもつことが期待できる．

(3) L の数を指定することで，学習に利用できるデータ量を指定して増やすことができる．

- モデル
  - TransformerCopy
- データセット
  - TEC_JL
- 性能評価
  - GLEU+
- 結果
  - 出版書籍ドメインのデータ(PB)を使用するのが良い
  - 擬似データはそれほど多く作らない方が良い