Vol.2021-NL-251 No.27
Vol.2021-SLP-139 No.27
2021/12/3

BERTのMaskしないMasked Language Modelによる Data Augmentation

有田智也 [1,a)]　松井くにお [1,b)]

概要：自然言語処理では簡単な Data Augmentation の手法として，文中の名詞や動詞を類義語で置き換える方法が用いられることがある．類義語取得の方法としては BERT などの事前学習モデルでの学習時に利用されている，Masked Language Model を利用する手法がある．しかし Masked Language Model により予測される語は類義語ではなく，文脈的な自然な語が予測されるため，類義語置換の方法として適切でない．そのため本研究では BERT に入力文を Mask せずに入力し，類義語を予測する手法を提案する．実験では国会予算委員会での議員の発言に，要約として適当な発言か適切な発言でないかの 2 つのラベルを付与したデータに対して，WordNet，fastText，BERT，ELECTRA で Augmentation を行い，ラベルの予測精度を比較した．実験の結果 WordNet で Augmentation した場合，最も正解率が高かった．また BERT に Mask せずに文章を入力し類義語を予測した場合，Mask した場合に比べて 2.3%正解率が向上した．

キーワード：BERT，Data Augmentation，抽出型要約，ニューラルネットワーク

1. はじめに

日本の選挙では，候補者や政党は自らの政策や方向性を明記したマニフェストを，ネットやメディアを通じて有権者に示し，投票する際の参考にしてもらう形式が取られている．しかしマニフェストには法的拘束力がないため，消費税増税などの賛否が分かれる政策をマニフェストに記載せず，実権を握った後に不意打ち的に強行できる．また実現する気がない政策を有権者の関心を集めるために掲げることも可能である．そのため投票する際にはマニフェストだけを参考にするのではなく，議員の普段の発言を観察し，発言内容とマニフェストの内容に齟齬が無いかを確認する必要がある．しかし国会会議録は文章量が多く，普段の議員の発言すべてに目を通すことは困難なため，適切な長さに文章を要約するシステムが必要である．

そこで本研究では議員の発言を学習させ，抽出型要約で要約することを目指す．今回作成するニューラルベースの抽出型要約では教師データとして，発言に対して要約文として適切かのラベル付けがされたデータが必要である．しかしラベル付けされたデータは公開されていないため，データを新規に自作する必要がある．データを自作する場合は時間やコストの関係で，大量のデータを用意するのが難しい場合がある．学習データが少量の場合モデルが過学習することを防ぐために，学習データに対して何かしらの処理を加え，元のデータと似た新規のデータを作成する手法である，Data Augmentation が用いられている．そこで国会会議録の議員の発言に，要約として適切かのラベルを付与した少量のデータに対して，WordNet，fastText，BERT，ELECTRA を利用して Augmentation を行い，分類精度を評価した．

2. 関連研究

2.1 Data Augmentation

Wei ら[1]は EDA(Easy Data Augmentation)という言語モデルを利用しない，簡単な Data Augmentation 手法を提案した．EDA では訓練データの文章に対して，ストップワード以外のランダムな単語を類義語で置換する，文中の単語の類義語をランダムな位置に挿入する，ランダムに選択した二つの単語の位置を入れ替える，ランダムに選択した単語を削除する，の 4 つの操作を文長に応じた回数実行することにより Augmentation を行っている．Kumar ら[2]は BERT などの事前学習モデルを利用して Augmentation を行う手法を提案した．具体的には一部の単語をランダムに[MASK]トークンに置換した文章を事前学習モデルに入力し，置換前の単語を予測することで Augmentation を行っている．

2.2 事前学習モデル

fastTextはBojanowskiら[3]が提案したWord2Vecを改善したモデルであり，単語をサブワード単位で分割したデータを学習している．そのため活用形によって変化しない基幹部分の埋め込み表現を共有できるようになり，未知語に対しても適切な単語ベクトルを与えやすくなっている．またWord2Vecに比べて，高精度かつ高速で単語ベクトルを抽出できる．BERTはJacob Devlinら[4]が提案した事前学習モデ

1　金沢工業大学
Kanazawa Institute of Technology
a)　c6000578(@)planet.kanazawa-it.ac.jp
b)　kmatsui(@)neptune.kanazawa-it.ac.jp

ⓒ 2021 Information Processing Society of Japan

ルであり，複数の双方向Transformerエンコーダを積み重ねることによって構成されている. BERTはあらゆる自然言語処理のタスクにFine-tuningすることが可能なモデルになっている. BERTに文を入力する際には単語列の始まりを表す[CLS]トークンを付与し, 複数文を入力する場合には文と文の間に区切りを表す[SEP]トークンを付与する．BERTは事前学習としてMasked Language Model(以下MLM)とNext Sentence Prediction(以下NSP)の二つのタスクを解くことで学習している．MLMでは入力文から[CLS]，[SEP]を除いた単語を一定の割合で[MASK]トークンに置換した文をBERTに入力し，[MASK]トークンに置き換えられる前の単語を予測する．NSPでは[SEP]トークンで接続された2文が連続しているかを, [CLS]トークンの出力をもとに予測する. 前述した2つのタスクを解くことでBERTは双方向から文章を捉え，文脈に応じた単語埋め込み表現を出力することができる．Clarkら[5]はBERTの事前学習の問題を指摘し，MLMを改善したELECTRAを提案した．MLMでは入力文の一部をmaskし，置き換えられる前の単語を予測している．しかしmaskされていない単語に関しては，入力をそのまま出力することで元単語を予測できるため，入力文の一部しか学習に利用することが出来ない．ELECTRAではReplaced Token Detection(以下RTD)という手法を採用することで入力文全体から学習できるように改善した．RTDではGeneratorとDiscriminatorを利用し，Generatorではmaskした単語を穴埋めすることで，入力文に似た文章を出力できるように学習し，DiscriminatorではGeneratorから生成された文章が元の単語と同じ単語かを予測することで，入力文章全体から学習できるように改善している．Generatorは学習の時のみ利用され，Fine-tuningする際にはDiscriminatorのみを利用する．

3. 提案手法

従来の BERT を用いた Data Augmentation 手法ではまず Tokenizer を利用して入力文を単語列に分割する．次に単語列中の動詞，名詞からランダムに選択した単語を[MASK]トークンに置換し，単語と ID の対応をもとに ID 列に変換する．次に ID 列を BERT に入力することで，モデルの出力として mask される前の単語の出現確率が得られる．出現確率が高い単語を複数抽出することで類義語群を取得し，得られた類義語群からランダムに選択した類義語で[MASK]トークンを置換することによって新しい文を作成する．

本研究で提案する unmask BERT による Data Augmentation 手法を図 1 に示す．提案手法ではまず既存手法と同様に Tokenizer を利用して，入力文を単語列に分割し，単語と ID の対応をもとに ID 列に変換する．次に入力を mask せずにそのまま BERT に入力し，名詞，動詞の中からランダムに選択した類義語候補における, 単語の出現確率を取得する．得られた出現確率から確率が高い単語を複数抽出することで類義語群を取得し，類義語群からランダムに取り出した単語で，類義語取得対象の単語を置換することで新しい文を作成する．

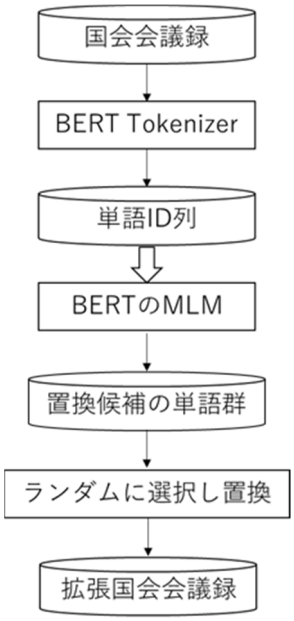

図 1　unmaskBERT での類義語置換

4. 文章要約実験

4.1 使用するデータセット

本研究では国立国会図書館が公開している国会会議録データ[6]の内 2020 年度の予算委員会会議録から，議員への質問の回答文 40 件を抽出し，要約に適切な文に 1，適切でない文に 0 のラベルを付与したデータを作成した．ここで作成した 40 件の学習データで学習した結果がベースラインになる．WordNet，fastText を利用した Augmentation では各文から名詞，動詞をランダムに抽出し，抽出された単語が名詞の場合は処理を行わず，動詞の場合は原形に変形し，それぞれのモデルに入力することで類義語候補群を取得した．得られた類義語候補群の中から BERT の語彙に存在する単語を選択して置換した文を，拡張国会会議録に追加する．ELECTRA を利用した Augmentation では入力文中の名詞，動詞をランダムに[MASK]トークンに置き換えた文を ELECTRA の Generator に入力し，mask される前の単語を生成し，類義語候補群を取得する．得られた類義語候補群の中から BERT の語彙に存在する単語を選択して置換した文を，拡張国会会議録に追加する．

各文に対してそれぞれ 5 箇所を 3 つの類義語で置換しているため，Augmentation 後の文数は 640 件になっている．

4.2 実験設定

本研究で行うタスクは議員の発言が要約として適切かのラベルを付与したデータを，WordNet，fastText，BERT，

ⓒ 2021 Information Processing Society of Japan

unmask BERT，ELECTRA を利用して拡張したデータで行う 2 値分類である．

日本語 WordNet は国立研究開発法人情報通信研究機構が公開しているモデル[7]を利用した, fastText は@Hironsan が公開しているモデル[8]を利用した．BERT の事前学習済みモデルは東北大学が公開している BERT のモデル(cl-tohoku/bert-basejapanese-whole-word-masking)[9]を利用した．ELECTRA の Generator と Discriminator は Cinamon 社が公開している ELECTRA の Generator と Discriminator[10], [11]を利用した．

分類のために構築したモデルを図 2 に示す．ELECTRA Discriminator は 512 単語以下の ID 列を受け取り各単語に対応する 256 次元のベクトルを出力する．各単語のベクトルの中から[CLS]トークンに対応する出力を文ベクトルとして全結合層に入力し，全結合の重みと ELECTRA Discriminator の最終層の重みを学習させた．

学習のための利用したプログラムの設定はバッチサイズ 32，最適化手法に Adam，損失関数に Cross Entropy Loss を利用した．ELECTRA の最終層の学習率は 5e-5，全結合層の学習率は 1e-4 としてモデルを学習させた．実験のために作成したプログラムではモデルの全結合層の重みが毎回ランダムな値で初期化されるため，初期化された値によって分類精度が異なる結果になる．そのため同じ訓練データ，学習プログラムで 3 回実験を行い，検証データでの正解率が最も高かったエポックでの，テストデータでの正解率を実験結果として利用した．

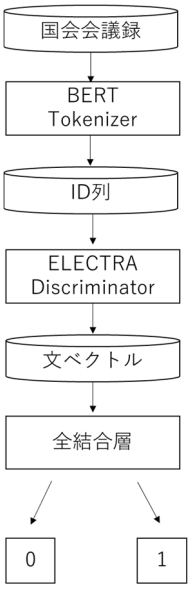

図 2　分類モデルの概要図

4.3 実験結果と評価

実験の結果得られた，各 Augmentation 手法による分類の正解率を表 1 に示す．実験の結果，unmask BERT を用いることで, BERT への入力を mask した場合に比べて正解率が 2.3%向上した．また WordNet による類義語置換が最も正解率が向上し，ベースラインに比べて 9.8%向上した．

Augmentation 手法の違いにより，取得される類義語を確認するために“全国でこの適格消費者団体がどんどん増えて、各様々なブロックでそれぞれ活動しております。”という文章の“増え”から各手法によって得られた同義語を表 2 に示す．表から unmask BERT を利用した場合，BERT を利用した場合に比べて，感覚的に類義語として自然な単語が抽出されていることがわかる．そのため文の意味の一貫性を保ちつつもバリエーションのある表現ができ，精度が向上したと考えられる．また WordNet から得られた類義語は，活用形を変化させた形ではなく，語幹の部分が変化しているため，他の Augmentation 手法よりバリエーションに富んだ表現が得られている．

表 1　各 Augmentation 手法による分類の正解率

Augmentation手法	正解率
baseline	0.542
WordNet	**0.640**
fastText	0.634
BERTのMLM	0.608
BERTの非MLM	0.631
ELECTRA	0.600

表 2　各 Augmentation 手法で得られた“増え”の類義語

Augmentation手法	同義語1	同義語2	同義語3
WordNet	高騰	増加	上がる
fastText	減る	増え	増えやす
BERTのMLM	し	っ	分かれ
unmaskBERT	増え	増やし	広がっ
ELECTRA	増え	続い	離れ

5. 考察

5.1 各モデルから得られた語の可視化

実験結果より，ある単語と意味的に近い単語で置換された場合，効果的な Data Augmentation になると考えられる．そこで各 Augmentation 手法で生成された単語の埋め込み表現を，可視化して図に表した．入力は“全国でこの適格消費者団体がどんどん増えて、各様々なブロックでそれぞれ活動しております。”とし，類義語を抽出する語を“増え”にして各モデルで 5 つずつ類義語を取得した．得られた類義語を ELECTRA の Discriminator に入力し 256 次元の単語ベクトルを抽出する．得られた単語ベクトルを主成分分析に入力し 2 次元に圧縮することで，図示できるように次元を落とした．元単語の単語ベクトルと各手法での Augmentation 後の単語ベクトルを可視化した結果を図 3 に

ⓒ 2021 Information Processing Society of Japan

Vol.2021-NL-251 No.27
Vol.2021-SLP-139 No.27
2021/12/3

示す．また“趣旨がよくわかりませんので”の“わかり”を対象として各手法で Augmentation を行い，元単語の単語ベクトルと Augmentation 後の単語ベクトルを可視化した結果を図 4 に示す．図 3 では同義語置換前の単語は（-5,-4）付近の位置にプロットされており，図 4 では同義語置換前の単語は（-2,1）付近の位置にプロットされている．図 3，図 4 を見ると unmaskBERT から得られた類義語群は，BERT から得られた類義語群に比べて，埋め込み表現が元の単語周辺に固まってプロットされる傾向があり，事前学習モデル的にも，より元の単語と意味的に近い語が得られていることが見て取れる．また WordNet を用いて得られた類義語群は，unmaskBERT から得られた類義語群に比べて元単語から離れた位置にプロットされる傾向があるため，活用形を変化させた同義語より，語幹を変化させた同義語の方がよりロバストなモデルの構築に役立っていると考えられる．

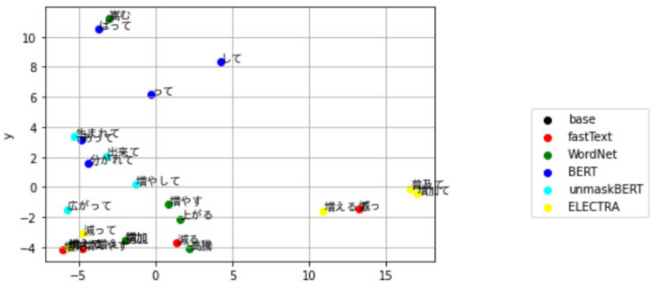

図 3　各 Augmentation 手法で得られた”増え”の類義語の単語ベクトルを PCA で可視化した結果

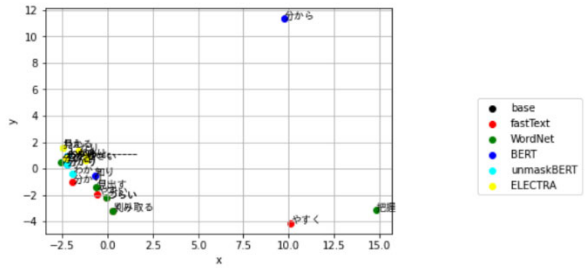

図 4　各 Augmentation 手法で得られた”わかり”の類義語の単語ベクトルを PCA で可視化した結果

6. おわりに

本論文では BERT への入力を mask せずに MLM を行う unmask BERT を提案した. unmask BERT を利用することで，mask した入力した場合に比べて，人間の感覚的により意味が似た語が抽出されることが判明した．また単語の埋め込み表現を可視化したところ，分類精度が BERT を用いた Augmentation より高かった，fastText, WordNet, unmaskBERT から得られた類義語は，同義語置換前の単語周辺にプロットされる傾向があり，Augmentation では元単語と類似した単語で置換することが重要であると判明した．

今後は類義語に置換する候補を選択する手法について検討していきたい．また人手で Augmentation したデータでの精度評価や埋め込み可視化を試み，人の知識が入ることよる影響について考察していきたい．

参考文献

[1] Jason Wei, and Kai Zou． EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks EMNLP-IJCNLP 2019 short paper

[2] Varun Kumar, Ashutosh Choudhary and Eunah Cho．Proceedings of the Second Workshop on Life-long Learning for Spoken Language Systems, pages 18–26 Online, December 7, 2020

[3] Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas． Enriching Word Vectors with Subword Information． arXiv preprint arXiv:1607.04606

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Association for Computational Linguistics, pp4171-4186

[5] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher and D. Mannin． ELECTRA: PRE-TRAINING TEXT ENCODERS AS DISCRIMINATORS RATHER THAN GENERATORS． In ICLR 2020

[6] 国立国会会議録検索システム検索 API https://kokkai.ndl.go.jp/api.html(引用日:2021/10/22)

[7] 日本語 Wordnet． http://compling.hss.ntu.edu.sg/wnja/(引用日:2021/10/22)

[8] vector.zip - Google ドライブ https://drive.google.com/file/d/0ByFQ96A4DgSPNFdleG1GaHcxQzA/view?resourcekey=0-dtRh36dTEOh9RETFJDOQbw(引用日:2021/10/22)

[9] Hugging Face． https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking(引用日:2021/10/22)

[10] Hugging Face． https://huggingface.co/Cinnamon/electra-small- japanese-generator(引用日:2021/10/22)

[11] Hugging Face． https://huggingface.co/Cinnamon/electra-Small-japanese-discriminator(引用日:2021/10/22)

ⓒ 2021 Information Processing Society of Japan