Vol.2020-NL-244 No.4
2020/7/3

BERTの教師無しデータへの適用

築地 毅[1,a)]　鈴木 晴也[1,b)]　柴原 一友[1,c)]　藤本 浩司[1,d)]　池田 龍司[2,e)]　尾﨑 和基[2,f)]
森田 克明[2,g)]　松原 敬信[3,h)]

概要：本稿では，BERT を利用した教師無しデータへの適用について論ずる．近年ディープラーニングの技術が確立し始めており，特に画像認識分野において，既存の技術では困難だった特徴の自動抽出を実現したことにより，非常に高い精度を上げるようになってきている．自然言語処理においてもディープラーニングの研究は広く行われているが，近年 Google により発表された BERT の功績は大きく，教師あり学習のタスクに対して，既存の成果を大きく上回る成果を上げている．本稿では，教師あり学習の精度を大きく高めた BERT を教師無しデータに適用することで，既存手法の性能向上につながる可能性があるという仮説を主張する．本稿では，特許文書を対象に，教師あり学習を行わずに特許の類似性を図る実験を行った．実験の結果，人手で付与した特許分類フラグに対し 61.9 %の正解率となり，BERT を活用することで教師データを与えずとも，特許の類似度を表現できることを示した．

Application of BERT to Unsupervised Data

Tsuyoshi Tsukiji[1,a)]　Haruya Suzuki[1,b)]　Kazutomo Shibahara[1,c)]　Koji Fujimoto[1,d)]
Ryuji Ikeda[2,e)]　Kazuki Ozaki[2,f)]　Katsuaki Morita[2,g)]　Takanobu Matsubara[3,h)]

Abstract: In this paper, we discuss application of documents to unsupervised data using BERT. In recent years, the technology of deep learning had begun to be established, and in the field of image recognition, the automatic extraction of features that was difficult with existing technologies has led to very high accuracy. Deep learning has been widely studied in natural language processing, but in recent years, BERT, by Google, has achieved a great deal of success and has far exceeded the existing achievements for supervised learning tasks. In this paper, we assert that applying BERT, which greatly improves the accuracy of supervised learning, to unsupervised data may lead to better performance than existing methods. We did an experiment on similarity of patents for patent documents without supervised learning. As a result, the accuracy rate was 61.9% for the manually assigned patent classification flag, and it was shown that the similarity of patents could be expressed without using training data by using BERT.

1 テンソル・コンサルティング株式会社
Tensor Consulting Co.Ltd.
2 三菱重工業株式会社 ICT ソリューション本部 EPI 部
EPI Department/ICT Solution Headquarters/Mitsubishi Heavy Industries,Ltd.
3 三菱重工業株式会社 バリューチェーン本部 プロジェクト部
Project Department/Value Chain Headquarters/Mitsubishi Heavy Industries,Ltd.
a) tsuyoshi.tsukiji@tensor.co.jp
b) haruya.suzuki@tensor.co.jp
c) kazutomo.shibahara@tensor.co.jp
d) koji.fujimoto@tensor.co.jp
e) ryuji_ikeda@mhi.co.jp
f) kazuki_ozaki@mhi.co.jp
g) katsuaki_morita@mhi.co.jp
h) takanobu_matsubara@mhi.co.jp

1. はじめに

近年，ディープラーニングの研究が広く行われている．特に画像認識分野において，既存の技術では困難であった特徴の自動抽出を実現したことにより，今までの成果を大きく超えた成功を収めている [2][5]．ディープラーニングの適用範囲は広く，囲碁においてプロに勝利する成果をあげた，AlphaGo の成果は広く知られている [3]．自然言語処理においてもディープラーニングの適用は広く研究されていたが，近年 Google により発表された BERT の功績は極めて大きく，教師あり学習のタスクに対して，既存の成

果を大きく上回る成果を上げている [4]．教師あり学習の精度を大きく高めた BERT を教師なしデータへ適用する方法が確立されれば，既存手法以上の性能につながる可能性があると考えられるため，BERT を教師なしデータへ適用する試みの意義は極めて大きいと筆者らは考える．

そこで本稿では，特許文書をテストデータとして利用し，BERT を教師無しデータへ適用する手法を提案する．まず 2 章で関連研究として BERT について述べ，3 章で本稿で提案する手法について述べる．4 章で評価と考察を行い，5 章でまとめと今後の課題について述べる．

2. 関連研究

本章では，本稿の理解で必要な箇所に絞って，BERT の概略について説明する．なお，ディープラーニングの基本的な知識については岡谷の文献 [7] に，BERT の詳細については Devlin らの論文 [4] や，柴原らの書籍 [8] に譲る．

2.1 BERT の構成

BERT は図 1 のとおり構成されている．図 1 の下部から文章を入力する．入力インターフェースは 768 要素からなるベクトルが 512 個用意されており，文章が一単語ずつ分散表現に変換され，一つのノード（ベクトル）に一つずつ入力される*1．但し，一般的に左端のノードは，[CLS] という特殊な単語が入力される．BERT 内部は複数の transformer[1] から構成され，それぞれ内部計算の後に上部の出力層から，それぞれの入力単語に対応した出力が分散表現のベクトルとして出力される．ファインチューニングで分類問題を解きたいのであれば，左端の C の部分を取り出して，softmax などにより分類を行う．本稿では，出力層の C の部分を [CLS] 部分，[CLS] 部分を除いた部分を，入力した単語に対応した部分であることから，単語対応部分と呼ぶことにする．

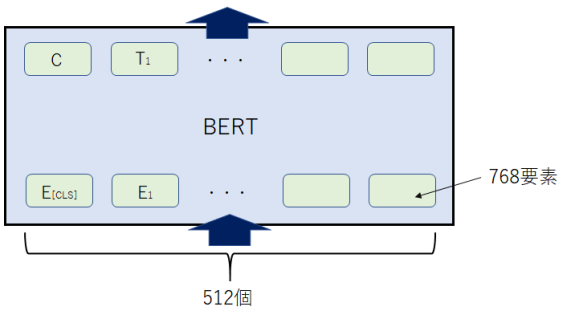

図 1 BERT の構成

*1 日本語の場合，分かち書きを行う必要があるため，何らかの形態素解析ロジックにより，形態素解析を行う必要がある

2.2 BERT の学習

BERT を用いた教師あり学習は，大きく下記の二つのフェーズに分けられて実行される．

(1) 事前学習（穴埋め問題，文章接続問題）の実施
(2) ファインチューニングの実施

まず，事前学習として，大量の文書データによる学習が行われる*2．事前学習では，穴埋め問題と文章接続問題の学習が行われる．BERT に用意されている学習器により，与えられた文書データから自動的に大量の問題と正解が生成され，それぞれ学習が行われる．大量に穴埋め問題と文章接続問題を学習することで，「文章のある位置に当てはめられる単語を理解する」学習と，「ある文章につなげられる文章を理解する」学習を経て，「基本的な文書構造を理解する」ことが期待される．

次にファインチューニングで，解きたいドメインのデータを利用して，解きたい問題を学習する．事前学習で「基本的な文書構造」が学習されているから，一般的に事前学習で利用した文書データよりも少ない件数で，精度の良い学習を行うことができる．一般的に，ある領域で学習したモデルを，別の領域に適用して効率的に学習する手法は転移学習と呼ばれており，BERT は転移学習の考え方から構築されている．

2.3 BERT による教師あり学習の成果

BERT が教師あり学習において効果的であった理由について述べる．既存手法においては，単語の出現頻度であったり，単語ごとの分散表現を入力データとして利用したりすることが一般的であり，単語間の演算が行わない場合がほとんどであった．BERT においても入力データは 2.1 のとおり，単語ごとの分散表現であるが，大量のデータからなる事前学習と transformer の計算により，内部的に単語間の関係が獲得されていると考えられている．その結果得られる出力は，いわば「文章全体を考慮した高度な分散表現」であると考えられており，この高度な情報を含む分散表現が，BERT の精度に寄与していると考えられている．

3. 提案手法

本章では，BERT の文章全体を考慮した高度な分散表現を活用することで，教師あり学習を行うことなく，類似文書を分類する手法を提案する．本稿においては，特殊性の高い単語を多く含む特許文書を事例として利用する．

*2 一般的に事前学習では，本来解きたい問題とは異なる種類の文書データで行われることが一般的である．理由としては解きたい種類と同じ文書データを大量に用意することは困難であるケースが多いためである．文書の種類としては，特許文書，ニュース文書，論文，ブログなどが考えられる．本稿ではこの種類のことをドメインと呼ぶことにする

3.1 発想

本節では，本稿での提案手法の説明に先立って，提案手法の発想について述べる．

一般的に 2.2 で述べた通り，事前学習により「基本的な文書構造」を獲得し，「文章全体を考慮した高度な分散表現」を出力できるようになる．つまり，事前学習された BERT を用いれば，ファインチューニングを行わなくとも「文章全体を考慮した高度な分散表現」を獲得できると考えられる．この「文章全体を考慮した高度な分散表現」であれば，入力データを精緻に表現できるだろうから，入力文書データの類似度を図る指標として利用できるだろう，という発想に基づく．

ところで，事前学習において評価したいドメインのデータがまったく使われないというケースも当然起こりうる．一般的には大量に獲得できる文書データを，事前学習に用いるケースが多い．例えば近年であれば，SNS などから収集することは，実運用上比較的容易であると想像できる．一方，特許文書のような特殊性の高い文書は，実運用上事前学習に含まれないケースが多いと考えられる．しかも，特許文書は特許特有の単語を多く含むため，事前学習に特許文書が含まれないことで、精度に悪影響を及ぼしやすいのではないかと考えられる．

仮に，事前学習において評価したいドメインのデータを利用することができれば，「評価したいドメインのデータに特化した，文章全体を考慮した分散表現」が当然得られるだろう．つまり「評価したいドメインのデータに特化した，文章全体を考慮した分散表現」であれば，評価したいドメインの入力データを，さらに精緻に表現できるだろうと考えられる．

3.2 提案手法

本節では 3.1 の発想に基づき，BERT の分散表現を活用することで，教師あり学習を行うことなく，類似特許を効果的に分類する手法について提案する．3.2.1 では，BERT による分類手法の基礎となる技術を提案し，3.2.2 では事前学習にドメインを考慮することで，より精緻に分類することを目的とした手法について提案する．

3.2.1 分散表現を利用した分類

まず，分散表現を利用した分類方法について提案する．本稿では，下記のとおり BERT の出力部分を利用した，二つの方法を提案する．

- CLS 部分を評価する方法
- 単語対応部分を評価する方法

3.1 で述べた通り，ファインチューニングを行わなくても，充分な事前学習が行われているのであれば，入力データごとに得られる「文章全体を考慮した高度な分散表現」は，入力データの類似度を測る指標として利用できるだろうと考えられ，それを検証によって明らかにしたい．そこで本稿では，ファインチューニングを経ずに，事前学習済みの BERT の出力部分のベクトルを利用して，それらの距離から類似度を評価する方法を提案する．具体的な評価指標については，4.1 で述べる．

[CLS] 部分については，768 要素の一つのベクトルであるから，[CLS] のベクトルをそのまま利用する．単語対応部分については，[CLS] および [SEP] を取り除いた 510 個の 768 要素のベクトルからなるため，本項では式 (1) のとおり 510 個のベクトルの平均ベクトル $\bar{\boldsymbol{w}}$ を利用することにする．ただし，$\boldsymbol{w}_i$ は単語対応部分の i 番目のベクトルを表しているものとする．

$$\bar{\boldsymbol{w}} = \frac{1}{510}\sum_{n=1}^{510} \boldsymbol{w_i} \tag{1}$$

3.2.2 ドメインを考慮した事前学習の活用

本節では，事前学習を行うデータのドメインを，評価時のドメインと同じものにする方法を提案する．3.1 で述べた通り，一般的に BERT では事前学習で利用されるデータのドメインは，必ずしも評価したいデータと同じドメインとは限らない．一般的にはファインチューニングを施すことを前提としているため，ファインチューニング時にドメインに合わせた調整が行われるためである．本稿ではファインチューニングを行わずに，類似特許を評価することを目的としているため，事前学習時からドメインに合わせたデータを利用して学習することで，より効果的に類似度を測ることを提案する．

3.3 システム設計

本稿で提案するシステムのシステム設計を図 2 のとおり定める．本稿では，分かち書きを行う SentencePiece[6] と BERT を事前学習により調整し，調整した SentencePiece と BERT を利用して，特許文書の類似度の評価を行う．評価として入力する文書は請求項文書とする．本システムにおける評価軸としては，BERT の出力層部分においてどの要素を利用すると効果的かという観点が 3.2.1 に対する評価であり，どの事前学習データを利用すると効果的かという観点が 3.2.2 に対する評価である．

4. 評価

本章では，3.2 で提案した提案手法について評価を行い，その有効性について論ずる．

4.1 評価方法

本稿において，本手法の有効性を評価する指標として下記を採用した．

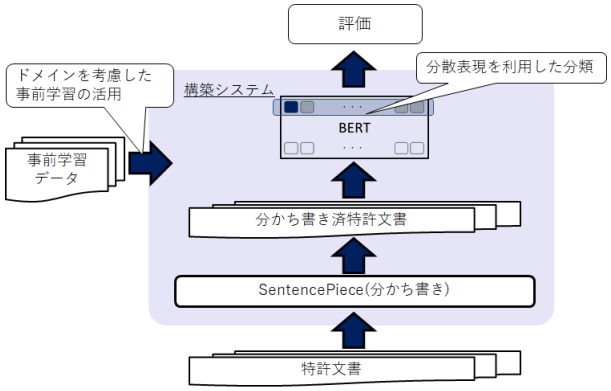

図 2　システム設計

正解データ　人手による特許分類フラグ
評価軸 1　主成分分析による，入力データと正解との関係の可視化
評価軸 2　k 近傍法による，入力データに対する正解率

まず正解データについて述べる．同一ドメインの特許を調査するために用意した複数の検索式による特許公報の検索結果ごとに特許分類フラグを付与したデータセットを正解データとして利用する．従って，一つの特許文書には，複数の特許分類フラグが付与されることがある．本稿では教師無しデータへの適用について評価したいので，あくまでもこの特許分類フラグは本手法の有効性を評価するためだけに利用して，ファインチューニングなどの教師あり学習は行わない．評価で利用した特許及び特許分類フラグの件数は表 1 のとおりである*3．

表 1　特許分類フラグの件数

特許分類フラグ	件数
class1	50
class1,class2,class3,class4	3
class1,class3,class4	1
class1,class4	1
class2	54
class2,class3	124
class2,class3,class4	93
class3	17
class3,class4	11
class4	12
class5	97
class6	65

次に評価方法および評価軸について述べる．まず評価したい事項に応じて事前学習を施した BERT を用意する．BERT の仕組みによれば，特許文書を入力として一つ与えると，3.2.1 で述べたとおりの出力が一つ得られる．入力する特許文書は，表 1 で述べた件数すべてであり，入力した特許文書件数だけの出力が得られる．

そして，主成分分析を利用して，特許文書から得られた全出力を 3 次元に縮約し，可視化して評価する．四角形一つが一つの特許文書を表しており，特許分類フラグごとに色分けがされている．要素が類似特許を正しく表現できているのであれば，同色の四角形が視覚的に固まるようマッピングされることが期待される．

次に k 近傍法を応用して，ある特許がどの特許分類フラグに分類されるかを定量的に評価する．評価対象とする特許以外については特許分類フラグが分かっているという設定とし，評価対象特許がどの特許分類フラグに対応するかを k 近傍法で評価する．k 近傍法によって得られた特許分類フラグと，実際の特許分類フラグが一致していれば正解，一致していなければ不正解と扱う．類似特許を正しく表現できているのであれば，類似特許同士の距離が近くなるため，正解率が高くなることが期待される．

4.2　分散表現を利用した分類の評価

本節では，分散表現を利用した分類の評価を行う．3.2 で述べたとおり，[CLS] のベクトルおよび単語対応部分のベクトルを，人手による特許分類フラグを正解データとして評価する．ただし，事前学習については，日本語 wikipedia で学習された BERT*4 を利用することにする．この日本語 BERT は事前学習として，wikipedia データ約 1,800 万件を利用した学習がすでに行われている．

4.2.1　分散表現を利用した分類：可視化による評価

[CLS] のベクトルを可視化した結果は図 3，単語対応部分のベクトルを可視化した結果は図 4 のとおりである．図 3 によると，同一特許フラグを持つデータが固まっている傾向が特に見られず，類似特許の表現力が弱いことが伺える．一方，図 4 によると同一特許フラグを持つデータ同士が固まっており，類似特許を表現できていることが伺える．以上のことから，単語対応部分のベクトルを利用することで，類似特許を評価できることがわかった．

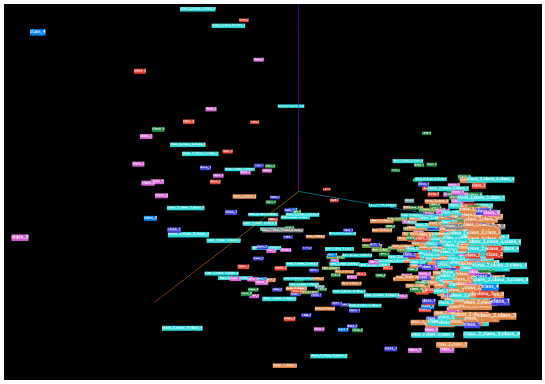

図 3　分散表現を利用した [CLS] の可視化

*3　具体的な内容については社内情報のため伏せる

*4　https://yoheikikuta.github.io/bert-japanese/

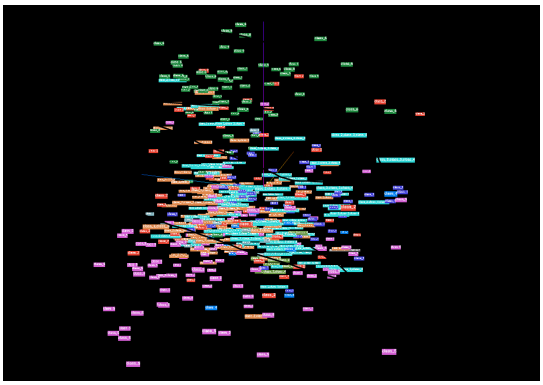

図 4　分散表現を利用した単語対応部分の可視化

4.2.2　分散表現を利用した分類：k 近傍法による評価

k 近傍法により評価した結果は表 2 のとおりである．表 2 によると，[CLS] による正解率が 45.8%，単語対応部分が 61.9%となり，可視化による評価同様に単語対応部分の方が優れた結果になることが分かった．

表 2　[CLS] と単語対応部分の k 近傍法による比較

手法	正解率	正解数
[CLS]	45.8%	242
単語対応部分	61.9%	327

4.2.3　分散表現を利用した分類の考察

以上の二つの評価により，単語対応部分のベクトルを利用してその距離を評価することで，教師あり学習を行うことなく，類似特許を評価できることが分かった．一般的に類似した文書は，類似した単語から構成されている．2.3 で述べた通り，単語対応部分については入力単語の分布を表現しているので，単語対応部分を一種の分散表現としてその距離を評価することで，入力データの類似性を評価することができたと考えられる．

また事前学習において，「文章のある位置に当てはめられる単語を理解する」学習と，「ある文章につなげられる文章を理解する」していることを，2.2 で述べた．詳細は，Devlin らの論文 [4] や，柴原らの書籍 [8] に譲るが，事前学習における「ある文章につなげられる文章を理解する」においては，[CLS] 部分の出力を利用し，ある文章とある文章がつながるかつながらないかの二択問題の学習を行っている．つまり，事前学習だけを施した BERT において [CLS] 部分は，「ある文章につなげられる文章を理解する」ためだけしか調整がされておらず，必ずしも「文章全体を考慮した高度な分散表現」になっているわけでないため，今回のような教師無しデータとして入力データの類似度を測る問題には適さなかったものと考えられる．

4.3　ドメインを考慮した事前学習の活用の評価

本節では，ドメインを考慮した事前学習の活用の評価を行う．本稿では，下記の学習条件で BERT および SentencePiece の事前学習を行う．そこで得られた BERT および SentencePiece を利用して，前述のとおりの評価を行い，その効果について考察を行う．

初期値　平均値 0，標準偏差 0.02 の切断正規分布
学習率　1e-4
学習データ　請求項
学習データ件数　2,637 件の特許データ

まず，上記の学習条件で BERT の事前学習を行った結果は図 5 のとおりであり，十分な学習が行われ，学習は収束していると考える．本稿では，この学習条件で得られた BERT を利用して，評価を行う．但し，4.2.3 のとおり，[CLS] より単語対応部分を利用する方が良いことが判明しているため，今後の評価はすべて単語対応部分を利用することにする．以降，本手法により得られた BERT のことを「特許事前学習 BERT」，前述の日本語 wikipedia データで事前学習された BERT のことを「wiki 事前学習 BERT」と呼ぶことにする．

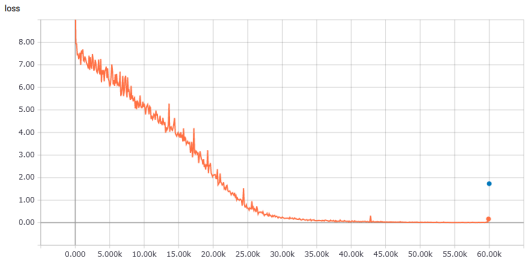

図 5　事前学習の収束曲線

4.3.1　ドメインを考慮した事前学習の活用：可視化による評価

wiki 事前学習 BERT で得られた単語対応部分のベクトル及び，特許事前学習 BERT で得られた単語対応部分のベクトルを可視化した結果は図 6 および図 7 である．なお，wiki 事前学習 BERT の結果は，前述の図 4 と同じものであるが，比較のため再掲する．その結果，特許事前学習 BERT も，同一特許フラグの距離が近くなる傾向があることが示され，wiki 事前学習 BERT 同様に，類似特許を評価できることが示された．一方，両者を比較して評価しても，両者に明確な差は見られなかった．

4.3.2　ドメインを考慮した事前学習の活用：k 近傍法による評価

k 近傍法により評価した結果は表 3 のとおりである．ここでは，ベースラインとしての比較のため，事前学習を施さない BERT も用意し，本項では「事前学習無し BERT」

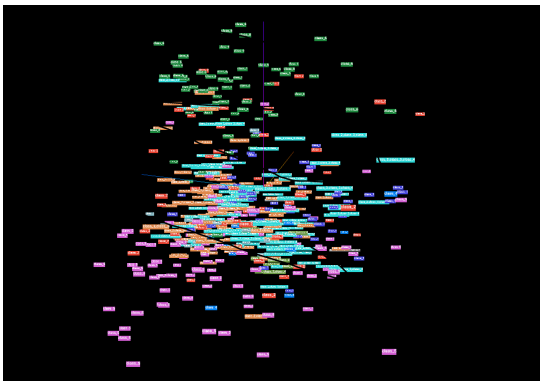

図 6 wiki 事前学習 BERT の可視化 (図 4) の再掲

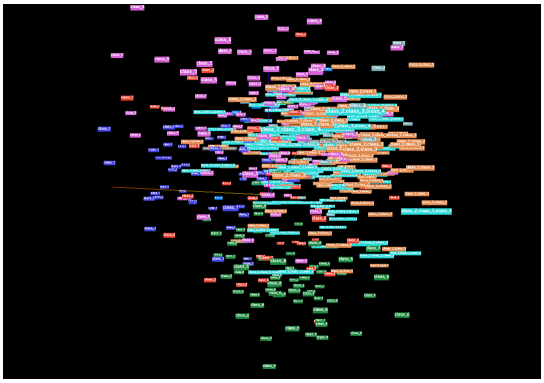

図 7 特許事前学習 BERT の可視化

と呼び，同様に k 近傍法による評価を行う．ただし，図 2 のとおり，本稿のシステム設計において，SentencePiece をシステム内に含んでいるから，SentencePiece の学習を行わないと，本来正しい分かち書きを行うことができない．そこで本稿においては，事前学習無し BERT の分かち書き部分については，特許事前 BERT で利用されている SentencePiece を利用することとして，本稿で評価すべき対象を BERT に限定できるように留意した．

表 3 によると，事前学習無し BERT による正解率が 36.4%，特許事前学習 BERT による正解率が 57.0%となった．特許事前学習 BERT は，事前学習無し BERT の結果と比較すると明らかに高い正解率を上げており，自明ではあるが事前学習することに意義があることを示すことができた．一方，wiki 事前学習 BERT の 61.9%と比較すると正解率は低く，ドメインを考慮した特許事前学習 BERT の優位性を示すことはできなかった．

表 3 ドメインを考慮した事前学習の k 近傍法による比較

手法	正解率	正解数
事前学習無し	36.4%	192
wiki 事前学習	61.9%	327
特許事前学習	57.0%	301

4.3.3 ドメインを考慮した事前学習の活用の考察

以上の二つの評価により，今回の検証の範囲では，ドメインを考慮した特許事前学習 BERT の優位性を示すことはできなかった．ここでは，今回の検証における課題点を述べ，次に分析結果から得られた本手法における有効性について主張する．

前述のとおり事前学習に用いたデータ件数は表 4 のとおりであり，wikipedia 学習と比較して，特許データの件数が明らかに少なかった．特許データというデータの特殊性のため，データを大量に集めることが困難であったためである．一方視点を変えるならば，wikipedia データの 0.01%程度のデータ量で，同等に近い結果を示すことができたともいえる．しかしながら，今回の手法の優位性を正当に主張するのであれば，データ件数を一律にして評価すべきであり，今後の課題であると考えている．

表 4 事前学習のデータ件数

事前学習データ種別	事前学習データ件数
wikipedia データ	約 1,800 万
特許データ	2,637

次に，本手法において明らかになった優位性について述べる．本稿において明らかになった優位性としては，特許特有の単語を評価できるようになったことがあげられる．本稿で提案するシステム構成では，分かち書きに Sentencepiece を利用している．すなわち，事前学習時のデータに依存した分かち書きの学習が行われる．ここでは，「特許事前学習 BERT」の分かち書き結果と，「wiki 事前学習 BERT」の分かち書き結果について，その差を論ずる．

それぞれの分かち書き結果の一部を表 5 に示した。例えば「有底筒状の」という単語は，特許事前学習 BERT の場合「有底筒状の」という一つの単語に，wiki 事前学習 BERT の場合「有」「底」「筒」「状の」という四個の単語に分かち書きされた．「有底筒状の」というキーワードは，特許文書の文脈においては特許特有の単語であるから，「有底筒状の」という一つの用語として表現されることが望ましい．このように，特許というドメインを考慮して学習することで，特許特有の単語を評価できるようになったことは，本項で提案する手法の優位性であると考えられる．一方，特許事前学習 BERT において「を有することを特徴とする請求項」が一つの単語として評価されてしまうなど，一般用語に弱くなる傾向も見受けられる．本稿においては前述のとおり，事前学習のデータ数が少なかったこともあり，一般用語を学習する機会すら乏しかったためと考えられる．そこで，特許データを増強する，あるいは特許データとその他の一般的なデータ (wiki データなど) を適切な配分で混合して利用するなどの対応が考えられ，今後の課題であると考えている．

表 5　分かち書き結果の比較一例

No	特許事前学習 BERT の分かち書き結果	wiki 事前学習 BERT の分かち書き結果
1	有底筒状の	有-底-筒-状の
2	輸送容器	輸送-容器
3	スペーサ	ス-ペー-サ
4	上端開口部	上-端-開口部
5	を有することを特徴とする請求項	を有する-こと-を特徴とする-請求-項

5. おわりに

本稿では，BERT を利用した教師無しデータへ適用について論じた．事前学習済みの BERT を利用し，単語対応部分を評価することで，教師無しデータへの活用が可能となることを示すことができた．今後の課題として，同一ドメイン文書を事前学習に利用する有効性については，データ数を確保できるドメインを考慮するなどして，データ数を考慮した追加実験を行う必要があると考える．また，BERT 以外の手法で類似の研究を行い，他の手法と比較した場合の本手法の有効性について検証する必要があると考えている．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin : Attention is all you need. In Advances in Neural Information Processing Systems, pp6000-6010 (2017).
[2] D.Ciresan, U.Meier, and J.Schmidhuber : D.Ciresan, U.Meier, and J.Schmidhuber : Multi-column deep neural networks for image classification, Proc. of CVPR, pp3642–3649 (2012).
[3] D.Silver et al.:Mastering the game of Go with deep neural networks and tree search, Nature, Vol.529-7587, pp.484–489 (2016).
[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova : BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805 (2018).
[5] Q.V.Le, M.Ranzato, R.Monga, M.Devin, K.Chen, G.S.Corrado, J.Dean, Andrew.Y.Ng : Building High-level Features Using Large Scale Unsupervised Learning, In ICML (2012).
[6] Taku Kudo, John Richardson : SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, arXiv:1808.06226 (2018).
[7] 岡谷貴之：深層学習, 講談社 (2015).
[8] 柴原一友, 藤本浩司：続 AI にできること，できないこと：すっきり分かる「最強 AI」のしくみ, 日本評論社 (2019).