1W-01

特許文書構造を利用したBERTによる事前学習

湯浅 亮也† 谷 和樹†† 田村 晃裕†† 伊藤 和真††† 大林 弘明††† 加藤 恒夫†
同志社大学理工学部インテリジェント情報工学科† 同志社大学理工学部情報システムデザイン学科††
トランスコスモス株式会社†††

1. はじめに

近年，自然言語処理の分野では，大規模コーパスから単語や文などのテキストの汎用的な分散表現を事前に学習する事前学習手法が数多く提案されている．特に，BERT [1]をベースとした事前学習手法を活用することで，様々な自然言語処理タスクにおいて最高精度が更新されている．

自然言語処理の応用先の一つとして注目されているのが，特許情報処理である．特許出願や特許調査などの支援技術として，特許文書に対する文書分類やクラスタリング，検索や機械翻訳などの自然言語処理への期待が高まっている．本研究では，特許文書のための事前学習手法を提案することにより，特許文書に対する自然言語処理の性能改善を目指す．

特許文書は，段落単位でまとめられて記述されており，【発明の詳細な説明】や【技術分野】などの見出しラベルにより構造化されている．しかし，従来の事前学習手法は，通常，文単位でトークンのまとまりを捉える．また，文章における各文の位置付けは考慮しない．そのため，特許文書の構造を反映した分散表現を学習することが困難である．

そこで本研究では，特許文書において段落単位でまとめられる特徴と各段落の見出しラベルで表される特徴の 2 つを考慮した，特許文書のための BERT による事前学習手法を提案する．具体的には，段落同士の連続性推定による事前学習と，各段落が属する見出しラベル推定による事前学習を行うことで，段落単位のまとまりとつながり及び各段落のラベル情報を反映した分散表現を学習する．

特許文書のクラスタリングタスクで提案の事前学習手法を評価した結果，従来の BERT に比べて，F 値が 0.04 高い結果となることを確認した．

2. 従来手法：BERT

BERT [1]は Transformer Encoder [2]に基づくモデルであり，様々な自然言語処理タスクにファインチューニング可能で汎用的な分散表現を獲得するための事前学習モデルである．事前学習では，Wikipedia や BooksCorpus といったラベルなしテキストデータを用いて，Next Sentence Prediction (NSP)と Masked Language Model (MLM)の二つの教師なし事前学習を行う．NSP では，同時に二つの文を入力し，入力された二文が連続するか否かの分類問題を解けるように学習することで，文の接続関係を考慮した分散表現を得る．MLM では，入力トークンをランダムにマスクし，マスク前のトークンを推定できるように学習することで，文脈を考慮した分散表現を得る．

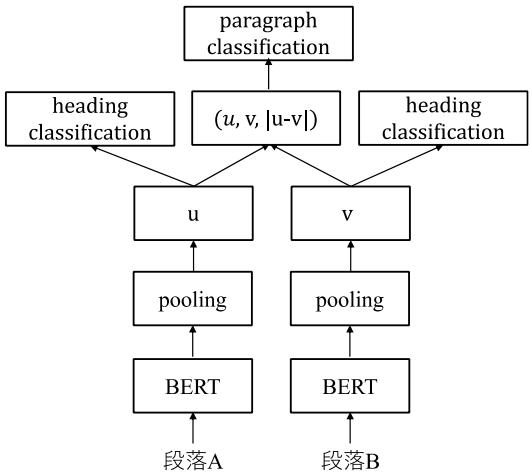

図 1：提案手法の概要図

【発明の詳細な説明】
【技術分野】【０００１】 段落 1
【背景技術】【０００２】 段落 2
【０００３】 段落 3
【発明の概要】

図 2：特許文書の構造例

これらの事前学習は文単位で行われるため，文より大きい単位（段落や章）のまとまりやつながりを捉えることが困難である．また，見出し情報などの各文の属性を陽に利用していない．

3. 提案手法

本研究では，特許文書のための BERT による事前学習手法を提案する．提案手法の概要を図 1 に示す．

特許で記載される内容は定型化されており，それらの内容は，見出しラベル（【技術分野】など）で整理され，段落単位（【０００１】など）で記載されている（図 2）．そこで提案手法では，BERT に段落単位で入力をする．提案手法で用いる BERT は全て重みを共有している．そして，二つの段落が連続するか否かを推定する二値分類問題（paragraph classification）と，各段落が属する見出しラベルを推定する多クラス分類問題（heading classification）を事前学習に導入することで，段落単位でのまとまりとつながり及び見出しラベルの情報を反映した分散表現を獲得する．

提案手法で導入する事前学習の詳細を説明する．提案手法では，二段落の連続性推定による事前学習を行うため，本来の特許文書内で連続する二つの段落の後方の段落を，50%の確率で，異なる特許文書

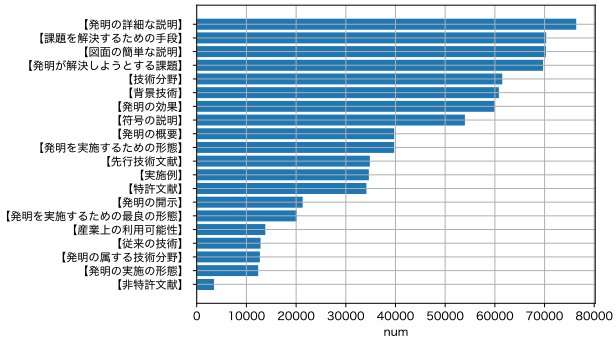

図 3：特許文書の見出しラベルの分布

から無作為に取り出した段落に置き換えたデータを学習データとする．段落間の連続性推定による事前学習は，Classification Objective Function を用いた Sentence-BERT [3]の構造で実現する．具体的には，pooling 処理で各段落内のトークンの平均ベクトルを求め，求めた 2 つの特徴ベクトルとその差の絶対値のベクトルを結合した後，softmax classifier で段落間が連続するか否かを 2 値分類する（式(1)）．

$$softmax\left(W_p(u, v, |u - v|)\right) \quad (1)$$

ここで，uは前方段落（段落 A）の特徴ベクトル，vは後方段落（段落 B）の特徴ベクトルである．

見出しラベル推定は多クラス分類で行う．特許文書 75,000 件中の見出しラベルの分布を図 3 に示す．図 3 の通り，特許文書内で頻出する見出しラベルは決まっている．そこで，頻出する特定の見出しラベルを対象にした多クラス分類を事前学習時に行うことで見出し情報を考慮した分散表現を学習する．各段落に対して，段落内トークンの平均ベクトルに基づき，N 個の頻出見出しのクラスと「頻出見出し以外」クラスの合計 N + 1 クラス分類を行う（式(2)）．

$$softmax(W_h u) \quad (2)$$

ここで，uは段落 A または B の特徴ベクトルである．

4. 実験

特許文書のクラスタリングタスクにより提案手法を評価する．提案手法は，東北大学 BERT [4]を初期値として特許文書で MLM により事前学習することで，特許分野に分野適応する．その際，段落同士の連続性推定と各段落の見出しラベル推定を行う．この事前学習モデルを用いて，特許文書の【背景技術】と【発明が解決しようとする課題】部分のテキストに対する特徴ベクトルを求め，求めた特徴ベクトルに基づいて凝集型クラスタリングを行う．距離尺度はユークリッド距離，マンハッタン距離，コサイン距離を用いた．提案手法の有効性を検証するため，BERT の初期値，従来手法で分野適応したモデル，段落間連続推定のみ導入した提案モデルのそれぞれを事前学習モデルに用いた場合の性能評価も行う．

全ての手法において，エポック数，バッチサイズ，学習率は，それぞれ，3，12，2e-5 とした．見出しラベル推定では，学習データにおいて頻度が多い上位 19 個の見出しを頻出見出しとした（N=19）．なお，図 2 のように，1 つの段落が複数の見出しラベルに属する場合は，直前の見出しラベルをクラスとした．事前学習時の学習データには，教師なし特許データ 5,000 件を使用し，クラスタリング時の開発データ及びテストデータには人手で正解を付与した特許データ 73 件，95 件をそれぞれ用いた．クラスタリングの性能は，Purity と Inverse Purity の調和平均を取った F 値で評価した．

表 1 ： 実験結果

手法	入力単位	分野適応	段落間連続推定	見出し推定	F 値
従来手法	文	なし	なし	なし	0.47
	文	あり	なし	なし	0.50
	段落	あり	なし	なし	0.51
提案手法(段落間連続推定のみ)	段落	あり	あり	なし	0.52
提案手法	段落	あり	あり	あり	**0.54**

5. 実験結果

実験結果を表 1 に示す．表 1 より，分散表現を特許分野に適応させることで特許文書のクラスタリング性能を改善できることが分かる．また，入力単位を段落単位にしたり，段落間の連続性判定を導入したりすることにより，段落単位のまとまりやつながりを考慮した分散表現を獲得することは，特許文書のクラスタリングにおいて有効であることが分かる．さらに，各段落が属する見出し情報を事前学習に取り入れることで性能が改善したことから，見出し情報は特許文書の特徴を表す際に重要な情報であることが確認できる．

6. まとめ

本研究では，特許が見出しラベルで整理され，段落単位で記載される点に着目し，それらの特徴を考慮した BERT による特許文書のための事前学習手法を提案した．特許文書のクラスタリング実験を通じて，本研究で提案した，段落同士の連続性推定による事前学習と，各段落が属する見出しラベル推定による事前学習がそれぞれ有効であることを確認した．

謝辞

本研究を進めるにあたり様々な助言をしてくださった愛媛大学の二宮崇先生，梶原智之先生，秋山和輝氏に感謝を申し上げます．

参考文献

[1] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proc. of NAACL-HLT 2019.
[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Proc. of NIPS 2017.
[3] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proc. of EMNLP-IJCNLP 2019.
[4] https://huggingface.co/cl-tohoku/bert-base-japanese-char-whole-word-masking