継続事前学習によるドメイン特化 LLM 構築のための MLM

高本　涼馬*，増田　大河，平川　翼，山下　隆義，藤吉　弘亘（中部大学）

Building Domain-Specific Large Language Models via Continual Pretraining Using MLM

Ryoma Takamoto, Taiga Masuda, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi (Chubu University)

１．はじめに

汎用的かつ大規模な自然言語処理モデルは，多様な言語タスクにおいて高い性能を示しており，様々なタスクで応用されている．Bidirectional Encoder Representations from Transformers (BERT) (1)は，自己教師ありの事前学習モデルとして広く用いられ，双方向の文脈情報を利用可能であることから，汎用的な言語理解モデルの基盤となっている．

しかし，BERT を含む多くの大規模言語モデルは Wikipedia や BooksCorpus などの一般的な文章を用いて事前学習されており，学術論文などの分野特有の語彙や表現に対しては十分な性能を発揮できないという課題がある．

本研究では，専門的な文脈や語彙を対象とし，専門的な表現の獲得が可能な学習手法を提案する．

２．大規模言語モデル

自然言語処理分野で著しい進化を遂げている Transformer (2)ベースの大規模言語モデル (LLM) は，大量のテキストデータセットに対して自己教師あり学習を行うことで，文脈に即した意味表現の獲得を可能にし，多様な下流タスクに対して高い性能を示している．

Transformer は，従来の RNN や CNN ベースの構造と異なり，自己注意機構 (self-attention) を用いることで，系列長に依存しない高効率な文脈処理を実現した．この構造は並列計算に適しており，大規模な事前学習に向いている．LLM は，この Transformer を基本構造とし，自己教師ありの事前学習によって獲得した文脈表現を，少量のデータによるファインチューニングによって多様なタスクに適用する．

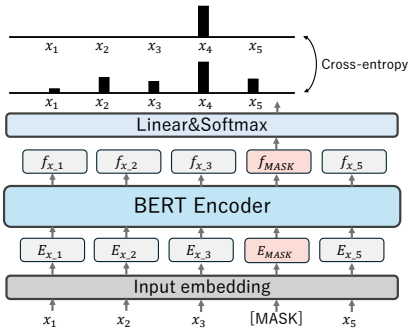

Fig.1.　BERT pre-training (MLM)

LLM の代表的な例としては，Bidirectional Encoder Representations from Transformers (BERT) や Generative Pre-trained Transformer (GPT) (3) が挙げられる．BERT は文の理解を目的としたエンコーダのみを用いた構造を持ち，文章の 1 部の単語をマスクトークンに置き換えた文章をモデルに入力し，マスクトークンに置き換えた単語を予測するタスクである Masked Language Modeling (MLM) (Fig.1)による事前学習を行う．これによりモデルは文脈を前後関係から適切に捉えることが可能となり，意味的に整合性がある内部表現を獲得できる．

一方，GPT はデコーダのみを用いた構造を持ち，次の単語を予測するタスクにより事前学習され自然言語生成タスクにおいて高い性能を発揮する．

このように，LLM は構造や学習方式の違いに応じて，自然言語を理解するタスクに強いモデルと生成タスクに強いモデルに大別される．本研究では，文の意味や語彙の表現獲得を得意とする BERT 系列の LLM を使用し，専門的な表現の獲得が可能な学習手法を提案する．

３．関連研究

BERT は事前学習に一般的な文章が用いられており，専門分野の語彙や文脈を十分に捉えることは困難である．この課題を解決するため，特定ドメインに特化した BERT モデルが提案されている．たとえば，BioBERT (4) は，BERT-Base モデルを生物医学分野の大規模コーパスである PubMed の要旨や PMC の全文を用いて継続事前学習を行うことで，生物医学特有の専門知識を効果的に獲得している．

一方，SciBERT (5) は，科学論文から抽出した約 3.1 億トークンのデータセットを用い，科学分野に特化した語彙（SciVocab）を構築し，BERT と同様の Transformer 構造で学習することで，科学文献の多様なタスクに対応可能なモデルを実現している．

さらに，SecureBERT (6) は RoBERTa (7) をベースに，サイバーセキュリティ関連の脅威インテリジェンスや脆弱性情報などの専門テキストを大量に用いた MLM による追加学習を行い，セキュリティ分野に特化した高精度な言語理解能力を獲得している．

これらのモデルはいずれも，BERT モデルを基盤として各分野のテキストを用いた継続事前学習により専門的な文脈や語彙に対応する性能を獲得している．

しかし，これらの手法には共通する課題が存在する．それは, MLM におけるマスク対象の選定時にランダムマスキングが用いられている点である．ランダムマスキングでは，文中

のトークンが無作為に選ばれるため効率的に専門分野の特徴の獲得ができない．そこで本研究では，専門分野に特化した MLM での継続事前学習を行うことで，効率的に専門分野の特徴の獲得を図る．

４．提案手法

本研究では，専門分野の語彙や文脈を効率的に獲得するために、BERT の事前学習で行われる MLM の改良を行い，Domain-Specific Masked Language Modeling (DS-MLM) を提案する．本節では，提案手法である DS-MLM のマスク手法と学習手法について詳述する．

＜4・1＞モデルの構成　提案手法のモデル構成は，BERT と同一構造であり，12 層の Transformer ブロック，各層に 12 個の自己注意ヘッド，隠れ層の次元数 768，語彙サイズ 30,522 からなる BERT-Base の構成に基づいている．また，トークンの埋め込み，位置エンコーディング，セグメント埋め込みの各モジュールも原著論文に準拠しており，事前学習済みの BERT モデルを初期値として用いる．

＜4・2＞DS-MLM による継続事前学習　提案手法である DS-MLM は，既存の事前学習手法に加えて，対象とする専門分野に特化させる継続事前学習を行うことで，当該分野における表現能力を向上させることを目的とする．DS-MLM のマスク処理方法を Fig.2 に示す．

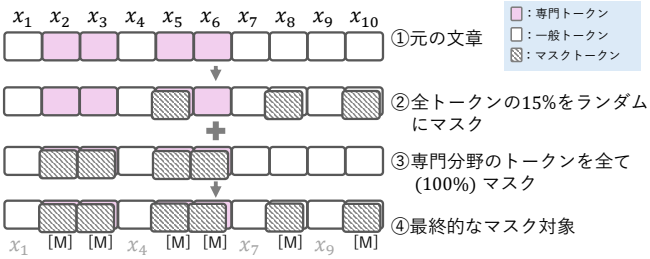

Fig.2.　Mask processing method for DS-MLM

DS-MLM におけるマスク処理は，2 段階の手順で構成される．第 1 段階では，従来の MLM と同様に，入力文中のトークンのうち 15%をランダムにマスクする．第 2 段階では，専門的な単語に該当するトークンをすべてマスク対象とする．これにより，一般的な文脈に加えて，専門的な単語や文脈における表現を重点的に学習することが可能となる．

５．実験

提案手法の有効性を検証するために MLM タスクでマスクされたトークンの予測精度を正解率により評価する．加えて，専門分野に特化させた MLM の継続事前学習が下流タスクにおいても有効であることを確認するため，2 文間の類似性判定を目的とする下流タスクでファインチューニングを行い，コサイン類似度に基づく二値分類の正解率を指標として評価を行う．

＜5・1＞実験条件　モデルは BERT の基本モデルである“bert-base-uncased”を使用する．エポック数は 50，データセットは，化学分野の論文から抽出した仮説文と結論文から構成される CRNLI データセットを用いる．化学的な単語の定義は化学的な文章で構成された CRNLI データセットに存在し，一般的な文章で構成されたデータセットの SNLI (8) には存在しないトークンとする．

＜5・2＞マスクの正解率の比較　比較対象として，以下の 3 つの手法を対象とする．

(i) 一般分野，化学分野関係なく，文章全体のトークンをランダムに 15%マスクする従来手法 (MLM)

(ii) 化学トークンのみを対象にランダムに 15%マスクする手法 (Chem-MLM)

(iii) 全トークンの 15%をランダムにマスクしつつ，化学トークンを追加で 100%マスク対象とする手法 (DS-MLM)

各手法における学習後のマスク予測精度を Table 1 に示す．マスク予測精度の評価は，化学トークンのみをマスクする場合，一般トークンのみをマスクする場合，全てのトークンをマスク対象とする場合の 3 つの観点で比較を行った．

Table 1 より，MLM と Chem-MLM を比較すると，Chem-MLM は化学トークンに対する予測精度は 66.1%と高い値を示したが，全体の正解率や一般単語の正解率はそれぞれ 31.5%，19.9%と低下した．この結果は，特定分野への過度な特化が一般分野における表現の獲得に悪影響を与えていることを示唆している．

一方で，MLM と，DS-MLM を比較すると，DS-MLM は化学トークンを重点的に学習しつつも，一般トークンに対してもマスクを行うことにより，バランスの取れた性能を達成している．このことから，DS-MLM は一般的な文脈情報を保持しながら，専門分野に特化した文脈の獲得にも成功していると評価できる．

Table 1　Percentage of correct answers with varying tokens masked during learning [%]

マスク手法	ALL	化学単語	一般単語
MLM	71.6	55.3	80.8
Chem-MLM	31.5	**66.1**	19.9
DS-MLM	**74.6**	62.1	**80.4**

＜5・3＞DS-MLM と MLM の予測単語による比較

DS-MLM と MLM における予測単語の比較で定性的評価を行う．定性評価を Fig.3 に示す．

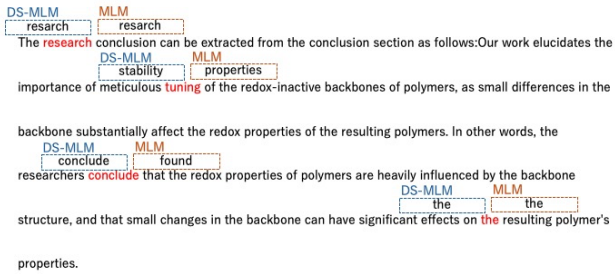

Fig.3.　Comparison of DS-MLM and MLM by predicted words

Fig.3 から，“resarch”，“the”などの一般的な単語は ML と DS-MLM の両方で予測できていることが分かる．一方，“conclude” という専門的な単語は DS-MLM で予測できているため，専門的な単語の特徴を捉えることができてい

る．しかし，“tuning”という専門的な単語はどちらの手法も予測できていないため，学習が十分ではないと言える．

＜5・4＞下流タスクによる比較　2 文の類似性を判定する下流タスクにおける分類精度の比較結果を Table 2 に示す．分類ラベルとの正解率を指標として用いる．

Table 2 より，DS-MLM は従来の MLM（89.80%）よりも高い精度（91.93%）を達成しており，Chem-MLM と同等の性能を示している．一方，Chem-MLM はマスク予測タスクにおいて全体の正解率が低かったにもかかわらず，本タスクでは高精度を示しており，ファインチューニングにより，化学分野の特徴をうまく抽出できたと考えられる．

以上より，DS-MLM および Chem-MLM はいずれも，専門的な文章に内在する特徴を抽出可能な学習手法であると考えられる．

Table 2　Comparison of classification accuracy in downstream tasks by each method [%]

	MLM	Chem-MLM	DS-MLM
正解率	89.80	91.93	91.93

６．おわりに

本研究では，専門的な表現の獲得に有効な学習手法として，DS-MLM による継続事前学習を適用し，精度の向上を確認した．さらに，得られたモデルの入力処理を分析した結果，専門的な単語の多くがトークナイザによって複数のサブワードに分割されていることが明らかとなった．このことは，専門的な語彙に対するモデルの理解を妨げる可能性があると考えられる．今後は，トークナイザの辞書の内容を変更することで専門的な文章の読解能力の向上を図る．また，その他のマスク手法による MLM の実験を行うことを検討する．

文　献

(1) J. Devlin, et al:“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”NAACL, pp. 4171‑4186, 2019.

(2) A. Vaswani, et al:“Attention Is All You Need” NeurIPS, vol. 30, pp. 5998‑6008, 2017.

(3) A. Radford, et al:“Improving Language Understanding by Generative Pre-Training,” OpenAI Technical Report, 2018.

(4) J. Lee, et al.:“BioBERT: a pre-trained biomedical language representation model for biomedical text mining” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.

(5) I. Beltagy, et al.:“SciBERT: A Pretrained Language Model for Scientific Text” EMNLP, pp. 3615–3620, 2019.

(6) E. Aghaei, et al.:“SecureBERT: Towards Pre-trained Language Models for Cybersecurity Text” Springer, pp. 39–56, 2023.

(7) Y. Liu, et al:“RoBERTa: A Robustly Optimized BERT Pretraining Approach” arXiv:1907.11692, 2019.

(8) R. Bowman, et al.:“A large annotated corpus for learning natural language inference” arXiv:1508.05326, 2015.