継続事前学習によるドメイン特化 LLM 構築のための Domain-Specific MLM

EP21059 高本 涼馬　　指導教授：山下 隆義

1. はじめに

学術論文の読解は，研究の進展や新しい知見を得るために重要である．しかし学術論文は分野独自の表現や複雑な文脈で書かれることが多く，論文内容の読解には多大な時間と労力が必要であり，研究活動の課題となっている．この課題を解決するため，Transformer [1] を用いた大規模言語モデル (LLM) である BERT [2] や RoBERTa [3] の活用が，テキスト分類や質問応答などのタスクで注目されている．これらの LLM は，人が日常的に使う一般文章を対象としており，専門単語を多く含む学術論文を読解することが出来ない．そこで本研究では，化学分野を対象とし，専門的な単語からなる文章から特徴を抽出可能な言語モデルの構築を目的とする．

2. BERT

Bidirectional Encoder Representations from Transformers (BERT) は，Transformer を基盤とした事前学習モデルである．BERT はトークナイザを用いて入力文章からトークンを作成する Input embedding と，作成したトークンの特徴を抽出する BERT Encoder により構築される．BERT の学習は事前学習とファインチューニングの 2 段階で行われる．図 1 に BERT の事前学習の一つである Masked Language Modeling (MLM) を示す．

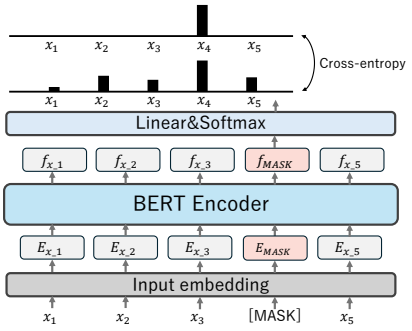

図 1：BERT の事前学習

MLM は，文章の一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．マスクトークンは通常文章全体の 15%の割合でランダムな位置に配置する．これによりモデルは文章全体を考慮して単語を予測することが可能となる．

3. 提案手法

学術論文は分野特有の表現や専門単語が多く，一般的な文章で学習されたモデルでは読解できない．そこで，専門的な単語の特徴がより多く抽出されるようマスクした Domain-Specific MLM (DS-MLM) を提案する．DS-MLM のマスク処理方法を図 2 に示す．DS-MLM のマスク処理は 2 段階で行われる．はじめに，一般的なマスク処理と同様に，文章全体の 15%のトークンをランダムにマスクする．専門的な単語に該当するトークンをすべて（100%）マスクする．この 2 段階のマスク処理により，専門分野特有の単語やその表現を重点的に学習することが可能となる．

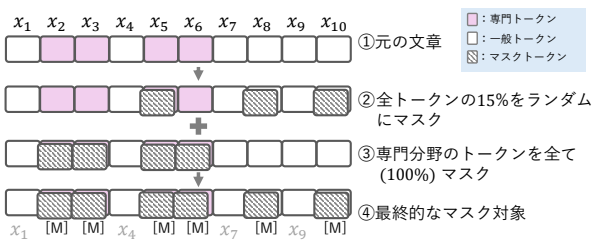

図 2：DS-MLM の概要

4. 評価実験

提案手法の有効性を検証するために MLM タスクでマスクした単語の正解率の評価を行う．また，専門的な文章の特徴が抽出されているかを調査するため，2 つの文章の類似性を求める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き正例，不例を分類した際の正解率で評価する．

4.1. 実験条件

データセットには，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．専門的な単語を専門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[4] には存在しないトークンと定義する．

4.2. マスクの正解率の比較

比較実験では，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%マスクした従来手法 (MLM) と化学トークンのみを対象にランダムに 15%マスクした手法 (Chem-MLM) と比較する．学習時にマスクするトークンを変化させた時のマスク正解率の比較を表 1 に示す．表 1 より，Chem-MLM は化学単語の正解率は 66.1%と高いスコアを示したものの，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．これは，化学分野に特化した学習が一般分野の特徴に悪影響を与えていることを示す．これに対し，化学トークンの割合を増加させつつ一般トークンもマスク対象に含めた DS-MLM は，全体の正解率が 74.6%と精度が高い．これにより，DS-MLM は一般分野の特徴を維持しつつ，化学分野の特徴を抽出できたといえる．

表 1：学習時にマスクするトークンを変化させた正解率 [%]

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

4.3. 下流タスクによる比較

2 つの文章の類似性を求める下流タスクでの結果を表 2 に示す．分類結果とラベルの正解率を評価指標として用いる．表 2 より，DS-MLM の正解率が MLM を上回っており，分類性能が向上していることが分かる．Chem MLM も分類性能が向上しており，専門分野の事前学習が有効であると言える．これらの理由により，DS-MLM の学習法が専門的な文章の特徴を上手く抽出していることがわかる．

表 2：ラベルの正解率の比較 [%]

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

5. おわりに

本研究では，化学分野のような専門的な文章を読解できる事前学習手法を構築し，精度向上を確認した．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．

参考文献

[1] A.Vaswani *et al.*, “Attention Is All You Need”, NIPS, 2017.

[2] J.Devlin *et al.*, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, ICLR, 2019.

[3] Y.Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, NIPS, 2019.

[4] R. Bowman *et al.*, “A large annotated corpus for learning natural language inference”, arXiv, 2015.