逆翻訳とゼロショット学習に基づく BERT での件名自動付与
～TRC MARC を使った実験～

門脇夏紀†　　岸田和明‡

† 慶應義塾大学非常勤講師
kadowaki72@keio.jp

‡ 慶應義塾大学文学部
kz_kishida@keio.jp

抄録

本研究では，図書への件名の自動付与に，マルチラベル分類用の BERT モデルを適用する実験を試みた。その際，逆翻訳による訓練レコードの追加の効果と，BERT のマスク語予測機能を用いたゼロショット学習の可能性に焦点を当てた。実験には，TRC MARC データを用い，一部の件名を意図的に選択して，それに対する分類器（BERT モデル）を構築し，評価を行った。その結果，逆翻訳による訓練レコードの増加には，分類器の性能を向上させる効果が観察され，また，ゼロショット学習に関しては，ある程度，件名の自動付与を実現する可能性が見出された。

1. はじめに

GPT とともに，Transformer に基づくアルゴリズムとして，BERT がよく知られている。その発表は 2018 年であり，それ以来，テキスト分類に BERT を応用する試みが数多くなされてきた。文献に対する件名（標目）の自動付与に関しても，MeSH を対象とした You ら[1]の例がある。

BERT をはじめとする機械学習の手法に基づいて，図書や論文に件名を自動付与するのは，それほど容易ではない。まず，NDC 番号の付与がシングルラベル分類に相当するのに対して，件名の場合にはマルチラベル分類となり，作業的に複雑である。また，ラベルとしての件名の異なり数が多く，なおかつ，訓練データにおいて，その出現文献数が偏っているという，実際上の問題もある。

本実験で使用した TRC MARC のデータ（約 1 年の間に作成された MARC レコード群）における出現図書数での件名の分布を図 1 に示す。『基本件名標目表』（BSH）に含まれる件名に限定しているものの，図 1 には，3,031 個の件名が含まれ，それらが出現図書数（縦軸）の順に横軸に並べられている。それらが出現する図書の延べ冊数は 25,913 で，最も出現していた件名である「人生訓」でさえ，そのうちのわずか 547 冊に含まれるのみだった。すなわち，最頻出の件名でさえ，全体のわずか 2.1%で使われているのに過ぎない。

図 1 の分布は Zipf の法則に従った形状をしており，件名群は，おおよそ，

A) 標本サイズがそれなりに大きい
B) 標本サイズがかなり小さい
C) 該当レコードがほとんどない

の 3 つのグループに大別される。この種のデータを一括して，機械学習アルゴリズムに投入しても，十分な分類器を構成できるとは考えにくい。

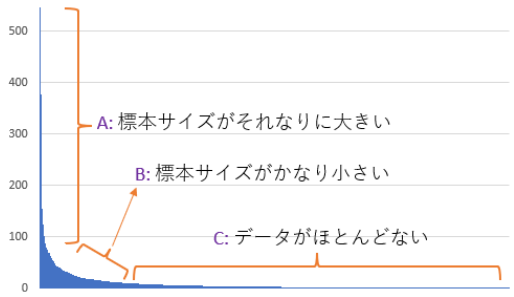

図 1 各件名の出現図書数の分布

Tunstall らによる Transformer の解説書[2]には，B)に対する方策として逆翻訳（Back translation）による訓練レコードの追加，C)についてはゼロショット学習（Zero-shot learning）の応用が紹介されている。本研究の目的は，BERT を使った件名の自動付与に対して，これらの方法が有効であるかどうかを確認することにある。本稿では，TRC MARC のデータに含まれる一部の件名に対して，これらの 2 つの方法の効果を実験で確かめた結果を報告する。

2. 関連研究

You ら[1]のほか，最近，BERT を MeSH の件名付与に用いた試みとして，Lin ら[3]がある。彼らは，MeSH タームの付与のために，4 つの BERT モデルに基づくアンサンブル学習を提案した。

一方，Chou と Chu[4]は米国議会図書館件名標目表（LCSH）の自動付与に BERT を利用することを探究している。なお，マルチラベル分類そのものに関しては，岸田ら[5]がレビューしている。

3. 実験データ

本研究で使用するデータは，2019 年 12 月から約 1 年の間に作成された TRC MARC のレコード群である。ただし，今回の実験では，逆翻訳とゼロショット学習の効果の初歩的な確認に留めることから，レコードの一部のみを使用した。

具体的には，8 個の件名から成る集合を設定し，それに限定したマルチラベル分類器を BERT で実装した。図 1 に関して説明したように，レコード全体では，3,031 個の件名が出現している。これらを 1 度に取り扱う分類器を構築するのは無理と判断し，まずは 8 個の件名に限った分類器を作って，その性能を細かく検討することとした。

データを調べつつ，試行錯誤の末，次の 2 つの件名集合を設定した。

集合 L: 人生訓，料理，人間関係，コミュニケーション，健康法，話しかた，栄養，食生活

集合 M: 相続税，税務会計，工作，手芸，童謡，遊戯，生理学，解剖学

ここで，L は「大規模」，M は「中規模」を意味している。すなわち，集合 L 中の各件名は十分な数の図書に出現しているのに対して，集合 M 中の件名はそれほどでもない。本研究では，まず集合 L に対して，BERT でのマルチラベル分類の性能を確認したのち，それとは独立に，集合 M に対して，逆翻訳による訓練データの追加やゼロショット学習の効果を確かめる。

これらの件名の選択にあたっては，件名ペアが共起する図書数に注意した。つまり，マルチラベル分類の性能を確認する際に，単一の件名のみが付与された図書だけが分析対象となっては，意味がない。そこで，複数の件名を持つ図書が何件かは含まれるよう，件名ペアの共起図書数に着目しつつ，慎重に，集合 L と M とを構成した。

これらの集合に対し，書名と内容紹介フィールド中のテキストデータを使って，別個の BERT モデル（すなわち分類器）を構築し，評価した。その際のレコード件数を表 1 に示す。なお，「訓練」用と「検証」用のデータと，「評価」用のデータとでは，その性質が異なる。訓練および検証データは，集合 L と M それぞれで，8 個の件名のみが出現する図書から構成されている。一方，評価データはそうではない。それぞれの具体例（1 冊の図書での件名）を次に示す。

訓練または検証：「料理，栄養，健康法」

評価：「料理，喫茶店」

これらの件名の中で「喫茶店」のみ，集合 L に含まれていない。

すなわち，8 個の件名で「閉じた」レコード集合で分類器を学習し，「開かれた」レコード集合で評価を行うこととした。後者の場合，評価データからは，8 個以外の件名（上記の例では「喫茶店」）は削除される。

表 1 BERT モデルの実験に使用するレコード数

	訓練	検証	評価
集合 L	1026	100	100
集合 M	100	28	100

「閉じた」集合に含まれる図書は集合 L では 1,126 件で，それを訓練用 1,026 件，検証用 100 件に分割した。一方，集合 M の場合には，128 件だったので，100 件を訓練用，28 件を検証用とした。評価データに関しては，「開かれた」集合（L に関しては 874 件，M に関しては 367 件）から，それぞれ 100 レコードを無作為抽出した。

4. BERT によるマルチラベル分類の実験(1)

最初に，集合 L および M それぞれのデータに対して，通常のマルチラベル分類を試みた（データ中には，件名が 1 つのみの図書も数多く存在する）。具体的は，AutoModelForSequenceClassification[6]に，「cl-tohoku/bert-base-japanese」を組み込み，problem_type に「multi_label_classification」を指定した。この実装にあたっては，Transformer のチュートリアル[7]を参考にしている。実行時には，入力テキストの最大トークン数を 128，バッチサイズを 8，学習率を「2e-5」，減衰重みを 0.01 とした。

表 2 BERT での件名付与の結果（マクロ平均）

データ		正解率	精度	再現率	F 値
集合 L	検証	.810	.802	.868	.812
	評価	.630	.621	.747	.663
集合 M	検証	.679	.813	.652	.678
	評価	.450	.473	.276	.329

表 2 に実験結果を示す。これらは，検証データで損失関数の値を確認し，集合 L では最終的なエポック数を 5，集合 M では 15 とした結果である。な

お，評価指標の計算には，scikit-learn のモジュールを用いた。

検証データ（閉じた集合）に対しては，標本サイズが大きい集合 L のほうが優れた結果を示した（F 値については，集合 L が 0.812，集合 M で 0.678）。評価データ（開いた集合）に関しても，集合 L と集合 M との間で，顕著な差異が観察された（F 値で 0.663 と 0.329）。ただし，評価データでは，指標の値が両集合ともに低下した。「閉じた」集合に対する図書と「開いた」集合のそれとでは質的な相違があり，前者のみで学習した分類器は，後者では十分に機能しないのかもしれない。

集合 L に対しては指標の値は比較的高く，TRC MARC データに対して，BERT はある程度機能すると判断できる。一方，具体的な失敗例を以下に示す。

①予測：「健康法，食生活」　正解：「料理」

②予測：「遊戯」　正解：「童謡，遊戯」

①は 1 つも正解できなかった例である。一方，②では，正解となる 2 つの件名のうち，1 つは予測できている。

なお，表 1 のデータ以外に，集合 L および M に無関係なレコード（すなわち 8 個の件名を 1 つも含んでいないレコード）100 件を無作為抽出して，各分類器を適用してみた。その結果，例えば，「社会科」のみを件名とする図書に対して「人生訓」が付与されるなどの過剰付与が何件か生じた。その数は集合 L では，34 件だったのに対して，集合 M では 9 件のみであった。もし現実の状況で各分類器を単独で使用するならば，「開いた」集合の図書だけでなく，「無関係」な図書に対して，付与の誤りが生じる可能性がある。

5. BERT によるマルチラベル分類の実験(2)

次に，集合 M のデータに関して，逆翻訳の適用を試みた。今回は，「日本語→英語→日本語」のパターンで逆翻訳することとし，翻訳には Google 翻訳を利用した。

逆翻訳の例を以下に示す。

①元の文：「ミシンなしでかんたん！かわいい手芸どうぶつ。」

②逆翻訳：「ミシンなしで簡単！かわいい手作り動物たち。」

逆翻訳により，「かんたん」→「簡単」，「手芸どうぶつ」→「手作り動物たち」のように，表現が増加していることが分かる。

訓練データ 100 件に対して逆翻訳を行い，その正解ラベルとして，元の図書のものをそのまま複製したうえで，それらを訓練データに追加した。すなわち，訓練データは 200 件に増えることになる。

第 4 節と同じ条件で，この訓練データを使って，マルチラベル分類を実行した結果を表 3 に示す。検証データにおける精度の値を除き，評価指標の値は大幅に改善された。その精度に関しても，ともに 0.813 であり，劣化したわけではない。すなわち，今回の事例では，逆翻訳による訓練データの追加には一定の効果が認められた。

表 3 逆翻訳による件名自動付与の結果

データ		正解率	精度	再現率	F 値
集合 M	検証	.893	.813	.813	.792
	評価	.640	.842	.552	.641

6. BERT によるマルチラベル分類の実験(3)

最後に，BERT のマスク語予測モデルに基づくゼロショット学習による件名付与の実験結果を報告する。

6.1 transformers の fill-mask の利用

具体的には，Tunstall ら[2]に従い，Hugging Face の transformers で提供される pipeline で，タスクとして「fill-mask」を指定した。もし，付与対象の図書の書名と内容紹介を並置したテキストデータが変数 text に格納されているならば，

```
from transformers import pipeline
model_ckpt = 'cl-tohoku/bert-base-japanese'
pipe = pipeline('fill-mask', model=model_ckpt)
prompt = '以上の内容は[MASK]に関するものである。'
output = pipe(text + prompt)
```

を実行すれば，BERT がプロンプト中の[MASK]にあてはまる語を予測し，その結果を変数 output に戻してくれる。デフォルトでは語数は 5 で，それぞれにスコアが付随する。それらの語を$t_1, \ldots, t_m$，スコアを$s_1, \ldots, s_m$と表記しておく（$m = 5$）。

仮のテキストデータに対する結果を図 2 に示す。この例では，図 2 中の書名と内容紹介に対して，「検索」などの 5 つの語が，マスクされた部分に当てはまるものとして予測されている。

6.2 マスク語と件名との類似度の計算

実験では，マスクに対する予測語の集合と各件名との類似性を測定し，最も近い件名を当該図書に付与することにした。類似度の計算には，BERT に内蔵されている分散表現（768 次元）を用いた。分散表現は tokenizer 中の vocab を参照すれば取

得可能である。

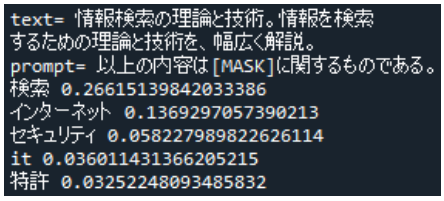

図 2 BERT によるマスク語予測の例

まず，マスクに対して予測された語の分散表現を取り出した。ここでは，t_jに対する分散表現をv_jと書く（$j = 1, \ldots, m$）。そして，5 個の予測語全体のベクトルV_Mを以下のように求めた。

$$V_M = \frac{1}{m}\sum_{j=1}^{m} s_j v_j$$

ここで，s_jはスカラーで，予測語の重みとして機能する。

一方，件名の場合には，最初に，BERT の tokenizer で語分割する。例えば，「税務会計」という件名に対して tokenizer を適用すると，

{'input_ids': [2, 19727, 6787, 3], 'token_type_ids': [0, 0, 0, 0], 'attention_mask': [1, 1, 1, 1]}

が出力される。ここで「19727」が「税務」，「6787」が「会計」を意味し，この id により，BERT モデルから分散表現を抽出できる。その後，上記のV_Mと同様に，BSH に含まれるすべての件名のベクトルを算出した。これをV_Sと表記する。ただし，件名の構成要素の重みはすべて 1 である。

最後に，V_MとV_Sの間の余弦係数を求め，その値の大きな件名を当該図書に付与する。以上が，本研究で提案するゼロショット学習での件名の自動付与方法である。

6.3 ゼロショット学習についての実験結果

集合 M に対する検証データ（100 件）に対して，本研究で提案するゼロショット学習での件名付与を行った結果を表 4 に示す。

表 4 ゼロショット学習での件名付与の結果

件名の採用	正解率	精度	再現率	F 値
上位 1 件	.179	.375	.156	.207
上位 3 件	.250	.375	.250	.282

注：集合 M の検証用レコード 100 件での評価

当然，ファインチューニングを実行した表 2 または表 3 の結果に比べれば，評価指標の値は大きく低下している。それでも，誤った例の中には，この方法の可能性を感じさせるものが散見された。一例を挙げれば，以下のとおりである。

予測：「人体」，正解：「解剖学」
予測：「テレビ ゲーム」，正解：「遊戯」
予測：「デザイン，和服，和裁」，正解：「手芸」

ゼロショット学習での方法の性能を向上させるには，今後，予測語と件名との照合を精緻化する必要がある。これには，件名標目表の階層性を利用することなどが考えられる。さらには，マスク語予測以外の仕組み（例えば生成 AI）を，より積極的に活用する方向性もあり得るかもしれない。

7. おわりに

今回の限定された実験では，BERT のマルチラベル分類を，件名の自動付与に適用する可能性が改めて示唆された。また，逆翻訳による訓練レコードの追加やゼロショット学習についても，一定の効果が認められた。

謝辞 株式会社図書館流通センターから，研究用として TRC MARC データを提供していただきました。感謝申し上げます。

引用文献

1) You, R., et al. “BERTMeSH: Deep contextual representation learning for large-scale high-performance MeSH indexing with full text.,” Bioinformatics, vol.37, 2021, pp.684–692.
2) Tunstall, L. et al. Natural Language Processing with Transformers: Building Language Applications with Hugging Face. Oreilly & Associates, 2022.
3) Lin, S-J, et al. “A BERT-based ensemble learning approach for the BioCreative VII challenges: full-text chemical identification and multi-label classification in PubMed articles,” Database, 2022, article ID baac056.
4) Chou, C. and Chu, T. “An analysis of BERT(NLP) for assisted subject indexing for Project Gutenberg,” Cataloging & Classification Quarterly, vol.60, 2022, pp.807-835.
5) 岸田和明ほか．ゲームソフトの評価レビューに対するマルチラベル分類における SVM と BERT の比較. 情報処理学会・第 147 回情報基礎とアクセス技術研究発表会. Vol. 2022-IFAT-147, 2022, p. 1-6.
6) https://huggingface.co/transformers/v3.0.2/model_doc/auto.html, (参照, 2023-07-12).
7) Regge, N. Fine-tuning BERT (and friends) for multi-label text classification. https://github.com/NielsRogge/Transformers-Tutorials/blob/master/BERT/Fine_tuning_BERT_(and_friends)_for_multi_label_text_classification.ipynb, (参照, 2023-07-12).