クイズ AI の実現に向けたジャンル推定における多ラベル分類に関する研究

浅野 綾太　　指導教員：小林 邦和

1 はじめに

本研究では, 競技クイズにおける問題のジャンル推定を取り扱う. 先行研究 [2][1] で得た結果や考察を取り上げ, この問題点や考察より研究テーマとした. クイズにおけるマルチラベル分類の実現が達成されれば, クイズ AI の実現, 強いては早押しに対応したクイズ AI の開発や, その他自然言語における様々なタスクに応用できると考えている. 本研究では BERT を用いたマルチラベル分類モデルを作成し, 複数の評価指標を用いて評価を行った. 結果としては F1 のスコアで 8 割を超える精度を記録したが, まだまだ改善点はあるため分析が必要である.

2 データセット

本研究ではクイズ大会「abc」の公式問題集を利用している. 公式問題集には 1 つの問題に対し 1 つのジャンルが付与されている. 本研究の目的はマルチラベル分類, すなわち 1 つのデータに対して複数のラベルが付与されているデータセットを用いた分類問題であることから, 再アノテーションを行った. 全 5778 問のうち,4970 問は 1 つ,715 問には 2 つ,93 問には 3 つのラベルが与えられてた.

3 分類モデル

本研究では BERT を用いてマルチラベル分類を行う BertForSequenceClassificationMultiLabel を用いる. これは BERT の文献 [3] で紹介されていたマルチラベル分類のモデルを基にクイズ問題のジャンル分類用に改変したものである. Tokenizer,model は共に cl-tohoku/bert-base-japanese-whole-word-masking という学習済みモデルを用いている. シングルマルチ分類ならば与えられた選択肢から 1 つを選ぶ形式だが, マルチラベル分類では選択肢それぞれを選ぶか選ばないかを決める. 本モデルでは文章が入力されたときに選択肢一つ一つに対して分類スコアを与え, 正値なら選択肢, 負値ならば選択しないという方法で分類を行う. ファインチューニングする際の損失関数は, 分類スコアに Sigmoid 関数を用いることで予測確率としている. これにより予測確率が 50% を超えると選択するということに対応している.

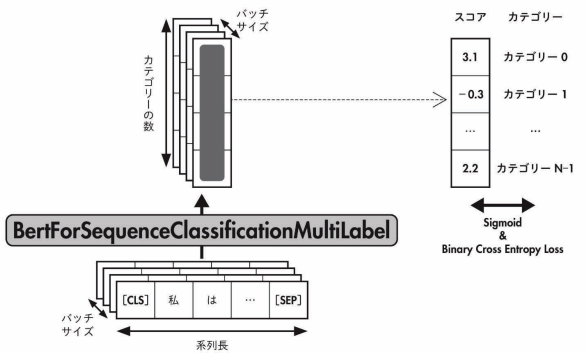

図 1 BertForSequenceClassificationMultiLabel の入出力関係 ([3] より引用)

4 評価指標

4.1 macro-F1,micro-F1

macro-F1 は, 複数のクラスを含むデータセットに対する,F1 スコアを計算する際の方法の一つである. macro-F1 では, 各クラスごとに F1 スコアを計算し, それらを平均して F1 スコアを計算します. macro-F1 は, データセット内でクラスごとのデータ数が均等である場合に有用である. また, 各クラスごとの精度を等しく重視する場合にも,macro-F1 が用いられることがある.

$$macroF1 = \frac{1}{N}(F_{1\ class1} + F_{1\ class2} + ... + F_{1\ classN}) \quad (1)$$

micro-F1 は, 複数のクラスを含むデータセットに対する,F1 スコアを計算する際の方法の一つである. micro-F1 では, 各クラスごとに正解と予測が一致したデータ数を集計し, それらを用いて F1 スコアを計算する. micro-F1 は, データセット内でクラスごとのデータ数が非常に偏りがある場合に, クラスごとに精度が大きく異なる場合に有用である.

$$microF1 = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \cdot recall_N \cdot precision_N}{recall_N + precision_N} \quad (2)$$

4.2 Exact-match ratio

日本語では「完全一致マッチング」と呼ばれる. Exact-match ratio は完全解答のみを正解とするため, 計算式は式 3 となる.

$$ExactMatchRatio = \frac{1}{N}\sum_{i=1}^{N} I[Y_i = T_i] \quad (3)$$

N 個からなるデータにおいて,i 番目の正解を Y_i, 予測を T_i として算出する. I[・・・] は,・・・が真なら 1, 偽なら 0 となる. 0 から 1 の範囲で値を取り, 高いほど良い. しかし, この評価指標では 1 だけ間違えた場合と全く異なる予測をした場合も同じ間違いとカウントしてしまうという弱点がある.

4.3 Partial-match ratio

Exact-match ratio が「完全一致マッチング」と呼ばれるならば,Partial-match ratio は「部分一致マッチング」といえる. Partial-match ratio は, 正解ラベルに含まれるラベルを全て出力したものを正解とする. 計算式は 4 で求めることができる. 0 から 1 の範囲で値を取り, 高いほど良い.

$$PartialMatchRatio = \frac{1}{N}\sum_{i=1}^{N} I|Y_i \supset T_i| \quad (4)$$

5 計算機シミュレーション

5.1 シミュレーション環境

本研究では Google Colaboratory の GPU 環境を用いて実験を行った. シミュレーション環境を表 1 に示す.

表 1　シミュレーション環境

環境	仕様
OS	Ubuntu18.04.6 LTS
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	NVIDIA Tesla T4 (15109MiB)
プログラム言語	Python3.8.16
メモリ容量	13GB
フレームワーク	pytorch 1.13.0 pytorch_lightning 1.6.1

5.2　問題設定

先行研究 [1][2] ではシングルラベル多クラス分類によるクイズ問題のジャンル推定が行われた．本研究においては, 先行研究の結果から得た分類結果からマルチラベルによる分類が必要ではないかという結論に至った．先行研究でも利用していたクイズの問題データに再アノテーションを施した上で, マルチラベル多クラス分類を行い, その性能評価を行う．入力データにはクイズの問題文を BERT の学習済みモデルでテンソル化し, 正解データも 12 次元のバイナリ化する．checkpoint を設定し, 学習の途中で最も損失が少ないタイミングのモデルを保存する．

5.3　パラメータ設定

前章で紹介した BERT モデルは殆どがデフォルトの設定である．その他詳細なパラメータを表 2 に示す．

表 2　ハイパーパラメータの設定値

パラメータ	内容
入力の最大系列長	64
ミニバッチサイズ	32
系列のトリミング方法	末尾からトリミング
学習率	1e-5
エポック数	50

5.4　結果

BERT によるマルチラベル分類の結果を sklearn の classification report を用いて評価した．表 3 にある通り, ほぼ全てのジャンルで 8 割近くの F1 を得ることができた．また, 表 4 では Exact-match ratio と Partial-match ratio も合わせて記した．完全正解を意味する Exact-match ratio で約 8 割の精度を出せており,Partial-match ratio では 9 割の問題で正解ラベルを選択することに成功している．

改善点としては, 言葉の問題のみ recall と F1 が他と比べて低い結果になった．言葉のジャンルは多くの問題でクロスオーバーしていることから recall が低くなったと考察できる．このように, クロスオーバーしているか否かで精度に影響してしまう点は完全する必要がある．

6　おわりに

クイズ大会「abc」の問題集を用いて再アノテーションしたデータセットを用いて,BERT によるマルチラベル分類の研究を行った．結果としては各評価指標で満足の行く数値を得たが, 課題も多く残っている．特に損失関数に Simoid 関数を用いているため,「該当するジャンル：なし」という出力をすることがあったため, これに関してはモデルの調整を行い早急に解決したい課題である．

表 3　BERT モデルでの評価

ジャンル	precision	recall	f1-score	support
音楽	0.85	0.80	0.82	55
日本史	0.84	0.86	0.85	49
世界史	0.88	0.88	0.88	50
生活	0.79	0.78	0.79	170
科学	0.87	0.85	0.86	145
芸術	0.87	0.68	0.76	88
スポーツ	0.92	0.95	0.94	87
言葉	0.83	0.58	0.69	103
文学	0.93	0.87	0.90	75
地理	0.93	0.80	0.86	98
芸能	0.88	0.87	0.88	79
公民	0.85	0.73	0.79	75
micro avg	0.86	0.80	0.83	1074
macro avg	0.87	0.80	0.83	1074
weighted avg	0.86	0.80	0.83	1074
samples avg	0.87	0.84	0.84	1074

表 4　評価指標

評価指標	評価値
macro-F1	0.83
micro-F1	0.83
Exact-match ratio	0.77
Partial-match ratio	0.88

また, 本研究では BERT のモデルを 1 つに決めて実験を行ったが, 他の BERT モデルや更に大規模な GPT,T5 などのモデルでも実験し, それぞれのモデルに差があるのかという検証も研究の余地がある．

参考文献

[1] 淀川 翼, 伊東 栄典:”機械学習手法を用いたクイズ問題のジャンル推定”, 火の国情報シンポジウム論文集, Vol.2020,pp.1-–4 (2020)

[2] 浅野 綾太, 小林 邦和:”クイズ問題のジャンル推定における機械学習手法の比較検討”, 言語処理学会 第 28 回年次大会 (NLP2022), pp.1982–1985 (2022)

[3] 近江 崇宏, 金田 健太郎, 森長 誠, 江間見 亜利 (編集: ストックマーク株式会社):「BERT による自然言語処理入門: Transformers を使った実践プログラミング」, 株式会社オーム社 (2021)