E-030

自然言語処理を用いた文章のカテゴリー分類における比較検討
A Comparative Study on Categorization of Sentences using Natural Language Processing

元木英理香 [1]　高岡詠子 [2]
Erika Motoki　Eiko Takaoka

1. はじめに

10000 個の文章データに対して，分類数を複数設けて文章分類を行った．ネットニュース記事，新聞記事など様々なラベル付き文章データセットを用意した．どのデータセットにおいても，二値分類が最も精度が良く，分類数を増やすとそれに従って精度が下がる，という仮説を立てた．仮説を確かめるために，データを分類数に合わせて整形し，ラベル付き文章データセットを各種アルゴリズム(BERT [1], ALBERT [2], XLNET [3])に当てはめた．得られた結果を元に仮説が検証されたか，分類数，アルゴリズム毎に主だった特徴は見られるのかに関して考察をし，同時にデータセットに見合った分類数，アルゴリズムの提案をする．

2. 提案方法

カテゴリーラベルが付けられた文章データ 5 種類を用意した．各データの詳細は表 1 に示すとおりである．

表 1 比較検討に用いたデータセットの一覧

名前	カテゴリーラベル数	文章データ数	詳細
データ1	93	126418	記事投稿型SNSmediumに掲載された記事のタイトルをカテゴリー別にまとめたもの
データ2	93	126418	記事投稿型SNSmediumに掲載された記事のサブタイトルをカテゴリー別にまとめたもの
データ3	102	1484340	アイルランドの新聞社irish timesで掲載された記事のタイトルをカテゴリー別にまとめたもの
データ4	66	369047	covid19に関するネットニュースのタイトルをサイト別にまとめたもの
データ5	66	369047	covid19に関するネットニュースの記事をサイト別にまとめたもの

それぞれの文章データにおいて，カテゴリーラベルの多い順に並べ，6 通りの分類パターンに必要なデータ数を保持するカテゴリーを選んだ．分類パターンに適した形になるようにデータの抽出，整形を行った．6 通りの分類パターンは以下に示す通りである．

パターン I. 5000 個の文章データ 2 種類による 10000 個のデータ 2 値分類
パターン II. 2500 個の文章データ 4 種類による 10000 個のデータ 4 値分類
パターン III. 2000 個の文章データ 5 種類による 10000 個のデータ 5 値分類
パターン IV. 1250 個の文章データ 8 種類による 10000 個のデータ 8 値分類
パターン V. 1000 個の文章データ 10 種類による 10000 個のデータ 10 値分類
パターン VI. 625 個の文章データ 16 種類による 10000 個のデータ 16 値分類

各パターンで用意した合計 10000 個の文書データのうち，1000 個をテストデータ，9000 個を評価データとした．学習率は 2e-5，エポック数は 4 とした．分類手法の比較検討には，HuggingFace の提供する Transformers ライブラリ [4]より 3 種類のアルゴリズムを用いた．

I. BERT (BertForSequenceClassification)
II. ALBERT (AlbertForSequenceClassification)
III. XLNET (XLNetForSequenceClassification)

表 1 に示す 5 種類のカテゴリーラベル付き文章データを用いて，上記の 6 通りの分類パターンを，3 つのアルゴリズムでそれぞれ試行し，分類精度の比較を行った．

3. 結果

図 1, 2, 3 は，アルゴリズム毎の精度の比較結果である．縦軸が精度で，数値が大きいほど精度が高いことを示す．分類数を増やした場合に精度が下がったのは 5 種類の文章データのうちデータ 3 とデータ 5 の 2 種類みであった．データ 1 とデータ 2 はパターン III の 5 値分類において精度が一番高かった．データ 4 はいずれのアルゴリズムにおいてもパターン II の 4 値分類よりも，パターン III の 5 値分類を行った際に精度が上がっている．このことから，文章データによって最も高い精度の値を叩き出す分類パターンが異なったことが分かる．

4. 考察

結果より，分類数を増やすとそれに従って精度が下がる」という仮説に反して、精度の高い分類パターンがデータによって異なることがわかった．その原因として，今回の試行では 10000 個の文章データを 6 通りの分類パターンで比較検討する際に以下に示す 2 つの懸念点があった．

1 上智大学大学院理工学専攻　Graduate School of Science and Technology, Sophia University
2 上智大学理工学部　Faculty of Science and Technology, Sophia University

第一に，いずれの文章データセットにおいても，カテゴリー毎に文章数にばらつきがあった．例えばデータ 3 では最も多いカテゴリーで 57 万個，少ないカテゴリーでは 100 個のデータを保持していた．6 通りの分類パターンを行う際のデータ抽出において，必要な数に近いカテゴリーのデータを選択し，その上で分類に必要な分のみランダムに抽出をしていた．故に現状では分類パターンによってデータの組み合わせがバラバラである．そこで 6 つの分類パターンに用いるデータの組み合わせをなるべく揃えることが必要と考えられる．

第二に，いずれの文章データセットにおいても，比較検討を行う文章データの長さが統一されていなかった．例えばデータ 4 では多くの文章データが単語数 200 以下であったのに対し、データ 3 はほとんど単語数 50 以下であった．文章データの長さの依存性も考えられることから，文章の長さを統一した状態での比較も行うことが必要とされる．これらの考察を実証するために追実験として新たに 39 のカテゴリーそれぞれに 25000 個の文章データが格納されている文章データセットを用意した．文章長を統一するためにランダムに選んだ 2 つのカテゴリーそれぞれ指定された文章長のデータのみを格納したファイルを作る．その上で 39 のカテゴリーからまずパターン I の分類を実行するために文章データを 5000 個抽出し，2 値分類を行う．その後パターン II の分類のために，パターン I で用いた 2 つのカテゴリーと，新たに 2 つのカテゴリーからそれぞれ 2500 個抽出し 4 値分類を行う．このように用いるカテゴリー数を徐々に増やして実験をする．追実験は現在実験中である．発表時にはその結果を報告する予定である．

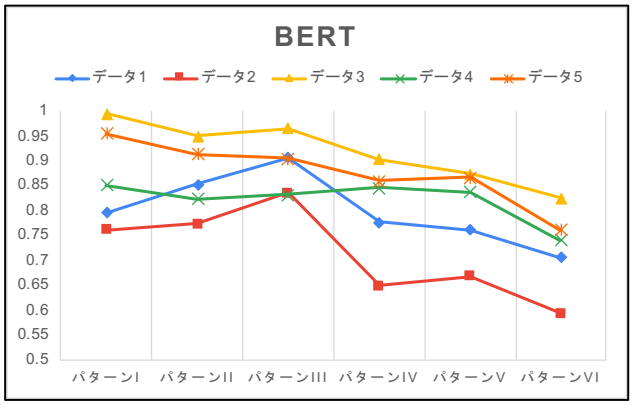

図 1 BERT を用いて 5 つのデータセットの比較を行った結果

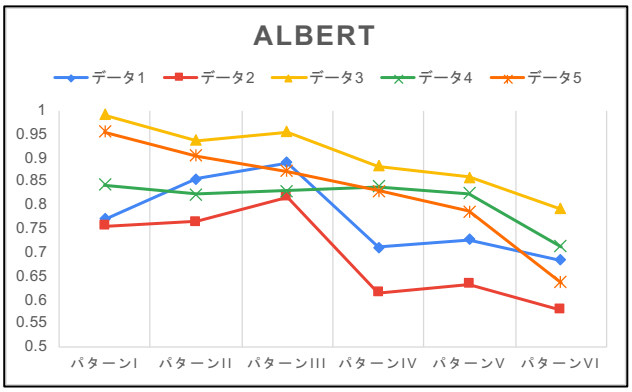

図 2 ALBERT を用いて 5 つのデータセットの比較を行った結果

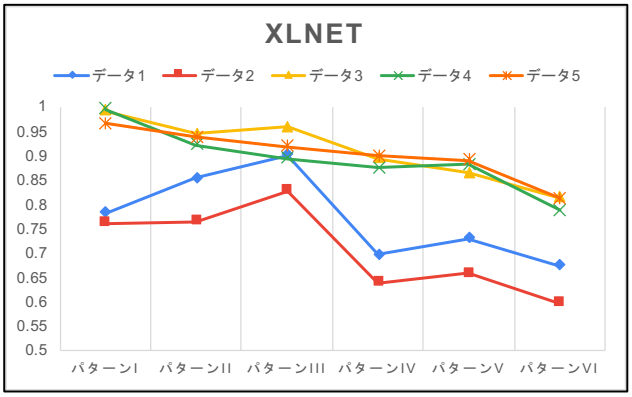

図 3 XLNET を用いて 5 つのデータセットの比較を行った結果

参考文献

[1] D. Jacob, M.-W. Chang, K. Lee , K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Google AI Language, 2018.

[2] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma , R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations,” Google Research; Toyota Technological Institute at Chicago, 2019.

[3] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov , Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” Google AI Brain Team; Carnegie Mellon University, 2019.

[4] Huggingface, "Transformers," [Online]. Available: https://huggingface.co/transformers/index.html.