氏名:鈴木刀磨， 試験区分:情報科学区分， 希望研究室:自然言語処理研究室

NAIST において取り組みたい研究について

1　これまでの修学内容について

私はこれまでデジタルプラットフォーマー(DPF)にまつわる独占禁止法の事例研究を行ってきた。DPF とは情報技術やデータを活用して第三者に「場」を提供する事業者であり、代表的なものとして Google や Amazon、Facebook が挙げられる。DPF は効率性や安全性の面で利用者に多大な便益をもたらす一方、情報の非対称性から DPF 利用者が不利益を被るおそれのある場面が現出するようになった。例えば Amazon が運営する EC サイトに利用者が商品を出品する際に、他のEC サイトと同等かそれ以下の価格に設定することを義務付けたこと事例がある[1]。Amazon による同行為は公正な競争を阻害するおそれがあるとして公正取引委員会が Amazon に対し事前告知を行い、Amazon は同行為を撤回する旨の確約計画を申請した。DPF の独占状態による利用者の不利益を回避するために、DPF に対し行政が主体となり法的対策を講じることが求められている。

2　NAIST において取り組みたい研究分野

2.1　はじめに

奈良先端科学技術大学院大学で取り組みたい研究テーマは「NLP モデルの軽量化」である。より具体的には Transformer 言語モデルの埋め込み層削減による軽量化を研究する。

2.2　研究の背景

Transformer [2]が発表されて以降、BERT [3]や GPT-3 [4]といった大規模な事前学習済み言語モデルが様々なタスクで State-of-the-art を更新している。しかし大規模事前学習言語モデルは極めて高いパラメータ数を持っており、学習や推論には高い計算能力を必要とする。オフライン環境下やエッジデバイスでの使用を念頭に置くと、精度をなるべく下げずにモデルサイズを削減する技術は重要となる。

また、モデル圧縮は DPF の IT 市場の寡占を解消する意義も持つ。DPF が高い競争力を持つ一因は、現代の IT 技術を支えるサーバ等のデバイスが極めて高価であることによって生じる高い参入障壁にある。したがって中小規模の事業者は IT 技術の導入や運用に関して DPF 等のテック企業に全面的に依存することになる。モデルの軽量化を進め、安価なデバイスでも動作を可能にすることは、DPF の IT 市場の寡占状態を解消する上で一定の意義があろう。

2.3　関連研究・課題

2.3.1　関連研究

NLP モデルの軽量化には、モデル圧縮によるものと、設計段階からモデルサイズを抑えるものがある。

既存の事前学習済みモデルの圧縮の取り組みとして代表的なものとしては量子化 [5]、枝狩り [6]、知識蒸留(KD) [7]がある。量子化はパラメータをより小さいビットで表現することでメモリ使用量を減らす手法である。また枝狩りは影響の小さいノードを削除しパラメータを削減する。量子化と枝狩りはパラメータ削減に重きを置くアプローチと言える。KD は大規模な教師モデルの出力を模倣するように小さな生徒モデルを学習させる手法であり、精度向上に重きを置くアプローチである。DisitilBERT [8]では BERT の Transformer 層を大幅に削減した生徒モデルに KD を行い、高精度モデルを作成した。TinyBERT [9]では事前訓練とファインチューニングの 2 段階で KD を行うことで高い精度を保持している。MobileBERT は BERT の隠れ層の次元を低くした生徒に KD を行うことで軽量化と精度の両立を実現した。

次に設計段階からサイズを抑える手法として ALBERT [10]と ALONE [11]がある。ALBERT は単語埋め込み表現の次元と隠れ層の次元を分解することで学習時の軽量化を実現している。また ALONE は機械翻訳タスクで提案された手法であり、共有埋め込みですべての単語を表現し、単語埋め込みのパラメータ数を削減している。

2.3.2　課題

上記で紹介した各手法を組み合わせる可能性は示唆されている [11]ものの、組合わせの効果についての横断的な研究は少なく、未だ明らかになっていない。また各手法の適用の順番によっても効果に違いがあることを示す研究 [12]もあり、今後の精査が必要だ。

また小型化したモデルの多くでは、総パラメータ数に対する埋め込み層の割合が高くなっており、さらなる小型化のために埋め込み層を削減することが課題となる。

2.4　研究手法

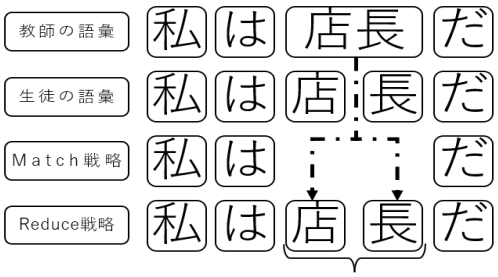
埋め込み層を削減した小型モデルに、KD を行うことで精度も両立させることを目指す。教師モデルとして東北大学の乾研究室が公開している、BERT-large モデル [13]を利用する。また生徒モデルとして Transformer 層、隠れ状態の次元、語彙のサイズをそれぞれ変えたモデルを複数準備する。

研究手順は大きく 2 ステップに分けられる。まず、(1)埋め込み層を削減した生徒モデルに KD を行う。次に、(2)生徒モデルを教師モデルと比較する。

2.4.1　埋め込み層の削減と KD

語彙を減らした生徒モデルに KD を行う。KD は生徒モデルが教師モデルの語彙予測の確率を学習するために行われるため、教師モデルと生徒モデルの語彙が違う場合、KD をナイーブに適用することはできない。そこで配列長次元アライメントを用いた蒸留 [14]を日本語に適用することから始める。同研究では、生徒と教師で完全に一致する語彙を関連付け(Match 戦略)、教師と生徒で異なる分割がされたサブワードにおいて生徒の中間/出力表現を集約し教師のサブワードに対応させる(Reduce 戦

略)ことで KD を可能にする。



しかし、同研究はロシア語を対象としているため、日本語に適用する際に考慮すべき点を二点述べる。

A) 日本語は単語同士の区切りが明確でないため、使用する Tokenizer によってサブワード化における差異が大きい。そのため同一の単語列に対し教師のトークン列より生徒のトークン列の方が短いとは限らず、左から貪欲に並べる Reduce 戦略が適用できない。

B) 文字種、字体(旧字体)、送り仮名などの違いによる表記ゆれが発生し、処理すべき語彙が不必要に増加する [15]

そして上記の対応策として以下を考案する。

A) K.Alina らの手法 [14]は、生徒の配列に合わせるために教師の配列を分割する一方向的なものである。そこで同手法を拡張し、Reduce 戦略で一致しない場合には、生徒のトークンをさらに分割し、再度配列を合わせるという双方向的な配列アライメントによって対処が可能になると考える。しかし Tokenizer が違ってもトークン化への影響が少ないことを示唆する研究 [16]もあり、どの程度の影響があるかは精査する必要がある。

B) UniDic [17]等の解析用辞書を用いて表記ゆれをまとめあげることは考えられる。しかしまとめあげにより処理が増大し、また「父」と「ちち」のように表記ゆれ自体がニュアンスを含む [18]こともある。そのため表記ゆれについては特別な対応は行わず、タスクの必要に応じて調整するのが良いと考えた。

2.4.2 比較方法

日本語理解ベンチマークである JGLUE [19]を使用する。パラメータ数・推論速度・精度を教師モデルと比較し、タスクごとに生徒モデルの性能を調査する。また量子化や枝狩りを重ねた効果も検証する。

簡単のためパラメータ数と推論速度が相関すると仮定すると、生徒モデルの性能比較結果は以下のように分類される。

I. 推論速度が向上し、教師モデルと精度が同等

II. 推論速度が向上し、教師モデルより精度が悪いが、語彙を削減しない蒸留モデルと精度が同等

III. 推論速度が向上し、教師モデル、語彙を削減しない蒸留モデルより精度が悪い

ⅠもしくはⅡに該当する生徒モデルに対し量子化や枝狩りを重ねて適用し、さらなる性能の向上を調査する。

2.5 予想される結果・今後の課題点

直接比較はできないものの、ロシア語で行われた K.Alina らの研究 [14]に比べ、日本語の生徒モデルの埋め込み層パラメータ数（語彙数）は大きく、精度は悪くなることが予想される。それは 2.4.1 で指摘した A,B に因るところが大きいと考える。汎用言語モデルの圧縮において A,B の影響を完全に取り除くことは困難だが、タスク特化モデルにおいては固有表現と知識ベースを紐づける [20]などタスクに応じた調整が推奨される。

以下に今後の課題を述べる。本研究では BERT をベースに KD を行う。しかし ALONE など、埋め込み層の構造が BERT とは大幅に異なるモデルに対しても KD を行う手法を研究することはモデルのさらなる圧縮において重要である。この問題をさらに一般化すると、どの埋め込み層を持つモデル間でも KD を行うための汎用的な手法も研究の必要性があるといえる。その方向性として、CANINE [21]のように、全ての語彙を文字単位に分割し再構成することも可能性があろう。しかし計算量が膨大になる点や、単語ベースでの意味が失われ得ることについては考慮が必要である。

参考文献

[1] 公正取引委員会, "(令和 2 年 9 月 10 日)アマゾンジャパン合同会社から申請があった確約計画の認定について," 2020. [オンライン]. Available: https://www.jftc.go.jp/houdou/pressrelease/2020/sep/200910.html. [アクセス日: 8 6 2022].

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser , I. Polosukhin, "Attention Is All You Need," NIPS, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee , K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Association for Computational Linguistics, 2018.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child , A. Ramesh, "Language Models are Few-Shot Learners," NIPS, 2020.

[5] Y. Gong, L. Liu, M. Yang , L. Bourdev, "Compressing deep convolutional networks using vector quantization," arXiv, 2014.

[6] S. Han, J. Pool, J. Tran , W. J. Dally, "Learning both weights and connections for efficient neural network," NIPS, 2015.

[7] H. Geoffrey, V. Oriol , D. Jeff, "Distilling the knowledge in a neural network," NIPS, 2015.

[8] V. Sanh, L. Debut, J. Chaumond , T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," NIPS, 2019.

[9] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang , Q. Liu, "TinyBERT: Distilling BERT for Natural Language Understanding," Association for Computational Linguistics, 2019.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma , R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," ICLR, 2019.

[11] S. Takase , S. Kobayashi, "All Word Embeddings from One Embedding," NIPS, 2020.

[12] C. Xu, W. Zhou, T. Ge, K. Xu, J. McAuley , F. Wei, "Beyond Preserved Accuracy:Evaluating Loyalty and Robustness of BERT Compression," Association for Computational Linguistics, 2021.

[13] 東北大学 乾研究室, "BERT models for Japanese text.," 2021. [オンライン]. Available: https://github.com/cl-tohoku/bert-japanese. [アクセス日: 8 6 2022].

[14] A. Kolesnikova, Y. Kuratov, V. Konovalov , M. Burtsev, "Knowledge Distillation of Russian Language Models with Reduction of Vocabulary," arXiv, 2021.

[15] 山本和英, "日本語の表記ゆれ問題に関する考察と対処," JAPIO, 2015.

[16] 築地俊平 , 新納浩幸, "Tokenizer の違いによる日本語 BERT モデルの性能評価," 言語処理学会, 2021.

[17] 国立国語研究所, "「UniDic」国語研短単位自動解析用辞書," 2021. [オンライン]. Available: https://clrd.ninjal.ac.jp/unidic/. [アクセス日: 8 6 2022].

[18] 宮西由貴 , 山本和英, "表記ゆれのまとめ上げによる統計的機械翻訳の改善," 言語処理学会, 2016.

[19] 栗原健太郎, 河原大輔 , 柴田知秀, "JGLUE: 日本語言語理解ベンチマーク," 言語処理学会, 2022.

[20] 高橋寛治, 甫立健悟 , 奥田裕樹, "ニュース記事中の組織名の曖昧さ解消," 人工知能学会, 2020.

[21] J. H. Clark, D. Garrette, I. Turc , J. Wieting, "CANINE: Pre-training an Efficient Tokenization-Free Encoder," Association for Computational Linguistics, 2021.