Aug AnaloGPT: 大規模言語モデルを用いたアナロジー生成によるデータ拡張

李　宰成　山田　武士
近畿大学
2010370225t@kindai.ac.jp　yamada@info.kindai.ac.jp

概要

データ拡張 (Data Augmentation) とは、新たなデータ収集なしに、訓練データの多様性を増やし、モデル性能を高める手法を指す。この手法は、自然言語処理 (NLP) において、学習データ不足やデータ不均衡問題に対処する上で重要な役割を果たす [1]。しかし、ほとんどのデータ拡張手法は数百例程度のデータセットに対してのみ有効である [2]。本研究では数千例からなるデータセットにおいても学習データ不足に対処可能なデータ拡張として、大規模言語モデルを用いた文のアナロジー生成に基づく手法 (Aug AnaloGPT) を提案する。文のアナロジー生成とは対象ドメインが異なるが文の論理構造や意味の関係性が類似している類比文を生成することである。提案手法を JGLUE の JNLI[3] のデータセットに適用したところ、言い換えに基づく既存のデータ拡張手法を凌ぐ性能向上を確認できた。

1　はじめに

1.1　データ拡張

データ拡張 (Data Augmentation) とは、新たなデータ収集なしに、訓練データの多様性を増やし、モデル性能を高める手法を指す。近年、自然言語処理 (NLP) におけるデータ拡張への需要と関心は高まっている。これは、自然言語処理によって、適応すべきタスクやドメインが増えているからだと考えられる [1]。

しかし、ほとんどのデータ拡張手法は数百例程度のデータセットのような極端にデータ量が少ない場合でのみ有効である [2]。

本研究では数千例からなるデータセットに対する学習データ不足に対処するために、大規模言語モデルを用いた文のアナロジー生成に基づくデータ拡張手法を提案する。文のアナロジー生成とは対象ドメインが異なるが文の論理構造や意味の関係性が類似している類比文を類推によって生成することである。また、提案手法では Gemini Pro[4] を用いてアナロジー生成を行うことで、データを拡張する。

1.2　類推

類推とは知りたいことを、それとよく似た既知のことに対応づけて考えることを指す。認知科学において、知りたいことをターゲットと呼び、既によく知っていることをベースと呼ぶ [5]。そして、ベースの要素をターゲットの要素に対応付けることをベースからターゲットへの写像という。写像を行う際に、ターゲットとベースのドメイン性を考慮し、関係性が類似している要素同士を対応づけることが必要になる。そのため、類比文の関係にある二つの文は論理構造や意味の関係性が類似している。

1.3　アナロジー生成によるデータ拡張

従来の自然言語処理における類推の研究は単語レベルに限定され、文の類推はあまり注目されていなかった。一方近年になり、大規模言語モデルが文のアナロジーを生成できることが示唆されている [6]。そこで、本研究では、大規模言語モデルの Gemini Pro[4] を用いて、類比文を生成することでデータ拡張を行う。JGLU の JNLI[3] データセットに対し、アナロジー生成によるデータ拡張を適用すると、既存の大規模言語モデルを用いた言い換え手法を凌ぐ性能向上を確認できた。

2　関連研究

自然言語処理におけるデータ拡張手法で代表的なものに言い換えベースの手法がある。言い換えとは与えられたテキストの意味を変えずに違う言葉で表現したテキストを生成することである。本研究の提

案手法と類似した言い換え手法に単方向翻訳という、元のテキストを一度他の言語に翻訳し、逆翻訳手法と違って、元の言語に戻すことのない言語を横断した方法がある [7]。この方法は多言語のデータセットでよく用いられ、言語が翻訳によって変換されたとしても、意味は保たれているため言い換え手法となると考えられている。

提案手法のアナロジー生成によるデータ拡張は言語を横断する代わりに、ドメインを横断する。単方向翻訳では意味が保存されるのに対し、提案手法では元々のテキストの論理構造や意味の関係とそれによって生じる印象が保存される。そのため、抽象的な言語理解や曖昧なニュアンスを捉えることが必要なタスクに対するデータセットに対して有効であると考えられる。例えば、レビューテキストからレーティングを予測するレーティングタスクなどの場合、製品カテゴリーをドメインとみなし、各々のレビューテキストが属するドメインをベースドメインとし、属していない他のカテゴリーをターゲットドメインとみなすことで提案手法を適用できる。

また、近年大規模言語モデルを用いた言い換えベースのデータ拡張手法も提案されている [8]。

3 実験

3.1 データセット

提案手法によって拡張するデータセットとして言語モデルの日本語理解能力を評価するベンチマークの JGLUE の JNLI[3] を選んだ。JNLI は自然言語推論のデータセットで、前提文が仮説文に対して、「含意」「矛盾」「中立」の関係のうち、いずれかのラベルが付与されている。このデータセットは現在、訓練データと検証データのみ公開されているため、訓練データから 2500 個のデータをランダムに抽出し、そのうち 1000 個を 訓練データ、1500 個を検証データとし、元々の検証データをテストデータとした。この訓練データに対してデータ拡張を行う。拡張した訓練データをもとに BERT をファインチューニングし、自然言語推論タスクの性能向上を図る。また、BERT には東北大学の乾研究室が公開している bert-base-japanese-v3[1)] を使用する。さらに、比較を行う既存手法として、大規模言語モデルを用いたテキスト言い換えによるデータ拡張 (Aug GPT)[8] もこのデータセットに適用する。また、この既存手法で用いる大規模言語モデルは Gemini Pro である。

表 1 JGLUE の JNLI に対してデータ拡張を行い、訓練データ 1 つあたりの拡張するデータ数の場合分けによる、自然言語推論タスクの正解率

拡張されたデータ数	1	2	3	4	5	6
ベースライン (拡張なし)	0.79	-	-	-	-	-
Aug GPT(言い換え)	80.8	80.5	79.7	80.4	79.0	78.9
Aug AnaloGPT(言い換え)	81.0	81.6	83.3	82.4	83.0	85.5

3.2 提案手法

類推を行う際には、ベースドメインとターゲットドメインを決定する必要がある。本研究で拡張するデータセットでは、大規模言語モデルがターゲットドメインを自ら決めるようにプロンプトを設計した。ゆえに、事前にどのドメインに写像を行うかは決まっておらず、モデルの挙動によって生成される文が大きく変わることが予想される。また、提案手法と既存手法の拡張データサイズの影響を測定するため、6 通り (訓練データ 1 つあたり 1 個から 6 個のデータを拡張する) の実験した。

3.3 結果・考察

表 1 に結果を示す。提案手法の性能はベースラインと既存手法を一貫して上回っている。さらに、データ拡張を行うデータ数が増えるほど性能が向上している。反対に、既存手法は増えるほど性能が下がっている。これは同じ意味の拡張データが増えてしまうと、そのデータはドメインも同じになるので、ドメイン固有の具体性にモデルが過適合していると考えられる。それに対し、提案手法はドメインを横断するため、ドメインの固有性を超えた抽象的な推論能力獲得に成功していると考えられる。

4 おわりに

本研究では、大規模言語モデルを用いたアナロジー生成に基づくデータ拡張手法 (Aug AnaloGPT) を提案し、自然言語処理 (NLP) におけるデータセットの多様性を増加させ、モデル性能の向上を目指しました。この手法は、文の論理構造や意味の関係性が類似している類比文を生成することによって、特に数千例以上のデータセットにおいて有効であることが示されました。JGLUE の JNLI データセットへの適用により、従来の言い換えに基づくデータ拡張手法を上回る性能向上が確認された。

1） https://huggingface.co/cl-tohoku/bert-base-japanese-v3

5 参考文献

参考文献

[1] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A Survey of Data Augmentation Approaches for NLP. In**Findings of the Association for Computational Linguistics: ACL-IJCNLP**,Volume 3, pages 968–988, Online, Association for Computational Linguistics. 2022.

[2] Itsuki Okimura, Machel Reid, Makoto Kawano, Yutaka Matsuo. On the Impact of Data Augmentation on Downstream Performance in Natural Language Processing. In**Proceedings of the Third Workshop on Insights from Negative Results in NLP**,, pages 88–93, Dublin, Ireland. Association for Computational Linguistics,2022.

[3] Kentaro Kurihara, Daisuke Kawahara, Tomohide Shibata. Jglue: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pages 2957–2966, 2022.

[4] G. Team, et al. Gemini: A Family of Highly Capable Multimodal Models. In **arXiv:2312.11805**, 2023.

[5] 鈴木宏昭. 類似と思考. ちくま学芸文庫, 2020.

[6] Cheng Jiayang, et al. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, p. 11518–11537, 2023.

[7] Bohan Li, Yutai Hou, Wanxiang Che. Data Augmentation Approaches in Natural Language Processing: A Survey. **AI Open** ,Volume 3, Pages 71-90, 2022.

[8] H. Dai, et al, AugGPT: Leveraging ChatGPT for Text Data Augmentation. In **arXiv:2302.13007**, 2023.