7ZC-06

自然言語処理におけるデータ拡張の強度とモデル性能の関係

和田 翔熙 †　　森本 尚之 ‡
京都大学 †　　三重大学 ‡

1 はじめに

機械学習においてデータが僅少である場合にはモデルの過学習により、汎化性能が低下してしまう恐れがある。このような場合、元のデータに対してデータの一部に変化を与えることで新たなデータを生成するデータ拡張はモデルの性能向上に有効な手段である。画像処理の分野においては広く利用されており、モデルの性能向上に貢献している。このことから自然言語処理においても EDA [1] などの様々なデータ拡張法が提案されている。

データ拡張戦略の設計は手法の組み合わせ、拡張を実行する確率、拡張の強度の最適化によって行われる。したがって戦略設計には強度の操作が有用である。しかし、拡張強度とモデル性能の関係について調査した文献は少ない。そこで、本研究では自然言語処理におけるデータ拡張の強度として「忠実さ」「多様さ」を利用し、モデル性能との関係に注目して実験を行った。

実験の結果、弱い拡張と強い拡張の併用によりモデル性能が向上するケースが見られた。

2 手法

2.1 拡張強度の指標

拡張強度の評価としてデータ拡張の品質を測るために広く使われる「忠実さ」、「多様さ」を用いた。忠実さは拡張データが元のデータにどれだけ忠実かを指す。よってこの指標として cos 類似度と BERTScore [2] を利用した。多様さは拡張データによってどれだけ語彙が多様になったかを指している。したがって、この指標として BLEU [3] を利用した。

cos 類似度　cos 類似度は文章の類似度を測定する目的で広く利用されており、ベクトル化したテキスト s, s' を用いて計算される。

BERTScore　BERTScore はトークン単位の類似度を計算するものである。

BLEU　BLEU は用いられている語彙の被り具合を計算するものである。

2.2 実験設定

2.2.1 データ拡張手法

データ拡張手法として、テキストの一部を類義語に置換する手法である Synonym Replacement (SR)、テキストの一部の位置を入れ替える手法である Random Swap (RS)、テキストの一部を削除する手法である Random Deletion (RD) を用いた。

2.2.2 データセット

livedoor ニュースコーパス [5] は、9 種類のカテゴリからなるニュース記事のデータセットである。本研究では、このデータセットからサブセットを 10%, 50%, 100% の大きさ（以下, データサイズと呼ぶ）で抽出し、文書分類を行った。

2.2.3 モデル

日本語での事前学習を行なったモデルである BERT [4] を利用して実験を行った。ベースラインとして元のデータのみを学習データとして用いる条件のほかに以下の 4 条件で 5 エポック

Relationship between Data Augmentation Intensity and Model Performance in Natural Language Processing
† Wada Shoki, Kyoto University
‡ Morimoto Naoyuki, Mie University

表 1　データサイズ 100% の実験結果

手法	r20	mix	t20	b20	base
RD	0.862	0.879	0.862	0.872	0.875
RS	0.901	0.899	0.894	0.897	0.881
SR	0.882	0.887	0.890	0.882	0.881

表 2　データサイズ 50% の実験結果

手法	r20	mix	t20	b20	base
RD	0.850	0.844	0.847	0.847	0.830
RS	0.857	0.842	0.850	0.847	0.844
SR	0.851	0.856	0.863	0.854	0.857

の学習を行なった。

top20 条件　元のデータに加えて、拡張データとして強度が弱いもの上位 20% をデータに加えたものを学習データとする

bottom20 条件　元のデータに加えて、拡張データとして強度が弱いもの下位 20% をデータに加えたものを学習データとする

mix 条件　元のデータに加えて、拡張データとして強度が弱いもの上位 10% と下位 10% をデータに加えたものを学習データとする

rand20 条件　元のデータに加えて、拡張データとして強度に関係なく拡張データの 20% をデータに加えたものを学習データとする

3　結果

実験の結果、弱い拡張だけでなく、強い拡張も利用することで性能が向上するケースが見られた。誌面の都合上、拡張強度の指標として cos 類似度を用いた場合についてのみ言及する。

表 1 と表 2 は、データサイズをそれぞれ 100%，50% とした際のモデル性能の評価 (Accuracy) である。

沖村ら [6] が主張するように、弱い拡張がモデル性能向上に寄与する傾向が本研究でも見られた（表 2)。しかし、表 1 で示すように一部のケースでは top20 条件よりも mix 条件で学習を行なった方が性能向上している。

このことから、拡張強度が弱いものだけが学習に有用なのではなく、強度が強いものと弱いものを混ぜたものも学習に対して有用であることが示唆された。

4　おわりに

本研究では自然言語処理におけるデータ拡張の強度として「忠実さ」「多様さ」を利用し、モデル性能との関係に注目して実験を行った。

実験の結果、一部のケースでは弱い拡張だけでなく強い拡張も利用することでモデルの性能が向上することが確認できた。

今後は様々なモデル、データセット、タスクでも同様の結果が得られるのか調査を続ける必要がある。

参考文献

[1] Jason Wei et al., “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”, Proc. EMNLP-IJCNLP, 2019.

[2] Tianyi Zhang et al., “BERTScore: Evaluating Text Generation with BERT”, Proc. ICLR, 2020.

[3] Kishore Papineni, “BLEU: a Method for Automatic Evaluation of Machine Translation”, Proc. ACL, 2002.

[4] 東北大学 乾・鈴木研究室. Pretrained Japanese BERT models, https://github.com/cl-tohoku/bert-japanese.

[5] NHN Japan 株式会社, “ライブドアニュースコーパス”, http://www.rondhuit.com/download.html\#ldcc (参照 2023-11-20)

[6] 沖村樹ほか,“自然言語処理におけるデータ拡張による性能改善への影響分析”, 人工知能学会全国大会論文集, 2022.