隠れ層補間によるデータ拡張を用いた障害レポート分類

山下 郁海[1] 小町 守[1] 真鍋 章[2] 谷本 恒野[2]
[1] 東京都立大学 [2] 富士電機株式会社
yamashita-ikumi@ed.tmu.ac.jp, komachi@tmu.ac.jp
{manabe-akira, tanimoto-kouya}@fujielectric.com

1 はじめに

近年，小規模なラベルありデータしか存在しないテキスト分類に対する研究が盛んに行われている [1, 2, 3]．これらの手法では対象となる分野のラベルなしデータを用意し，事前学習やデータ拡張に用いることで，大きな性能の向上が報告されている．

一方，本研究で扱う日本語の障害レポートはラベルありデータが少なく，また対象となる分野のラベルなしデータの用意も権利の関係で難しいデータである．また，障害レポートは特定の分野に関しての記述であり，例えば Wikipedia のような大規模なデータが入手可能なラベルなしデータとは大きく分野が異なる [4]．そのため，対象となる分野，もしくはそれに近い分野のラベルなしデータが必要になるような手法を適用することが難しい．

そこで本研究では，Chen ら [5] による隠れ層線形補間を用いたデータ拡張手法 TMix を用いて実験を行った．TMix は，入力文の隠れ状態ベクトルを線形補間することでデータ拡張を行う，追加のデータを必要としないデータ拡張手法であり，本研究で扱う障害レポートの分類問題に対しても適用が比較的容易である．

TMix を用いて障害レポートに対して分類実験を行った結果，TMix を用いない場合と比較して性能の向上が確認された．また，分類の対象とは異なる分野のラベルなしデータを用意し，TMix をもとにした半教師あり学習を行うことで，さらなる性能の向上を確認し，対象分野のラベルありデータと，対象となる分野とは異なる分野のラベルなしデータを線形補間して学習に用いることで性能向上が可能であることを示した．

2 関連研究

2.1 線形補間によるデータ拡張

線形補間によるデータ拡張は，近年画像分類の分野において盛んに研究されている．Zhang ら [6] は入力画像のベクトルとそのラベルをそれぞれ線形補間した上で学習を行うことで，性能向上を図る手法 mixup を 2017 年に提案している．その後，Berthelot ら [7] は mixup を用いた半教師あり画像分類手法 MixMatch を提案し，Verma ら [8] はこれまで入力のベクトルのみに行っていた線形補間を，隠れ状態ベクトルに拡張することで性能が向上することを報告している．

これらの研究はどれも線形補間を用いたデータ拡張の研究ではあるが，本研究とは異なり，全て画像分類の分野の研究であり，本研究では，これらの研究をもとに Chen ら [5] の提案した，テキストデータのための線形補間を用いたデータ拡張手法を用いて実験を行う．

2.2 小規模なラベルありデータに対するテキスト分類

小規模なラベルありデータに対するテキスト分類のための手法のうちの 1 つとして，ラベルなしデータを用いた事前学習の研究が行われている．Gururangan ら [3] は，分類の対象となる分野のラベルなしデータを用いて，Variational autoencoder を事前学習し，そこから得られる embedding を入力文の単語ベクトルと足し合わせて，テキスト分類の入力として用いることで分類性能の向上を図っている．

また，別の手法として，ラベルなしデータを用いた半教師あり学習の研究も行われている．Chen ら [1] と，Xie ら [2] は，どちらも対象となる分野の

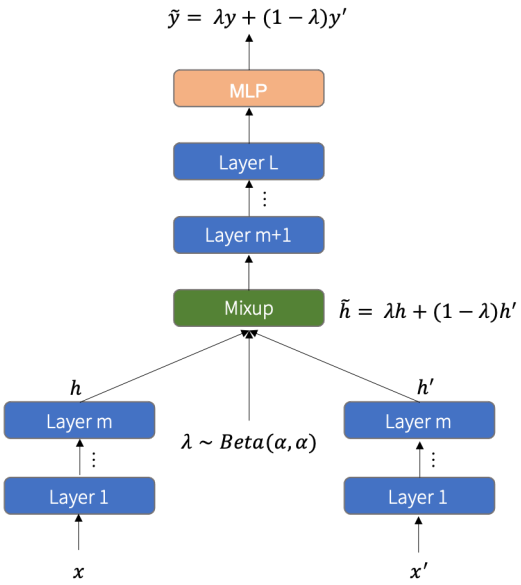

図 1　TMix [5] の概要図.

ラベルなしデータに対して逆翻訳をすることでデータ拡張を行い，それらのラベルなしデータを用いた半教師あり学習の手法を提案している．これらの研究は本研究とは異なり，どれも性能向上のために対象となる分野のラベルなしデータが入手できることを想定している．

3　障害レポート分類のためのデータ拡張

3.1　隠れ層の補間：TMix

本研究では，障害レポートの分類問題を解くにあたって，画像分類の分野におけるデータ拡張手法 mixup [6] をもとに，Chen ら [5] がテキストデータを扱えるように改良したしたデータ拡張手法 TMix を用いている．TMix の概要図を図 1 に示す．TMix は，BERT [9] のような多層モデルを用いる際に，入力文の隠れ状態ベクトルを線形補間することでテキストデータの拡張を行う手法である．図 1 では，まず 2 つの異なる文 x, x' を入力とし，第 m 層まで通常の手順でそれぞれの隠れ状態ベクトル h, h' の計算を行う．次に，2 つの文の隠れ状態ベクトルを線形補間し，以下の式に示す $\tilde{h}$ を得る．

$$\tilde{h} = \lambda h + (1-\lambda)h'$$

ここで補間係数 λ はベータ分布 $Beta(\alpha, \alpha)$ に基づいて選択される．α は補間係数をコントロールするためのハイパーパラメータである．その後の層での計算は，この新しく得た隠れ状態ベクトル $\tilde{h}$ を用いて行う．また，ラベルに関しても隠れ状態を線形補間する際に用いた係数と同じ補間係数 λ を用いて補間を行い，以下の式に示す $\tilde{y}$ を計算する．

$$\tilde{y} = \lambda y + (1-\lambda)y'$$

ここで y, y' はそれぞれ入力 x, x' に対応するラベルである．

学習データ中の入力文の組み合わせは無数に存在するため，TMix を用いることで，追加のデータを用いることなく，新しく拡張されたデータを多数作成することが可能である．また，学習文に対して全てのデータを補間するだけでなく，ランダムで補間するかどうかを選択することで，一部のデータに対して元のデータの隠れ状態を用いて学習を行うことも可能である．そのため本研究では，全てのデータを補間する設定と，補間するかどうかをランダムに選択し一部データを補間する設定の 2 つの設定で実験を行なっている．

3.2　TMix を用いた半教師あり学習

Chen ら [5] は TMix によるデータ拡張とラベルなしデータを組み合わせたテキスト分類に対する半教師あり学習の手法を提案しており，本研究でもそれにならい半教師あり学習を行う[1)]．

この手法ではまず，ラベルなしデータに対してその時点での分類器でラベルの予測を行う．その後，ラベルを予測したこれらのデータを追加のラベルありデータとして，通常のラベルありデータと同様に TMix を用いて扱う．この際に，ラベルありデータをもとに補間を行った場合と，新たにラベルを予測したラベルなしデータをもとに補間を行った場合で，loss の計算が異なる．すなわち，ラベルありデータを用いた場合は Cross-entropy が最小となるように，ラベルなしデータを用いた場合は，線形補間された入力文の隠れ状態ベクトルからモデルの予測した確率分布と，線形補間されたラベルの間の KL ダイバージェンスが最小となるように学習を行う．

また，モデルがラベルなしデータのラベルを予測する際に確信度の高いラベルを予測できるように，予測確率のエントロピー最小化を行い，以下のような Self-training loss を計算する．

$$L_{\text{self}} = \mathbb{E}_{\mathbf{x} \in \mathbf{X}_u} \max(0, \gamma - ||\mathbf{y}_u||_2^2)$$

[1)] Chen らの論文 [5] の記述と公開されている実装 (https://github.com/GT-SALT/MixText) には一部異なる部分や論文では言及されていない箇所があり，この節の記述は公開されている実装をもとにしている．

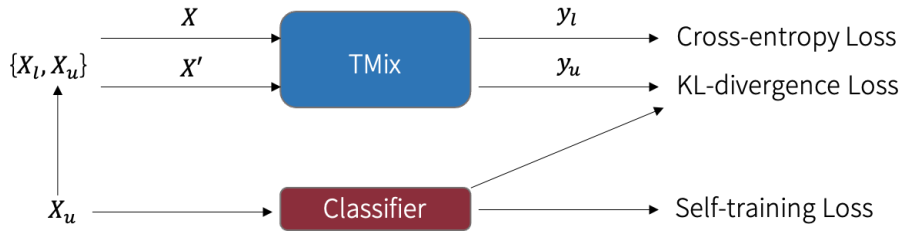

図 2　半教師あり学習の概要図.

表 1　障害レポートの各文とそれに対応するラベルの例

障害レポートの文	状況	原因	措置	その他
A 棟 1F の感知器より警報発報。	✓			
本箇所の感知器が頻繁に誤作動しているとの情報あり。				✓
原因は汚れだと思われるため、復旧及び感知器の交換実施。		✓	✓	

ここで，$\mathbf{X}_u$ はラベルなしテキストデータの集合であり，$\mathbf{y}_u$ はラベルなしデータに対して予測されたラベルである．また，γ は loss を調整するためのハイパーパラメータである．最終的な半教師あり学習の概要は図 2 のようになる．ここで，$\mathbf{X}_l$ はラベルありテキストデータの集合であり，$\mathbf{y}_l$ はラベルありデータに対して予測されたラベルである．

また，この手法ではラベルありデータとラベルなしデータを両方用いて補間を行うが，この際の補間のやり方に関して，ラベルありデータとラベルなしデータを混ぜて補間する設定と，別々にそれぞれのデータの中で補間する設定の 2 つの設定が考えられる．本研究では，この 2 つの設定両方で実験を行い，比較を行なっている．

3.3　障害レポート分類

本研究では障害レポートの各文に対してマルチラベル分類を行う．表 1 に実際の障害レポートの文の例を示す。ラベルには**状況，原因，措置，その他**の 4 つが存在し，各文に対してそれぞれ付与されている．例えば，1 行目の文に対しては状況のラベルが，2 行目の文に対してはその他のラベルが付与されている．一方，3 行目の文には原因と措置の 2 つのラベルが付与されている．このように，本研究で扱う障害レポートデータはマルチラベルデータである．

本研究では，マルチラベル分類の手法として勝又ら [4] と同じ，One-vs-Rest を用いた．分類モデルには BERT を用いた．具体的には，BERT の最終層に対して平均プーリングを行い，その後 2 層，128 次元からなる多層パーセプトロンを通し，その出力を用いて，各ラベルを付与するかしないかの確率を求める．この確率を用いて各ラベルごとに 2 値分類を行い，入力 x に対して各ラベルを付与するかどうかを予測する．

4　実験

4.1　実験設定

本研究では、設備保全に関する障害レポートのデータに対して実験を行う．このデータは今回新たに人手でアノテーションを行ったデータであり，データ中の障害レポートは 194 件存在し，総文数は 1,365 文である．この障害レポートを分割し，学習データに 869 文，開発データに 244 文，評価データに 252 文用いた．

半教師あり学習に用いるラベルなしデータには，Wikipedia を用いた．実際に学習に用いるデータは，元のデータからランダムに 2,500 文を抽出したものを用いた．このデータは，今回対象となる設備保全の障害レポートとは分野の異なるラベルなしデータである．これらのラベルありデータ，ラベルなしデータの単語分割には，共に MeCab [2] を用い，辞書として mecab-ipadic-2.7.0-20070801 を用いた．

本研究では事前学習済みの BERT を単にラベルありデータでファインチューニングし，分類を学習したものをベースラインとする．

本研究で使用する事前学習済み BERT モデルは Hugging Face[3] による自然言語処理ライブラリ Transformers[4] において，東北大学の鈴木らが公開している事前学習済み BERT[5] のモデルを使用した．分類には，Chen ら [5] の実装[6] をもとに，マルチラ

[2] http://taku910.github.io/mecab, v.0.996.
[3] https://huggingface.co/
[4] https://github.com/huggingface/transformers
[5] https://github.com/cl-tohoku/bert-japanese
[6] https://github.com/GT-SALT/MixText

表 2 分類実験の結果．太字はその列内で最もスコアが高いものを示す．

	状況			原因			措置対策			その他			
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F	平均 F
BERT	76.9	58.2	66.0	63.3	71.6	67.1	60.3	58.4	59.1	76.6	81.1	78.8	67.7
TMix	81.5	50.0	61.6	67.9	72.7	69.6	70.9	66.7	68.7	77.3	83.5	80.2	70.0
TMix (rnd)	**87.3**	53.1	65.1	63.2	83.0	71.7	72.4	67.9	69.9	80.3	83.5	81.9	72.1
TMix +Wiki	75.3	**63.3**	**68.7**	**68.5**	80.7	74.0	**74.1**	71.5	**72.2**	**81.3**	**84.7**	**82.9**	**74.4**
TMix (rnd) +Wiki	81.1	56.1	66.3	67.4	70.4	68.7	68.4	66.7	67.5	79.2	83.9	81.5	71.0
TMix (sep) +Wiki	81.8	58.2	67.9	66.7	77.3	71.6	67.0	**75.0**	70.8	81.0	83.5	82.2	73.1
TMix (rnd, sep) +Wiki	81.4	53.1	64.2	67.8	**86.4**	**75.7**	63.3	65.5	64.4	76.2	83.1	79.5	70.9

ベル分類が行えるように修正を行った．TMix を用いた学習の際の線形補間を行う層については，先行研究の結果をもとに 7, 9, 12 層目からランダムに選択を行った．また，言及していないハイパーパラメータについては元の実装のものを用いた．

本研究では，マルチラベル分類を各ラベルに対して付与されるかされないかの 2 値分類として扱い実験を行っている．そのため，評価は各ラベルごとに行った．具体的には，各ラベルに対して正しくラベルが付与できた場合を正解として，Precision, Recall，F-score を計算した．モデルの選択は，開発データに対して各ラベルに対する F-score の平均が最も大きいモデルを最も良い分類モデルとして行った．また，全ての実験結果はシード値を変更して 2 回実験を行い，平均したものである．

4.2 実験結果

実験結果を表 2 に示す．rnd と書かれているものは，学習中にデータに対して補間を行うかどうかをランダムで選択するモデル，sep と書かれているものは，学習中にラベルありデータとラベルなしデータをそれぞれ別々に補間するモデルである．また，+Wiki と書かれているものはラベルなしデータとして Wikipedia のデータを用いて半教師あり学習を行ったものである．

表 2 から，TMix を用いたモデルの平均 F がベースラインの BERT の平均 F を上回っており，隠れ層線形補間によるデータ拡張を用いることで，性能が向上していることが読み取れる．また，ラベルなしデータを用いて半教師あり学習を行ったモデルの全てがベースラインの BERT よりも平均 F が高く，少量かつ対象となる分野とは異なる分野のラベルなしデータを用いているのみにもかかわらず，性能が向上していることが確認できる．加えて，その中でも，Wikipedia のデータを用いて半教師あり学習を行ったモデル（TMix +Wiki）の平均 F が最も高いことが確認できる．

5 分析と考察

表 2 の実験結果から，ラベルありデータのみしか用いない場合（TMix, TMix (rnd)），学習中にデータに対して補間を行うかどうかをランダムで選択する方が性能が向上していることが確認できる．一方，同様のラベルなしデータを用いたモデルの中では，全ての入力データをまとめて補間するモデル（TMix +Wiki）の平均 F が最も高いことがわかる．これらの結果から，信頼度の高いラベルありデータのみで線形補間を行う場合は，全てのデータではなく一部のデータで補間を行う方が性能向上に役立つのに対して，信頼度の低いラベルなしデータはそれ単体で用いるのではなく，常にラベルありデータと補間して用いることで，より分類モデルの性能を向上させることができると考えられる．

6 おわりに

本研究では，対象となる分野のラベルありデータが少ないテキスト分類において，隠れ層線形補間を用いたデータ拡張を行うことで，追加のデータを用いることなく性能が向上することを示した．また，隠れ層線形補間をもとにして少量のラベルなしデータと共に半教師あり学習を行うことで，さらなる性能の向上が図れることを示した．

参考文献

[1] Jiaao Chen, Yuwei Wu, and Diyi Yang. Semi-supervised models via data augmentation for classifying interactive affective responses. In *AffCon, AAAI*, Vol. 2614, pp. 151–160, 2020.

[2] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. In *NeurIPS*, p. 14, 2020.

[3] Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. Variational pretraining for semi-supervised text classification. In *ACL*, pp. 5880–5894, 2019.

[4] 勝又智, 小町守, 真鍋章, 谷本恒野. 障害レポートの分類問題に対するデータ選択を用いた BERT モデルの精度向上. 言語処理学会 第 26 回年次大会, pp. 645–648, 2020.

[5] Jiaao Chen, Zichao Yang, and Diyi Yang. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *ACL*, pp. 2147–2157, 2020.

[6] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, Vol. abs/1710.09412, , 2017.

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. MixMatch: A holistic approach to semi-supervised learning. In *NeurIPS*, Vol. 32, pp. 5049–5059, 2019.

[8] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, Vol. 97, pp. 6438–6447, 2019.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.