文書分類におけるテキストノイズおよびラベルノイズの影響分析

池田 大志　　藤本 拓　　吉村 健
株式会社 NTT ドコモ
{taishi.ikeda.fz, fujimotohir, yoshimurat}@nttdocomo.com

1 はじめに

近年，自然言語処理技術を活用した製品やサービスが世の中に広く導入されつつある．例えば，文書データに対してラベルを付与する文書分類技術を応用することで，アンケートの内容を自動分類し，ユーザーの評判分析に利用することができる [1, 2].

文書データを自動分類するためには，まず文書分類モデルを作成する必要がある．文書分類モデルは，文書データと正解ラベルの対となるデータから作成される．この正解ラベルは，クラウドソーシングの利用やアノテーターを採用し，作業者が文書データの内容を確認することで，事前に定義したアノテーション基準に従って作成される．

また，上記の手順ではなく，文書分類用のデータ作成を目的とせず，例えば，既にグループ化されたアンケートなど，企業活動の中で蓄積された文書データから文書分類モデルを作成する場合，データ中には文書分類モデルの精度低下の原因となるノイズが含まれている可能性がある．このようなノイズを含むデータは，アノテーションの一致率が高い研究用のデータとは異なり，たとえ最高性能の文書分類モデルを用いて学習したとしても，期待通りの分類精度を得ることができない場合がある．また，分類精度向上のためにエラー分析を行ったとしても，何が原因で期待通りの精度が出ないのか，原因を追求することが難しい場合がある．これは，大きく分けて二種類のノイズが原因だと考えられる．一つ目は，運用上毎年更新されるラベル付与の基準の変更やラベル付与者の経験の差により引き起こる**ラベルノイズ**，二つ目は，音声認識経由で書き起こされた音声認識誤りを含む文書や OCR により変換された書類に変換誤りが含まれる**テキストノイズ**である．

文書分類に関する研究では，文書分類モデルを改善し，分類精度向上を報告する研究は多く存在するが，分類精度低下の原因を報告する研究は少ない [3, 4]．そこで，本研究では，分類精度は文書分類モデルだけではなく，学習データにも大きく依存すると考え，学習データに多様なノイズを混入することで，どのような状況下で分類精度は低下するのか，実用性の高い SVM および最高性能の BERT を用いて，その原因を分析する．

実験では，人工的に生成したノイズを学習データに混入させることで，文書分類モデルの正解率がどのように変化するのか検証を行う．実験の結果，ラベルノイズでは文書分類モデルによらず分類精度を低下させ，テキストノイズでは文書分類モデルによりノイズの影響の受けやすさが異なることがわかった．

2 文書分類モデル

本節では，本研究で用いる二種類の文書分類モデルの詳細ついて述べる．

2.1 SVM

一つ目は，Wang ら [5] により提案された Support Vector Machines (SVM) による文書分類モデルである．彼らの手法では，SVM の素性として，単語ユニグラムと単語バイグラムを利用しており，さらに各クラスの割合を素性の重みに与えることで，分類精度向上を実現している．手法はシンプルながらも，高速な学習が可能であり，強力なベースラインとして利用することができる．そのため，本研究では，実用的な文書分類モデルとして SVM を採用した．ここでは，scikit-learn[1] の LinearSVC を用いて手法の再実装を行った．

2.2 BERT

二つ目は，Devlin ら [6] により提案された Bidirectional Encoder Representations from Transformers (BERT) による文書分類モデルである．BERT は大規模な生コーパスで MASK language model と Next

[1] https://scikit-learn.org/stable/

sentence prediction model を pre-training し，対象タスクで fine-tuning することで様々なタスクにおいて最高性能を更新している [6]．また，Bataa ら [7] の研究では，BERT による文書分類モデルが日本語を対象とした文書分類タスクで最高性能を達成したと報告している．そのため，本研究では，最高性能の文書分類モデルとして BERT を採用した．本研究では，transformers [2] の bert-base-japanese-whole-word-masking を用いて文書分類モデルの実装を行った．BERT のハイパーパラメータは，開発データを用いて，エポック数を 1 ，バッチ数 を 32，最大系列長 を 256 と設定した．また，その他のハイパーパラメータは，デフォルト値を設定した．

3 データセット

本節では，本研究で用いるデータセットの詳細について述べる．まず文書分類のベンチマークとして利用する Yahoo movie review データセットについて述べ，次に学習データにノイズを混入する方法について述べる．

3.1 Yahoo movie review

本研究では，Bataa ら [7] の研究でも利用している Yahoo movie review を文書分類のベンチマークとして用いる．このデータセットは，日本語の映画に関するレビューに対して，その内容が肯定的な意見であれば Positive ラベル，否定的な意見であれば Negative ラベルを付与されたものである[3]．

データの統計情報として，ラベル数は 2，学習データの文書数は 29017，開発データの文書数は 1528，評価データの文書数は 7637，学習データの平均単語数は 181，学習データの語彙数は 57267 である．ここで，ノイズを混入しない場合 (WITHOUT NOISE) の各文書分類モデルの正解率 (Accuracy) は，SVM で **89.11%**，BERT で **91.29%** である．本研究では，後述するノイズ混入方法により，学習データにノイズを混入させることで，文書分類モデルの正解率がどのように変化するのか調査し，各ノイズの影響の分析を行う．

3.2 ノイズ混入方法

本研究では，学習データへのノイズ混入方法として，ラベルノイズとテキストノイズの二種類のノイズを考える．まず，**ラベルノイズ**では，文書データの内容と正解ラベルとして付与されたラベルがアノテーション基準に矛盾する，つまり文書データの内容に対して間違えたラベルが付与された場合を想定する．次に，**テキストノイズ**では，文書データに対して付与されたラベルは正しいが，文書データの一部が削除，または正解ラベルと関係のない文字や単語に置換されている場合を想定する．表 1 では，ラベルノイズとテキストノイズの混入方法の詳細を示す．ここでは，ノイズの混合率 (Noise rate) を定義することで，ノイズの量を調整し，様々なパターンのノイズ混じりデータを生成する．

4 実験

本研究では，学習データにノイズが混入する場合，表 1 のノイズが文書分類モデルの分類精度にどれほど影響を及ぼすのか，ノイズ混じりデータを用いて検証実験を行う．

4.1 実験設定

本研究では，学習データに対してノイズ混入ルールを適用する．また，文書分類モデルへのノイズの影響を調査するため，評価データと開発データに対しては，ノイズ混入ルールは適用しない．

4.2 ノイズの影響

表 2 では，ノイズを混入しない場合の SVM における正解率 (**89.11%**) と各ノイズ混入時の正解率の差分を降順に並べた結果を示す．ここでは，Noise rate を [0.1, 0.2, 0.3] に設定した正解率の差分を示す．同様に，表 3 では，正解率の差分を昇順に並べた結果を示す．表 2 と表 3 を比較すると，表 2 ではラベルノイズが上位を占め，表 3 ではテキストノイズが上位を占めていることがわかる．表 2 では，学習データに矛盾するラベルを混入させる INCONSISTENT_LABEL および DUPLICATE_LABEL が多く存在しており，矛盾するラベルの存在は分類精度に影響を及ぼしている

[2] https://github.com/huggingface/transformers
[3] https://github.com/dennybritz/sentiment-analysis

表 1: テキストノイズおよびラベルノイズの混入方法を示す．ここで，未知語は OOV と定義する．

Noise	Rule	Description
Text	OOV_FREQUENT_VOCABS	語彙頻度の高いものから順に Noise rate の割合で未知語に変換する
Text	OOV_INFREQUENT_VOCABS	語彙頻度の低いものから順に Noise rate の割合で未知語に変換する
Text	OOV_RANDOM_VOCABS	Noise rate の割合でランダムに語彙を選択し未知語に変換する
Text	OOV_RANDOM_WORDS	文書中の単語を Noise rate の割合で選択し未知語に変換する
Text	SUBSTITUTE_IRRELEVANT_WORDS	文書中の単語を Noise rate の割合で選択し語彙からランダムに選択された単語に変換する
Text	SUBSTITUTE_RANDOM_CHARS	文書中の単語の各文字を Noise rate の割合でランダムに選択し別文字に変換する
Text	DELETE_RANDOM_WORDS	文書中の単語を Noise rate の割合でランダムに選択し削除する
Text	SHUFFLE_RANDOM_WORDS	文書中の単語を Noise rate の割合で選択し語順を入れ替える
Label	DELETE_LABEL	学習データ中の文書を Noise rate の割合でランダムに選択し削除する
Label	DUPLICATE_LABEL	学習データ中の文書を Noise rate の割合でランダムに選択し異なるラベルを付与しデータに追加する
Label	INCONSISTENT_LABEL	学習データ中の文書を Noise rate の割合でランダムに選択し異なるラベルに変換する
Label	IRRELEVANT_LABEL	無関係なテキストにラベルを付与し，学習データの Noise rate の割合でデータに追加する
Label	IMBALANCE_LABEL	一部のラベルを選択し，そのラベルを Noise rate の割合で削除する

表 2: SVM(WITHOUT NOISE) における正解率 (89.11%) と各ノイズ混入時の正解率の差分 (降順)

	Noise	Rule	Noise rate	Differnce
1	Label	INCONSISTENT_LABEL	0.3	15.90
2	Text	OOV_FREQUENT_VOCABS	0.3	9.48
3	Label	INCONSISTENT_LABEL	0.2	8.02
4	Text	OOV_INFREQUENT_VOCABS	0.2	3.67
5	Label	DUPLICATE_LABEL	0.3	3.37
6	Label	INCONSISTENT_LABEL	0.1	3.33
7	Label	DUPLICATE_LABEL	0.2	1.91
8	Text	OOV_FREQUENT_VOCABS	0.3	1.11
9	Label	DUPLICATE_LABEL	0.1	0.96
10	Text	OOV_INFREQUENT_VOCABS	0.1	0.86

表 3: SVM(WITHOUT NOISE) における正解率 (89.11%) と各ノイズ混入時の正解率の差分 (昇順)

	Noise	Rule	Noise rate	Differnce
1	Text	SUBSTITUTE_IRRELEVANT_WORDS	0.1	-0.26
2	Text	SHUFFLE_RANDOM_WORDS	0.1	-0.2
3	Text	DELETE_RANDOM_WORDS	0.1	-0.2
4	Text	OOV_RANDOM_WORDS	0.1	-0.18
5	Text	SUBSTITUTE_RANDOM_CHARS	0.1	-0.05
6	Label	IRRELEVANT_LABEL	0.1	-0.03
7	Text	SUBSTITUTE_IRRELEVANT_WORDS	0.2	0.03
8	Label	IRRELEVANT_LABEL	0.2	0.08
9	Label	IMBALANCE_LABEL	0.1	0.09
10	Text	SUBSTITUTE_RANDOM_CHARS	0.2	0.14

ことがわかる．次に，表 3 では，テキストノイズが上位を占めているように，文書分類モデルに SVM を利用する場合，Noise rate = 0.1 の少量のノイズが文書データに混入していたとしても，文書精度への影響は少ないことがわかる．

4.3 SVM と BERT の比較

図 1 では，各ノイズの Noise rate を [0.1, 0.2, 0.3, 0.4, 0.5] に設定した正解率を示す．ここでは，ノイズを混入しない場合 (SVM: **89.11%**，BERT: **91.29%**) と各ノイズ混入時の正解率との差分の平均値を計算し，SVM と BERT の結果を比較することで，各文書分類モデルのノイズに対する頑健性を評価する．比較結果を表 4 に示す．表 4 より，ラベルノイズに関しては各文書分類モデルの正解率の差分はほぼ変わらないが，テキストノイズに関しては BERT の分類精度に大きな影響を及ぼしていることがわかる．これは，図 1 の左下，OOV_RANDOM_WORDS や OOV_INFREQUENT_VOCABS の正解率からわかるように，BERT では未知語の比率が高い場合，過学習が発生することがあり，One-hot な素性を用いる SVM とは異なり，BERT は学習過程でテキストノイズの影響を大きく受けていると考えられる．

表 4: ノイズ混入時の正解率の差分の平均値

Model/Noise	Label	Text
SVM	3.97	2.00
BERT	4.12	7.85

5 関連研究

本研究と同様に，Agarwal らや Andrew らは，人工的なノイズをデータに混入させ，文書分類におけるノイズの影響に関する研究を行っている [3, 4]．Andrew ら [4] では，付与されたラベルを取り替えるラベルノイズや，文書の一部を取り除くテキストノイズを学習データに混入することで，文書分類のおいてノイズが分類精度に影響を与えることを報告している．しかしながら，ノイズごとの影響分析は行っていない．

学習データにノイズの混入が想定される場合の対策としては，大きく分けて二種類の対策が考えられる．一つ目は，学習データからノイズを含む事例を取り除くこと [8, 9, 10] と，二つ目は，ノイズに頑健な分

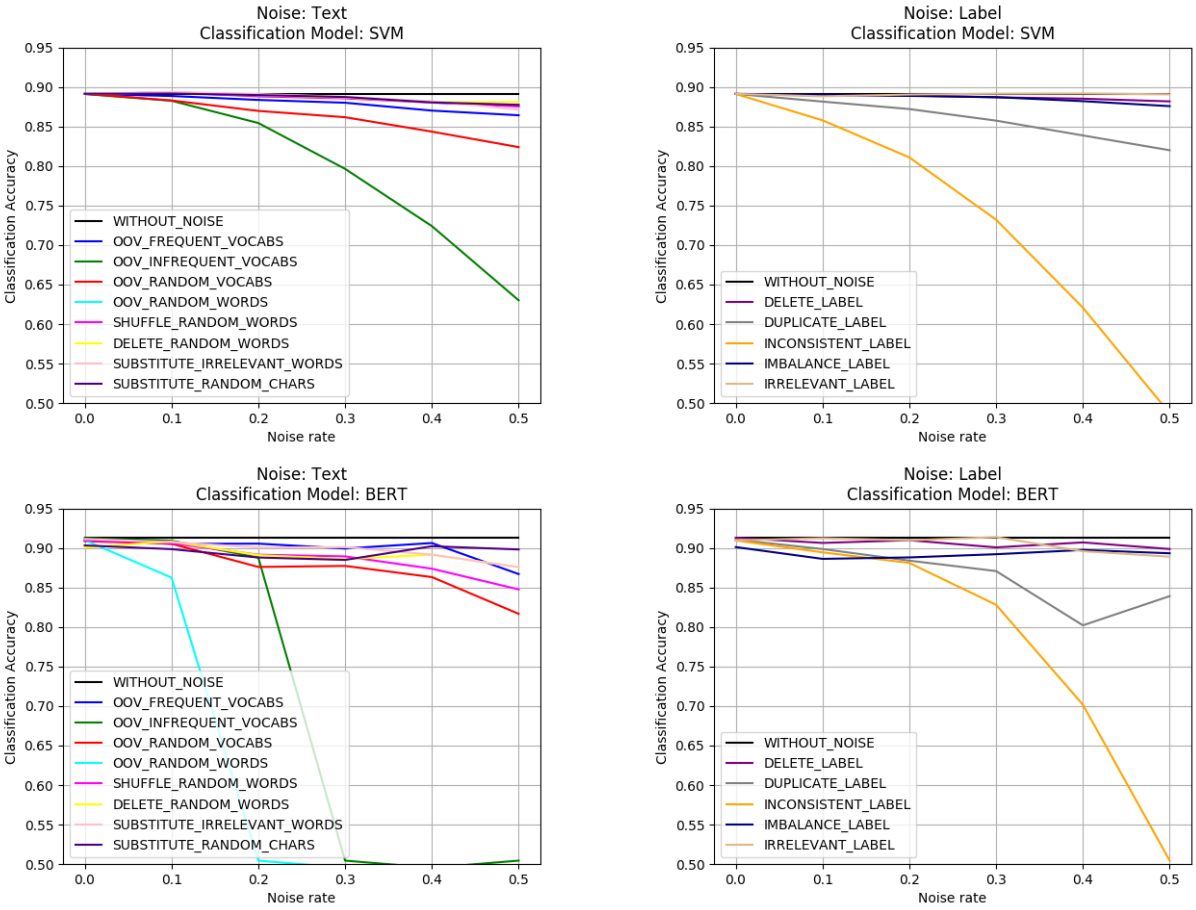

図 1: ノイズ混入時における文書分類モデルの正解率

類モデルを作成することである [4, 11, 12]．例えば，Jindal ら [11] の手法では，ノイズの生成をモデリングするネットワーク構造を用いることで，ラベルノイズに対して頑健な手法を提案している．

6 おわりに

本研究では，人工的に生成したノイズ混じりデータを用いて，文書分類におけるノイズの影響分析を行った．実験の結果，ラベルノイズは，文書分類モデルによらず分類精度を低下につながるため，今後の課題として，学習データからアノテーション基準に矛盾する事例を取り除く手法を検討したい．

参考文献

[1] 難波英嗣．人工知能による文書分類．情報の科学と技術, Vol. 66, No. 6, pp. 277–281, 2016.

[2] Ikuo Keshi, Yu Suzuki, Koichiro Yoshino, and Satoshi Nakamura. Semantically readable distributed representation learning for social media mining. In *Proceedings of ICWI*, pp. 716–722, 2017.

[3] Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. How much noise is too much: A study in automatic text classification. In *Proceedings of ICDM*, pp. 3–12, 2007.

[4] R. Andrew Kreek and Emilia Apostolova. Training and prediction data discrepancies: Challenges of text classification with noisy, historical data. In *Proceedings of EMNLP Workshop W-NUT*, pp. 104–109, November 2018.

[5] Sida Wang and Christopher Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of ACL*, pp. 90–94, July 2012.

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

[7] Enkhbold Bataa and Joshua Wu. An investigation of transfer learning-based sentiment analysis in Japanese. In *Proceedings of ACL*, pp. 4652–4657, July 2019.

[8] Fumiyo Fukumoto and Yoshimi Suzuki. Correcting category errors in text classification. In *Proceedings of ICCL*, p. 868, 2004.

[9] Andrea Esuli and Fabrizio Sebastiani. Training data cleaning for text classification. In *Proceedings of TIR*, pp. 29–41, 2009.

[10] Andrea Esuli and Fabrizio Sebastiani. Improving text classification accuracy by training label cleaning. *ACM TOIS*, Vol. 31, No. 4, p. 19, 2013.

[11] Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. An effective label noise model for DNN text classification. In *Proceedings of NAACL*, pp. 3246–3256, June 2019.

[12] Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. Learning with noisy labels for sentence-level sentiment classification. In *Proceedings of EMNLP-IJCNLP*, pp. 6285–6291, November 2019.