分類タスクにおける不確実性の高い文章の傾向調査

太田真人[1]　ファイサル・ハディプトラ[1]
[1] 電通国際情報サービス
{ota.m, faisal.hadiputra}@isid.co.jp

概要

信頼性のあるAIシステムの実現には精度だけでなく，予測の不確実性や説明性が必要である．予測の不確実性の推定は，分類を人に委ねるかの意思決定に役立つ．しかし，不確実性の高い文章の中に分類容易な文章が多いと，人はAIに不信感を抱く．そこで，分類タスクにおけるBERTモデルの予測の不確実性を軸に，文章傾向と精度を調査する．3つの分類タスクで実験をおこない，予測の不確実性の高い文章には精度が低く，分類困難な要因が多く存在することを示す．

1　はじめに

信頼されるAIシステムは，金融や医療など，人命や損害をともなう応用先で利用される．AIの信頼性の研究は，説明性や予測の不確実性が中心である[1]．予測の不確実性は，確率モデリングのベイズ推論をもとに推定されてきた[2]．自然言語処理と予測の不確実性の研究は，分布外検知[1]，文章要約の高品質化[3]，能動学習[4]といった品質・安全・効率化で取り組まれている．

信頼されるAIシステムの実現には，誤分類リスクが高い文章の傾向を調査する必要がある．人とAIの協調では，誤分類リスクの高い文章は，人間に判断を委ねる安全な設計をする[5]．人は誤分類が多いとAIシステムを信頼せず，システムから離脱する．しかし，人とAIの協調後も同様に，システムから離脱する恐れがある．それは，誤分類リスクの高い文章が人間にとって自明な文章が多いときである．したがって，私たちは，誤分類リスクの高い文章に分類困難な文章が多く含まれることを期待する．ここで，予測モデルの分類困難な文章は，訓練データ不足，クラス被覆，データバイアス，外れ値の文章を指す．

私たちは，誤分類リスクに予測の不確実性を用い，不確実性の高い文章傾向を調査する．先の予測モデルの分類困難な傾向は，予測の不確実性が高いときのみ顕著に存在し，精度が低い文章傾向とする．本研究の目的は，不確実性の高い文章中から分類困難な傾向を特定できるのか，また，どの程度存在するのかを明らかにする．予測の不確実性は，モデルの不確実性とデータの不確実性に分けられ，それぞれで調査する．データの不確実性は分類境界上で不確実性が高く，モデルの不確実性は訓練データ分布外で高くなるとされる[6]．実験は，事前学習済みモデルBERT[7]を用い，代表タスクとして，ネガポジ判定を2種類とニュース分類を1種類おこなう．

以下に本研究の貢献を示す．

- 3つの不確実性において，データの不確実性が誤分類リスクと最も関係があることを示す．
- ネガポジ判定では，データの不確実性から，ネガポジ両方の感情を含むハードサンプルやデータノイズを含む6つの傾向を特定し，モデルの不確実性から，低頻度・未知語や皮肉を含む5つの傾向を特定する．
- 定量化可能な不確実性の高い文章傾向の7種類中5種類は分類困難な要因であることを示す．

2　問題設定

本章ではBERT分類モデルの予測の不確実性の種類とその定量化方法を説明する．入力文章とラベルの組 $(x, y) \in (X, Y)$ を複数持つデータセットを $\mathcal{D}$ とする．予測はモデルパラメータ w を持つ事前学習済みモデルの出力 $f(x; w)$ とする．不確実性の推定の準備として，モデルパラメータの事前確率を $p(w)$ とし，予測分布を $p(y|x, w)$ とする．テストデータ $(x^*, y^*) \in \mathcal{D}_{test}$ に対する周辺予測分布は，パラメータの事後分布 $p(w \mid \mathcal{D})$ を用いて以下の式で与えられる．

$$p\left(y^* \mid x^*, \mathcal{D}\right) = \int p\left(y^* \mid x^*, w\right) \underbrace{p(w \mid \mathcal{D})}_{\text{posterior}} dw$$

2.1 不確実性の推定

本節では，不確実性の種類とその尺度を紹介する [6]．予測の不確実性の定量化はベイズ推論からおこなえる [8][2]．

2.1.1 全体の不確実性

全体の不確実性は，周辺予測分布 $p(y|x,\mathcal{D})$ の不確実性である．全体の不確実性の尺度は，周辺予測分布の分散やエントロピーで表される．周辺予測分布のエントロピーは，後に説明するデータの不確実性の予測エントロピーの期待値とモデルの不確実性の相互情報量との和として式変形できる [9][10]．

$$\underbrace{\mathbb{H}[Y \mid x, \mathcal{D}]}_{\text{predictive}} = \underbrace{\mathbb{I}[Y; \omega \mid x, \mathcal{D}]}_{\text{model}} + \underbrace{\mathbb{E}_{p(\omega|\mathcal{D})}[\mathbb{H}[Y \mid x, \omega]]}_{\text{data}} . \tag{1}$$

2.1.2 モデルの不確実性

モデルの不確実性は，モデルが特徴表現として獲得できていない，知識の欠如に由来する不確実性である．特に訓練データ分布と大きく異なる文章に対し，不確実性が高くなる．知識の欠如のため，該当する訓練データを増やすことで減少する不確実性として知られている．モデルの不確実性の尺度は，(1) 式の第一項で表す，パラメータ w と出力 y との相互情報量 $\mathbb{I}[Y; \omega \mid x, \mathcal{D}]$ として表される．

2.1.3 データの不確実性

分類境界上の複雑な入力に対し，データの不確実性は高くなる．訓練データを増やすだけでは減少しない不確実性であり，根本的に該当するデータを取り除くか修正する必要がある．データの不確実性の尺度は，事後分布 $p(\omega \mid \mathcal{D})$ からサンプリングされたモデルパラメータに対する予測エントロピーの期待値 $\mathbb{E}_{p(\omega|\mathcal{D})}[\mathbb{H}[Y \mid x, \omega]]$ として表される．

2.2 不確実性に基づく文章の分析方法

本節では不確実性に基づく文章の分析方法を説明する．テストデータから不確実性を軸に誤分類リスクの高い文章傾向を分析する．モデルの不確実性の尺度を用い，データ不足が要因の文章傾向を調べ，データの不確実性の尺度を用い，データの複雑さが要因となる文章傾向を調べる．分析方法はそれぞれの尺度の高い文章と低い文章を上位 k% を人が読み，頻出する文章パターンの傾向をまとめる．

3 実験

3 種類の文章分類データセットを用いて，データの不確実性とモデルの不確実性の高い文章傾向を調査する．

3.1 データセットとモデル

3 種類の文章分類データセットを用いる．感情分析に 2 クラス分類の Amazon 商品レビュー marc-ja [11] と 3 クラス分類の twitter の呟きデータ wrime [12]，9 クラス分類のニュース記事 livedoor-news を用いる．事前学習済み cl-tohoku/bert-base-japanese [1] を用いる．BERT のアーキテクチャは 12 層，隠れ次元数は 768，アテンションヘッド数は 12 ある 110M パラメータを持つ．事前学習には，約三千万データを含む日本語版の Wikipedia を使用している．トークナイザーにはワードピースレベルで MeCab を用いる．語彙数は 32,768 である．

3.2 実験設定

不確実性の定量化手法には Deep Ensembles [13] を用いる．Deep Ensembles は BERT の出力層の初期値の seed 値を変えて，$M = 5$ でアンサンブルする．BERT を微調整するために，エポック数 3，学習率は $5e-5$，バッチサイズは 16，文章長は marc-ja と livedoor-news で 512，wrime は 140 とする．

各不確実性と誤分類リスクとの関係調査に，[14] が提案した選択的予測に使われる RCC-AUC と Accuracy （Acc） を用いる．リスクカバレッジカーブ (RCC) は，予測棄却基準に応じたテストデータに対する累積誤分類数を示すグラフである．曲線下の面積が小さいほど，不確実性の推定値が誤分類リスクの基準に良いことを示す．予測棄却基準に全体の不確実性（TU），データの不確実性（DU），モデルの不確実性（MU）を用いる．それぞれの不確実性の尺度には，2.1 節で説明した指標を用いる．

本実験では，文章傾向の定性的調査にデータ分析経験 2 年以内の NLP の業務経験もない分析者 3 名がおこなう．これは，事前知識と経験をもとに傾向を発見するのを防ぐためである．分析者は marc-ja のテストデータ全体の 5%にあたる不確実性の高い文章 300 件と不確実性が低い 300 件から傾向を探す．

分析者が発見した不確実性の高い文章傾向を定量

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

化し，分類困難な要因か検証する．分類困難な要因は，不確実性の高い文章の傾向が不確実性の低い文章と比較し，出現頻度が高く，精度が低いとする．したがって，定量化可能な傾向は，不確実性の高い文章集合と低い文章集合に対し，出現数と精度を計算する．

3.3 実験結果

各データセットでの Deep Ensembles の Accuracy を表 1 に示す．wrime は twitter のデータで口語調のため，精度が低い．各データセットでの Deep Ensembles の RCC-AUC を表 2 に示す．RCC-AUC が小さいほど，誤分類リスクのある文章を不確実性の尺度で棄却できる．結果，DU が 2 つのデータセットで最も小さく，wrime は TU が最も小さい．

marc-ja における定性的傾向分析結果を表 3 と表 4 に示す．表 3 はデータの不確実性の高い文章から発見できた定性的な傾向を示す．ハードサンプルは文章の文脈的にラベル付けが困難なサンプルである．”ネガポジ・ポジネガ文” は，文章の序盤はポジティブだが，終盤ではネガティブな内容な文章を指す．”ネガポジ形容詞なし” は，明示的にネガポジに関する形容詞はでてこず，暗黙的に感情が表現されている文章を指す．”変換ミス” は，タイピングミスを指す．”別商品と比較” は，レビュー対象の商品ではなく，比較対象の商品を評価する文章である．データノイズは文章の文脈に依存しないエンコードミスといった表面上のノイズを表す．”同じ文字の繰り返し” は，1 文章の中に 3 回以上続けて同じ文字が使われる文章を指す．

表 4 はモデルの不確実性の高い文章から発見できた定性的な傾向を示す．カテゴリを未知語・低頻出と悪評や皮肉のような表現と分類する．未知語・低頻出は訓練データにほどんど存在しない珍しい文章を表す．”英単語” は，文中に英単語が 10 単語以上ある文章を表す．”固有名詞” は，文中の固有名詞の割合が 10%以上ある文章を表す．”特殊記号” は，文中の英数字以外の文字列が 10%以上ある文章を表す．悪評や皮肉のような表現は，文章全体ではポジティブだがネガティブと予測される言い回しである．定性的な調査では，”良い意味で” に続くフレーズがネガティブな内容なため，ネガティブと予測されている文章が散見された．

表 5 に各定性的分析結果から出現数と精度を示す．7 項目中 5 項目は予測モデルの分類困難な要因と考えられる．一方で，”同じ文字の繰り返し” と ”英単語” に関しては不確実性が高さに関係がなく，分析者のバイアスだった．

表 1 各データセットの精度比較結果

metric	marc-ja	wrime	livedoor-news
Acc	94.32	71.32	96.46

表 2 各データセットの RCC-AUC ↓ 比較結果

UE Type	marc-ja	wrime	livedoor-news
TU	47.51	**328.61**	3.67
DU	**47.37**	331.89	**3.62**
MU	62.10	378.82	4.54

4 考察

4.1 モデルの不確実性とデータの不確実性が誤分類と関係があるのか

両方の不確実性の推定が予測精度と関係はあるが，DU の方が顕著な傾向が見えた．表 2 の結果から，RCC-AUC では，DU が最も性能が高く，誤分類リスクを測る尺度であることがわかる．この結果は，[15] の事前学習みモデルを用いない LSTM の結果と類似する．事前学習済みモデルの微調整でも同様の結果が得られた．TU は DU と類似した結果を示すが，MU を含み，わずかに劣る結果になった．この結果は，MU が分布外サンプルの検出精度が高いが，本実験のテストデータには，分布外データが少ないことによるものと考えられる．今後は日本語文章の分布外検出精度を評価する．

4.2 モデルの不確実性やデータの不確実性が高くなる文章の傾向はあるのか

発見した傾向のカテゴリは既存研究と同様の結果が得られた [15]．表 3 と表 5 からデータの不確実性が高い文章の傾向カテゴリは，人間が読んでも分類が困難な文章やデータノイズである．しかし，本研究では，詳細に傾向を調べた．具体的な分類困難な文章は，表 5 の”ネガポジ・ポジネガ” や”ネガポジ形容詞なし” からわかる．特に，”ネガポジ形容詞なし” は不確実性の高い上位 300 件に 24 文章含む．一方で，不確実性が低い上位 300 件には 3 件しか含まれておらず，DU が高い顕著な傾向である．また，そのときの正答率も 41%の差がある．

同様にモデルの不確実性が高い文章の傾向は，訓練データに少ない単語や記号を含む文章である．分析者が見つけた傾向の ”固有名詞” と ”特殊記号” は，

表 3　データの不確実性の高い文章傾向の定性的な分析結果 (Marc-ja)

カテゴリ	傾向	文章例
ハードサンプル	ネガポジ・ポジネガ文	値段が安いのでこの程度でしょう。この商品は期待に添いました。
	ネガポジ形容詞なし	最後の Now, where was I?の日本語訳は、「さて、どこだっけ？」 ではなく、「さて、どこまで話したっけ？」だと思います。
	変換ミス	安いのに、何枚も CD 入っていてびっくりシタ！
	別商品と比較	「エイトミニッツミッション」の方が面白かった。ただの焼き増しだ。
データノイズ	エンコードミス	æªè³ãªå½ç©ã ¢ã-ãªAmazonãã''ã''ãª...
	同じ文字の繰り返し	朝、注文したら、夕方に配達。超早い・・・・・・・・・・・・・・・・・ これがほしいなぁ)))))))))))))))))))))))))))))))

表 4　モデルの不確実性の高い文章傾向の定性的な分析結果 (Marc-ja)

カテゴリ	傾向	文章例
未知語・低頻出	英単語	OPPO BDP-105DMR-BZT9300BDZ AX2700TBDZ ET1000PS...
	固有名詞	トニースコットが死んで、そのニュースで買いました。 タランティーノ×トニースコット。
	特殊記号	☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆☆
	新規クラス	詳しく具体的に書きましょう。商品購入前に知っておきたい情報などは 他のお客様に役立ちます。レビューは短すぎたり、長すぎたり...
悪評や皮肉のような表現	ポジをネガ判定 (”良い意味で”)	新譜を買うときにはいつもどきどき感がありますが、 良い意味で期待を裏切った作品です。

表 5　予測の不確実性の文章傾向の定量化の結果 (Marc-ja)．傾向の黒字は，不確実性が高いときにカウントが多く，正答率が低い分類困難な要因を表す．

傾向	不確実性	カウント (出現率)	正答率
ネガポジ・ポジネガ	High	57 (19%)	0.57
	Low	40 (13%)	1.00
ネガポジ形容詞なし	High	24 (8%)	0.59
	Low	3 (1%)	1.00
エンコードミス	High	4 (1.3%)	0.61
	Low	1 (0.3%)	1.00
同じ文字の繰り返し	High	48 (16%)	0.66
	Low	48 (16%)	1.00
英単語	High	7 (2.3%)	0.83
	Low	15 (5%)	0.98
固有名詞	High	6 (2%)	0.59
	Low	2 (0.6%)	1.00
特殊記号	High	20 (6%)	0.62
	Low	10 (3%)	1.0

訓練データに低頻出な文章と考える．”特殊記号” の場合，MU が高い 300 件中 20 件含まれ，精度は 62% と低い．一方で，MU が低い 300 件中に 10 件しかなく，精度が 100% と高いことから，”特殊記号” は MU が高い傾向と考える．

4.3　予測の不確実性に基づく文章分析は分類困難な要因を示せるか

予測モデルの分類困難な要因を示すため，予測の不確実性を定量化し，定性的な分析をおこなった．予測の不確実性が高いときのみ文章数が多く，精度が低い文章傾向を分類困難な要因とした．表 3 と表 4 から不確実性に基づく文章分析は，データの不確実性からハードサンプルとデータノイズを発見でき，モデルの不確実性から未知語や低頻出な単語を発見できた．また，表 5 の結果から，この傾向は不確実性が高い文章のみ精度が低く，出現頻度が多いことがわかる．さらに，表 5 の定量化可能な傾向に絞ると，分類困難な要因は 7 個中 5 個あった．この結果から，予測の不確実性から発見する傾向の多くは分類困難な要因を表す．

5　まとめ

予測の不確実性に基づく文章分析から分類が困難な文章傾向を実験的に示した．定量化可能な傾向の 7 種類中 5 種類は，不確実性が高い文章で誤分類率が高く，文章の出現頻度も高く，分類困難な文章傾向だった．結果として，不確実性の高い文章には予測モデルの分類困難な文章傾向が多く発見された．しかし，定量化できていない傾向も多く，人が全ての文章をグループ化したわけではない．グループ化されていない文章が人に与える影響の調査が今後の課題である．

参考文献

[1] Yibo Hu and Latifur Khan. Uncertainty-aware reliable text classification. In **Proceedings of the 27th ACM**

SIGKDD Conference on Knowledge Discovery amp; Data Mining, p. 628–636, 2021.

[2] Radford M Neal. **Bayesian learning for neural networks**, Vol. 118. Springer Science & Business Media, 2012.

[3] Alexios Gidiotis and Grigorios Tsoumakas. Should we trust this summary? Bayesian abstractive summarization to the rescue. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 4119–4131, May 2022.

[4] Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. Active Learning for BERT: An Empirical Study. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7949–7962, November 2020.

[5] Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Towards improving selective prediction ability of NLP systems. In **Proceedings of the 7th Workshop on Representation Learning for NLP**, pp. 221–226, May 2022.

[6] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. **arXiv preprint arXiv:2107.03342**, 2021.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, June 2019.

[8] David John Cameron Mackay. **Bayesian methods for adaptive models**. California Institute of Technology, 1992.

[9] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. **arXiv preprint arXiv:1803.08533**, 2018.

[10] Yarin Gal. Uncertainty in deep learning. 2016.

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

[12] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2095–2104, June 2021.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. **Advances in neural information processing systems**, Vol. 30, , 2017.

[14] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. **J. Mach. Learn. Res.**, Vol. 11, p. 1605–1641, aug 2010.

[15] Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 33, pp. 7322–7329, 2019.