ビジネス文書を対象に自然言語処理を用いた読み手にストレスを与える文章の検出に関する研究

Detecting Potentially Stress-Inducing Language in Business Documentation through NLP

松永 剛之進 [1)]
指導教員　青木 輝勝 [2)]

1) 東京工科大学大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻
2) 東京工科大学　コンピュータサイエンス学部

本研究では，文章の書き手が意図しない伝わり方を防ぐため，人工知能を用いて読み手にストレスや誤解を生む表現を自動で検出し，書き手に対して適切な表現への修正を支援する方法を提案する．ストレスと誤解の要因となる曖昧な表現や責任回避的な表現を検出し，文章の伝わり方を最適にすることを目指す．

人工知能，機械学習，自然言語処理，大規模言語モデル，BERT

1． 研究背景

情報過多の現代においては，文章を読む際にスkに書き手の意図していない伝わり方を防ぐために文章 DC9(Deep Communication 9)ヒューリスティック評価[1]の「情報の質」の項目(図 1)がある．この項目は，曖昧な表現や責任回避的な表現などを意図せず使用していないかを評価するものである．

本研究では，図 1 に示した内容をルールベースと機械学習ベースを用いて読み手にストレスを与える可能性がある文章の検出を行うシステムを開発する．

情報の質：読み手に誤解とストレスを与えない
- 曖昧な表現を避ける
- 重複情報を避ける
- 受け身の表現を避ける
- 責任回避と誤解される文章を避ける

図 1．情報の質の定義

2．研究目的

本研究の目的は，自然言語処理を用いて読み手にストレスを与える文章を検出し，書き手に対して改善のための具体的なフィードバックを提供するシステムを構築することである．情報の質を指標として，ストレスを感じる可能性のある表現を 5 つ定義した．この 5 つの表現に「問題がない表現」を加えた 6 つの表現をラベルとし，クラス(表 1)を設定した．

クラス 4 の曖昧な表現は動詞の目的語の有無で判定し，クラス 5 の受け身表現は主語が書き手であるにも関わらず受け身を用いる場合に該当する．クラス 4 と 5 は従来のルールベースで判定可能であるため，本論ではその他のクラス 0~3 を大規模言語モデルである BERT (Bidirectional Encoder Representations from Transformers) [2]を用いて機械学習ベースで検出する．

class	label
0	問題がない表現
1	過剰に丁寧な表現
2	責任回避的な表現
3	情報が重複する表現
4	曖昧な表現
5	受け身の表現

表 1．クラス

3．提案手法

3.1．検出システム

BERT を用いて，クラス 1~3 で多クラス分類の学習を行う．分類結果を用いて，ストレスを感じる要因となる表現を検出するシステムを構築する．このシステムは，一文ずつ文章を解析し，検出された文章とその理由（ストレスを感じる要因）をユーザに知らせるものである．

3.2．データセット

日本語事前学習モデルには東北大学のデータセット[3]を使用し，ファインチューニング用データセットは金融関連の文章 100 文を収集後，ChatGPT(Generative Pre-trained Transformer)[4]で各クラスに適した表現に言い換えた．1 クラスあたり 10 文を生成し，合計 4,000 文のデータセットを作成（各クラス 1,000 文）．訓練用に 6 割，検証用とテスト用にそれぞれ 2 割ずつに分割した．

4．実験

4.1．実験概要

本実験では，BERT モデルを訓練データと検証データを使用してファインチューニングした．テストデータを用いて正解率を測定し，モデルの精度を評価した．学習とテストは毎回ランダムにデータを分割して，10 回の試行を実施し，結果における精度の安定性と再現性を確認した．

4.2．実験結果

クラスごとの平均正解率，最大値，最小値を表 2 に示す．全体の平均正解率は 0.894 であり，クラス 1 が最も高い正解率(0.932)を示し，クラス 2 が最も低い正解率(0.861)であった．表 3 に各クラスでの推論結果を示す．特にクラス 1 とクラス 3 で高い精度が得られた．対照的に，クラス 2 では他クラスとの混同が見られ，誤分類が発生しやすい傾向がある．また，クラス 3 に他クラスが誤判定される傾向もあり，クラス 3 の特徴が他のクラスと重なる部分が多い可能性を示唆している．

class	avg accuracy	Max	Min
0	0.879	1.000	0.790
1	0.932	1.000	0.900
2	0.861	0.975	0.790
3	0.904	0.945	0.829
overall	0.894	0.959	0.856

表 2．モデルの精度評価

		predicted class				
	class	0	1	2	3	Total
ground truth	0	**1721**	8	37	234	2000
	1	14	**1873**	12	101	2000
	2	78	0	**1735**	187	2000
	3	29	1	146	**1824**	2000
	Total	1842	1882	1930	2346	8000

表 3．各クラスの推論結果

4.3．考察

全体の正解率が高い一方で，クラス間の誤分類も一定数存在することから，データの再調整や特徴抽出の改善により，モデルの改良が必要と考えられる．

5．今後の展望

実験により，従来のルールベースでは困難だったクラス 0〜3 の検出が機械学習ベースで可能であることが示された．今後，クラス 4 と 5 の実装を加え，「情報の質」を自動で判定するシステムを完成させる．

参考文献

[1] UCDA. 文章 DC9 ヒューリスティック評価, https://ucda.jp/solutions/text_dc9.html

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*, 4171-4186.

[3] 東北大学乾研究室. bert-base-japanese, https://onl.bz/VsMxmgV

[4] OpenAI. (2024). ChatGPT [Large language model]. OpenAI. Available at: https://www.openai.com/