

テキスト生成 AI による自由記述のラベル付けの安定性： AI と著作権に関するパブコメ分析から

隅谷 孝洋^{1,a)} 天野 由貴²

概要：テキスト生成 AI で、曖昧な自由記述文を「機械的に」分類したりラベル付けすることが可能になった。アンケート分析や対話分析など、この機能は広範囲にわたり便利に使うことができる。しかし、テキスト生成 AI からの回答は「毎回違う」と言われることもあり、研究ツールとして使うためにはその安定性の評価が必要だと考えられる。ここでは、文化庁が募った「AI と著作権に関する考え方について（素案）」に対するパブリックコメントの分析を題材として、いくつかの方法でラベル付をした場合の安定性を評価する方法について考察する。

キーワード：生成 AI, 自由記述文分析, ラベル付の安定性

Stability of Labeling Free-Text Responses with Text-Generative AI: Insights from Public Comments on AI and Copyright

TAKAHIRO SUMIYA^{1,a)} YUKI AMANO²

Abstract:

With the advent of text-generative AI, it has become possible to “automatically” classify or label ambiguous free-form text. This capability can be very useful in a wide range of applications, such as survey analysis and dialogue analysis. However, because text-generating AI can produce different answers each time, it is considered necessary to evaluate its stability if we want to use it as a research tool. In this study, using the analysis of public comments submitted in response to the Agency for Cultural Affairs’ “Draft Perspectives on AI and Copyright” as a case example, we explore methods for evaluating the stability of labeling when performed in several different ways.

Keywords: generative AI, analysis of free format sentence, stability of labeling

1. はじめに

教育工学においてはレポートやコメント、アンケート等で自由記述された文章を扱うことが多い。過去には定性的分析や、もしくはテキストマイニングを用いて定量的な分析に持っていくことが多かった。しかし、昨今の生成 AI の発展により、自由記述文にラベル付けや、何らかのレーティングなどを行ったりして定量的分析に持っていくことが可能になってきた。しかし、生成 AI を、人間との会

話を主目的としたチャットボットで使っていると、バラエティ豊かな表現になり、データ分析の道具として使うことは難しそうである。研究のツールとして使うには、その再現性や安定性について評価し、振る舞いを把握しておく必要があるだろう。ここでは昨年文化庁が募集した生成 AI と著作権に関するパブリックコメントを例として、生成 AI を用いた自由文のラベル付けの安定性について検討してみたい。

2. AI と著作権に関するパブリックコメント

2018 年、情報解析に著作物を用いる際の権利制限につい

¹ 広島大学, Hiroshima University

² 帝京大学, Teikyo University

^{a)} sumiya@hiroshima-u.ac.jp

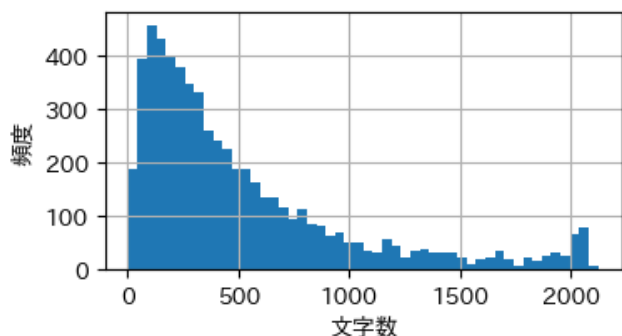


図 1 パブリックコメントの文字数の分布

ての法改正が行われた（30 条の 4、2018 年 5 月 25 日公布、2019 年 1 月 1 日施行）。著作物に表現された思想感情を享受することなく、情報解析に用いるためであれば、著作権者の利益を不当に害することのない範囲内で、著作物を自由に利用できる、という内容である。この法改正により、AI に他者著作物を学習させる際に著作権者の許諾を得る必要がなくなり、大量の学習を円滑に行えることになった。

その後、画像を生成する AI が開発され一般にも使われるようになった。また、2022 年末に ChatGPT 3.5 が画期的なテキスト生成 AI として公開され、にわかに注目を集めることとなった。2018 年の立法当初よりも AI ができることが想像以上に増え、利用が広まった結果、クリエイターを中心に日本の生成 AI に対する著作権法のあり方を問題視されることが多くなり、激しい議論が巻き起こってきた。

文化庁では、この新しい技術による作品を著作権法の文脈でどのように扱うべきかの検討を開始、文化審議会著作権分科会法制度小委員会で「AI と著作権に関する考え方について」という文書の素案を作成した。その内容について広く国民に意見を問うパブリックコメント募集を 2024 年 1 月 23 日から 2 月 12 日にかけて行った。

令和 5 年度第 7 回法制度小委員会^{*1}の資料によれば 24,938 件（うち団体・法人 73 件）のコメントが集まった。パブリックコメントとしては異例の件数で、この問題に対する関心の高さが窺える。集まったコメントは全て公開されるということだが、2025 年 1 月現在、団体・法人分と、個人分 6,000 件が前記小委員会のページで公開されている。

3. テキスト生成 AI によるパブリックコメントの分析の例

公開されている 6,000 件（通し番号は 6,000 までだが、実際には僅かに欠損して 5,994 件）のコメントを、テキスト生成 AI で分析する例を示す。

パブリックコメントは、およそ 2,000 文字までしか入らないように入力制限があったようだ。図 1 は、文字数の分布を示している。6 割が 500 文字以内で、中央値は 354 と

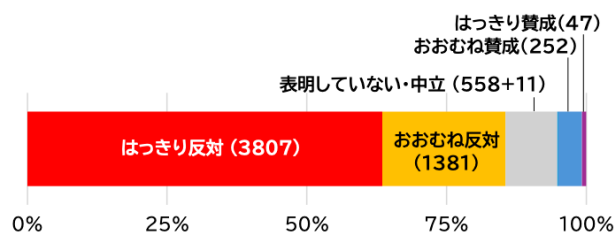


図 2 AI 活用に対する意見でパブリックコメントを分類

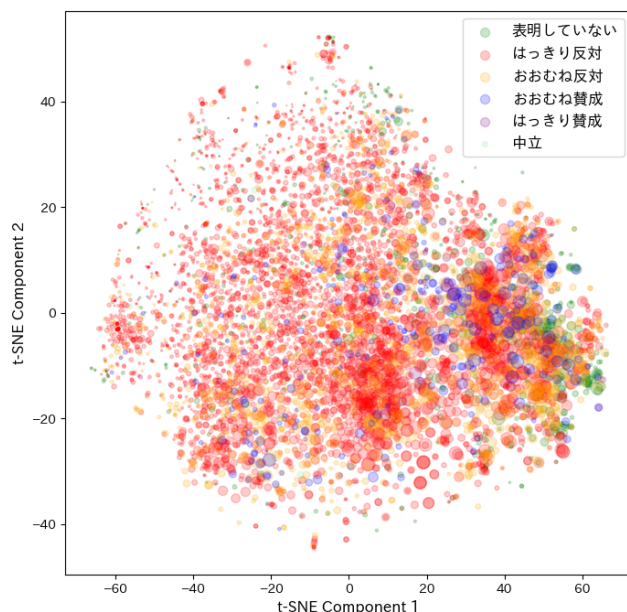


図 3 5,994 件のコメントを text-emb-3-large モデルで埋め込みベクトルにしたものを t-SNE で 2 次元にプロット。マーカーの色は AI 活用に対する意見、サイズは元コメントの文字長

なっている。

生成 AI を使った分析として、まず、テキスト生成 AI にコメントを与えてその内容に応じたラベル分けをすることができる。たとえば、寄せられたコメントが、生成 AI の活用に対して賛成の意見か反対の意見かを分類するため、「以下のコメントは、生成 AI の活用自体について『はっきり賛成』『おおむね賛成』『中立』『表明していない』『おおむね反対』『はっきり反対』のどれに当たるか判定をしてください。」という質問文を 5994 件のコメントひとつひとつにつけて、生成 AI に送り返答をさせる。その結果を示したのが 2 となる。

生成 AI に関連する自由記述文解析に有用な手法の一つに、文章を埋め込みベクトルに変換するものがある。全ての文章を同じ次元の数値ベクトルとして扱うことができ、文章間の類似度を定量的に扱うことができる。ここでは OpenAI が提供している text-emb-3-large というモデルを使い、5,994 件のコメントのそれぞれを 3,072 次元の埋め込みベクトルに変換した。

埋め込みベクトルをさらに t-SNE を用いて 2 次元に圧縮し、生成 AI の活用に対する判別結果をマーカーの色を

^{*1} https://www.bunka.go.jp/seisaku/bunkashingikai/chosakuken/hoseido/r05_07/

userとして送る文章は、日本の文化庁が作成した「AIと著作権に関する考え方について」（以下文化庁素案）に対するものとして市民から送られてきたコメントです。
このコメントについて、以下の判定をしてください。

意見：生成AIの活用自体に対しての意見を表明しているものであれば「一般的」、文化庁素案の具体的な箇所を示して意見を表明しているものや追加したい内容を記しているものは「文化庁素案」とし、法制度の整備など「考え方」の範囲を超えるものについては「法制度」としてください。

素案意見：文化庁素案に対して、素案全体を肯定的に捉えているもの（賛成）か否定的に捉えているものか（反対）、具体的な修正案を示したものの、特に意見を表明していないか、「賛成」「反対」「修正案」「なし」のいずれかを判定してください。

生成AI種類：どのタイプの生成AIに対する意見かについて、「文章」「画像」「動画」「音声」「音楽」「複数含む」「特定していない」のいずれか一つを選んでください。

生成AI活用：生成AIの活用自体について「はっきり賛成」「おおむね賛成」「中立」「表明していない」「おおむね反対」「はっきり反対」のどれに当たるかを判定してください。

返答はJSONフォーマットで返してください。JSONのキーは「意見」、「素案意見」、「生成AI種類」、「生成AI活用」です。

図 4 実験に用いたプロンプト

用いて表したのが図 3 である。マーカーの大きさは、コメントの文字数を表している。

4. ラベルの安定性

ここで問題になるのは、このラベル付けがどの程度の精度を持っているのかということになる。例えば、図 2 にあげたラベルの度数は、どの程度正しいのか。人手によるラベリングと比較して、生成 AI のラベルの精度を評価した研究はいくつもある [1], [2]。われわれの研究でも、SVM や BERT を用いたものよりも OpenAI の GPT (3.5, 4) を使ったもののほうが正確であるという結果が出ている。

一方、チャットボットを指向して作られた生成 AI の応答は、人との対話に使うため意図的に揺らぎを調整できるようになっている。この調整は *temperature*, *top_p* といったハイパーパラメータを通して行うことができる。*temperature* は、ニューラルネットワークの出力を確率分布に変換するために使われる Softmax 関数に含まれるパラメータで、小さくなるとニューラルネットワークからの出力で大きな値を持つものがより強調される確率分布になる。OpenAI が提供している ChatGPT などのチャットボットでは 0.7-0.8 に設定されているのでは、とされている。

この研究では最終的には生成 AI によるラベル付の安定性を数値的に評価したいが、ここではまずどの程度のバラツキがあるかを大雑把に掴むことを目的として、数値実験をおこなう。

数値実験では、同じ条件で繰り返し応答を生成させ、応答がどの程度変わるのか、変わらないのかを検証する。5,994

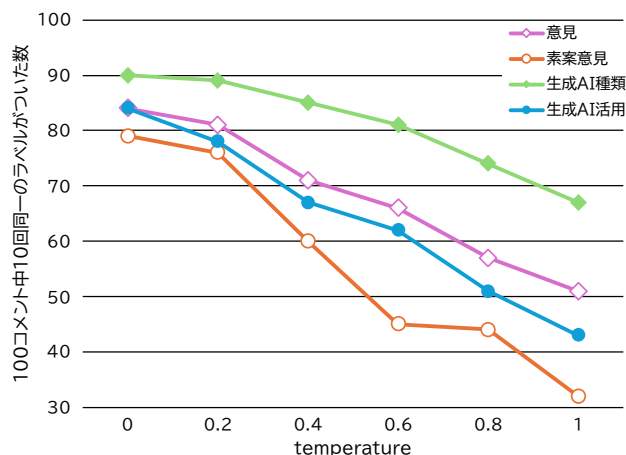


図 5 実験結果 (model=gpt-4o/top-p=1.0)

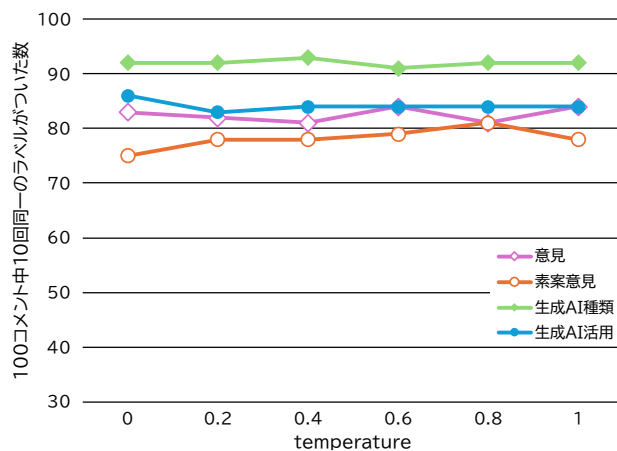


図 6 実験結果 (model=gpt-4o/top-p=0.0)

件すべてのコメントについて行うことはコスト上難しいため、100 件を無作為に選んで数値実験のサンプルとした*2。

テキスト生成 AI のモデルには gpt-4o-mini を用いた。top_p を 1.0 に固定して temperature を 0.2 刻みで 0.0 から 1.0 まで動かし、100 件のコメントのそれぞれに図 4 のプロンプトをつけ、ラベル判定結果を 10 回生成させた。結果が安定しているものであれば、10 回中同じラベルとして判定されるものの数が増えるはずである。図 5 は、この条件で 10 回とも同じラベルがつけられたものが 100 件のコメント中いくつあったかを示したものである。図 6 は同様の条件で、top_p を 0.0 に固定したものである。

4.1 結果

ここでは実験に用いた 100 コメント中 10 回の試行で 10 回とも同じラベルが得られたコメントの数を、安定性の指標とした。これはわかりやすいが、同じラベルが 10 回未満だったものに差をつけることができないので必ずしも良い指標とは言えない。それを踏まえた上で、以下の知見を得ることができた。

*2 サンプルデータは <https://home.riise.hiroshima-u.ac.jp/~sumi/ce178sample.html> で公開

- 安定性は質問の内容によってかなり変わる。この例だと、選択肢の多い「生成 AI 活用」や、素案の内容を把握していないと判断が難しい「素案意見」は安定性が低かった。逆にいうと、同様の手法で評価すれば、「ラベル付の難しさ」が可視化できるかもしれない。
- `top_p` の値がデフォルトの状態 (1.0) だと、`temperature` の値により相当安定性が変わるが、0.0 に変更した状態だと `temperature` の影響をほとんど受けず高い安定性を示した。
- しかし、`temperature` が低いところ (0.0 とか 0.2 とか) では `top_p` の値が 0 でも 1 でもそれほど変わらない安定性だった。OpenAI は `top_p` と `temperature` を「同時に動かさないほうがよい」と推奨しているので、テキストにラベルづけをする際には `top_p` はデフォルトのままで `temperature` を 0.2 以下くらいにするのがよいのではないか。
- 図は示していないが、プロンプトで質問する順序を変更すると安定性が変わる現象が見られた。複数のラベル付が必要な時は一度に聞くのではなく、別々に送信したほうがよいかもしれない。この点は今後実証してみる予定である。

参考文献

- [1] Weerts, H. J., Mueller, A. C., & Vanschoren, J. (2020). Importance of tuning hyperparameters of machine learning algorithms. arXiv preprint arXiv:2007.07588.
- [2] Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., & Baldwin, B. (2024). LLM Stability: A detailed analysis with some surprises. arXiv preprint arXiv:2408.04667.
- [3] Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406, 109-120.