大規模言語モデルを用いた傷害事件の関連法律予測

中下咲帆[1] 藤後英哲[1] 菊池英明[1]
藤倉将平[2] 則竹理宇[2]
[1]早稲田大学人間科学研究科 [2]LawFlow 株式会社
honkon@fuji.waseda.jp eitetsu@akane.waseda.jp kikuchi@waseda.jp
fujikura@lawflow.jp noritakebengo@keiji-bengosi.com

概要

大規模言語モデル（以下，LLM）の発達に伴い，高度な知識を必要とする法的な業務の自動化が期待されている．本研究では，特に法律相談場面に注目し，法的な相談に回答するための関連法律予測に取り組んだ．法的な相談事例と関連法律のデータセットを用いて GPT-3.5 をファインチューニングした結果，提案モデルの F1Score が GPT-4 に比べて僅差で下回った．また分析の結果，刑法 208 条の予測においてはベースラインを上回る正答数を示した一方，刑法 204 条についてはベースラインの正答数を下回った．学習時に使用したデータセットやハイパーパラメータの影響により，提案モデルが過学習を起こしている可能性が示唆された．

1 はじめに

1.1 背景

2022 年に登場した OpenAI の ChatGPT[1]といった LLM の急速な発展により，法律分野における LLM の活用が期待されている．同年，リーガルスケープ社が GPT-4 をベースに開発したシステムは日本の司法試験の一部科目において合格水準に達する回答率を出している[2]．また，法律文書処理の国際コンテスト COLIEE （Competition for Legal Information Extraction and Entailment）[3]では，関連条文を問題文に当てはめて Yes／No の解答を導き出すタスクにおいて，70%の正答率が出ている[4]．これらに関連して，司法試験の自動回答を目的とした研究は多く取り組まれてきた[5, 6]．一方で，法律分野においては法律相談の自動化のニーズも存在している．2021 年度に無料法律相談の件数は 31 万件に達しており，年々増加傾向にある[7]．弁護士ドットコムのチャット法律相談[8]や，Legal AI の AI 弁護士ツール[9]など国内での関連サービスが出てきている中，学術領域で法律相談システムの研究を行なっている事例は海外の研究がほとんどである[10]．そこで本研究では，日本の法律を対象とした法律相談システムの構築を目指す．

1.2 目的

法的なトラブルは大きく民事と刑事に分かれるが，本研究では共同研究先の専門分野である刑事事件に注目する．また，刑事事件の中には様々なカテゴリの事件が存在するが，仮説検証を簡易化するため，まずは刑事事件の中でも件数の多い傷害事件に注目し，コアとなる法律である刑法 204 条（傷害罪）と刑法 208 条（暴行罪）の予測に取り組む．図 1 は本研究が目指す相談者とシステムの対話イメージ図である．

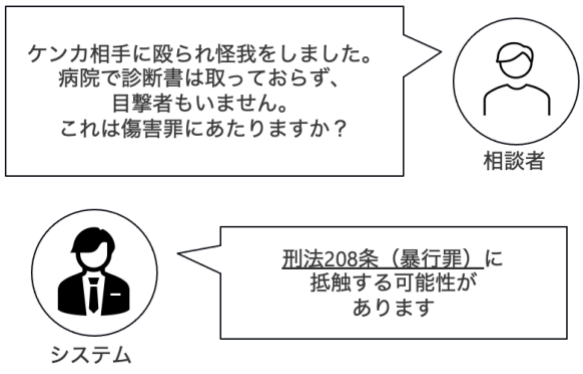

図 1. 対話イメージ

204 条と刑法 208 条の違いは，端的にいうと「暴行の結果，怪我を証明できるかどうか」という点にあるが，現状の GPT-4 はこの点を上手く処理しきれずミスをするケースが多い．したがって，本研究ではこのミスをなくすことを目標とし，的確な応答が

できる法律相談システムの構築を目指す．この目標が達成できれば，対象範囲を広げることで幅広い罪名予測が可能になるだけでなく，より詳細な量刑予測や法プロセス予測が可能になる．

2 事例-関連法学習モデルの構築

本研究では傷害事件の関連法律予測を目的とし，刑法 204 条（傷害罪）と刑法 208 条（暴行罪）の予測モデル（以下，事例-関連法学習モデル）を構築した．

2.1 事例-関連法データセット

事例-関連法学習モデルを構築するために，学習に使用するデータセットを構築した．データセットの構築にあたっては，弁護士が作成した法律に関する解説記事データを使用した．解説記事では，特定の法律に関する解説とそれが適用される事例が含まれている．今回は傷害事件に注目しているため，傷害罪や暴行罪に関連する記事だけを収集して使用した．収集した記事数は 142 件である．

次に，GPT-4 を用いて収集した記事から事例と関連法律の抽出を行なった．GPT-4 への入力は記事の全文である．表 1 に抽出した事例と関連法律の例を示す．

表 1. 抽出した事例と関連法律のペア

事例	関連法律
酒に酔って，喧嘩してしまい，相手を殴ってしまった	刑法 208 条，刑法 204 条

また，抽出した関連法律に誤りが生じていないか刑事事件専門の弁護士が確認し，修正を行った．抽出したペアデータは合計 137 件であり，刑法第 204 条と刑法第 208 条に該当するデータは，それぞれ 46 件ずつとなっている．

2.2 追加学習

事例-関連法学習モデルを構築するために，2.1 項の事例-関連法データセットを使用し，ファインチューニングを行う．ファインチューニングを行うモデルは OpenAI が提供する gpt-3.5-turbo である[11]．ファインチューニングのための前処理として，データセットを指定のフォーマットに整形し，学習に不要な情報の除去や表記揺れを解消した．表 2 に，学習時のハイパーパラメータを示す．

表 2. ハイパーパラメータ

Epochs	3
Batch Size	1
Learning Rate Multiplier	2

3 モデル評価

3.1 評価方法

事例-関連法学習モデルの評価を行うために，テストデータを作成し，F1Score を算出する．ベースラインはファイチューニングを行なっていない GPT-4 とし，事例-関連法学習モデルとの性能を比較する．

テストデータは，実際の法律相談事例に対して，弁護士が関連法律を回答した，事例-関連法ペアデータである．データ例を表 3 に示す．

表 3. テストデータ

事例	関連法律
ケンカ中に携帯を取られたため，相手に組みつき携帯を取り返そうとしたところ，相手が転倒し傷害罪で被害届をだされました． これは傷害ですか？	刑法第二百四条

テストデータは合計 44 件で，テストデータに含まれる刑法第 204 条（傷害罪）と刑法第 208 条（暴行罪）の数は，それぞれ 28 件ずつとなっている．表 4 にテストデータの内訳を示す．

表 4. テストデータの内訳

204 条のみ	208 条のみ	204 かつ 208 条	合計
16	16	12	44

3.2 評価結果

3.1 項の方法でベースラインと事例-関連法学習モデルの精度を比較した．表 5 にその結果を示す．

表 5. モデル評価結果

Model	F1Score
GPT-4	0.62
事例-関連法学習モデル	**0.61**

F1Score は事例-関連法学習モデルがベースラインを僅差で下回る結果になった．

3.3 結果分析

テストデータに対する予測結果を確認し，ベースラインと比較して事例-関連法学習モデルでどのような改善がみられたのか，または課題が残されているのかを分析した．表 6 にベースラインと事例-関連法学習モデルで刑法 204 条と刑法 208 条がそれぞれ何件ずつ予測されたのかを示す．

表 6. モデル予測結果

Model	刑法 204 条	刑法 208 条
GPT-4	32	16
事例-関連法学習モデル	10	34

事例-関連法学習モデルで改善がみられたのは，刑法 208 条の予測である．刑法 208 条が正解に含まれている場合に，ベースラインで予測できたが，事例-関連法学習モデルで予測できなかったというケースは 2 件であったのに対し，ベースラインで予測できなかったが，事例-関連法学習モデルで予測できたというケースは 13 件あった．したがって，刑法 208 条の予測については精度が向上していることが分かる．

一方で，事例-関連法学習モデルの課題は刑法 204 条の予測である．刑法 204 条が正解に含まれている場合に，ベースラインで予測できたが，事例-関連法学習モデルで予測できなかったというケースが 16 件であったのに対し，ベースラインで予測できなかったが，事例-関連法学習モデルで予測できたというケースは 1 件のみであった．したがって，刑法 204 条の予測については精度が低下していることが分かる．

3.4 考察

ファインチューニング時に使用した学習データの分布が均衡であったにも関わらず，事例-関連法学習モデルの予測結果が刑法 208 条に偏っていることから，事例-関連法学習モデルでは刑法 208 条に対する過学習が起きてしまっている可能性が考えられる．この過学習に対応するためには，まず，学習時のハイパーパラメータの探索をする必要がある．また，学習に使用したデータセットを拡張・改変することも有効だと考えられる．データセットの改変については，現在学習時に使用しているデータとテストデータの相談文の質が異なるため，これらの質が均質になるように改変することでモデルの予測精度の向上が期待できる．

4 おわりに

本研究では，法律相談システムの構築に向けた第一歩として，刑法 204 条（傷害罪）と刑法 208 条（暴行罪）の予測に取り組んだ．法的な相談事例と関連法律のデータセットを用いて GPT-3.5 をファインチューニングし，事例-関連法学習モデルを構築した．モデル評価では，事例-関連法学習モデルの F1Score がベースラインの GPT-4 に比べて僅差で下回る結果となった．結果分析では，刑法 208 条の予測においてはベースラインモデルを上回る正答数を示した一方，刑法 204 条についてはベースラインの正答数を下回った．モデルの精度向上に向けた今後の取り組みとしては，事例-関連法学習モデルは刑法 208 条を過学習している可能性があるため，学習に使用したデータセットの質と学習時のハイパーパラメータを見直し，精度向上を図っていく予定である．

参考文献

1. OpenAI. ChatGPT. (引用日:2024 年 1 月 7 日.) https://openai.com/blog/chatgpt.
2. 日本経済新聞. 生成 AI が司法試験「合格水準」 東大発新興、一部科目で. (引用日:2024 年 1 月 7 日.) https://www.nikkei.com/article/DGXZQOUC317WP0R30C23A5000000/.
3. COLIEE-2023. (引用日:2024 年 1 月 7 日.) https://sites.ualberta.ca/~rabelo/COLIEE2023/.
4. NII Today. 人工知能法学を識る. NII Today 第 97 号. (引用日:2024 年 1 月 7 日.) https://www.nii.ac.jp/today/97/4.html.
5. 藤田真伎, 狩野芳伸. 司法試験自動解答におけるルールベースと機械学習のシステム構築とアンサンブル. 言語処理学会第 28 回年次大会発表論文集, pp.1840-pp.1845, 2022.
6. 星野玲那, 狩野 芳伸. 司法試験自動解答を題材にした BERT による法律分野の含意関係認識. 言語処理学会第 26 回年次大会発表論文集, pp.577-580, 2020.
7. 日本弁護士連合会. 民事法律扶助援助実績件数. (引用日:2024 年 1 月 7 日.) https://www.nichibenren.or.jp/library/pdf/document/statistics/2022/6-2-2.pdf.
8. 弁護士ドットコム. チャット法律相談 (α). (引用日:2024 年 1 月 7 日.) https://chat.bengo4.com/.
9. Legal AI. AI 弁護士法律相談ツール. (引用日:2024 年 1 月 7 日.) https://legalai.co.jp/#/
10. Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, Li Yuan. ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases. arXiv:2306.16092v1 [cs.CL], 2023.
11. OpenAI. GPT-3.5. (引用日:2024 年 1 月 7 日.) https://platform.openai.com/docs/models/g pt-3-5