大規模言語モデルを用いた日本語判決書の自動要約

新保彰人　菅原裕太　山田寛章　徳永健伸
東京工業大学 情報理工学院
{shimbo.a.aa@m, sugawara.y.ag@m, yamada@c, take@c}.titech.ac.jp

概要

日本語判決書の自動要約の需要の高まりに伴って，大規模言語モデル（LLM）によって高品質な判決書の要約文を出力することが期待されている．本研究では One-shot 文脈内学習に用いるサンプルを近傍事例検索を用いて選ぶ手法を提案する．ベースライン手法と比較し，提案手法を用いることによって判決書要約の精度が高まることを示す．

1　背景

日本において，民事判決のデータベース化の構想が進んでいる [1]．従来データとして公開されていた民事判決は限定的であり，2017 年に下された民事判決のうち裁判所の公式サイトに掲載された判決はわずか 0.03% であった [2]．今後データベース化が進むと今までに比べて膨大な数の判決書にアクセスできるようになり，司法分野の業務に対して自然言語処理の活用による自動化・効率化の必要性が高まる．自動要約はそのような活用の一つである．

今日，判例集に掲載される判決書の多くには，事案の概要・裁判所の判断・判断の理由を法律の専門家が要約した**判示事項**（例：表 3）が付与されている．判示事項は弁護士や裁判官などが事例を分析する際の手掛かりとなるが，今後利用可能になると見込まれる大量の判決書全てに対して人手で判示事項を記述するのは困難になると考えられる．このような背景から判決書の自動要約の需要が高まっている．

近年，LLM は要約を含む多様なタスクにおいて高い精度を記録している．Few-shot 文脈内学習は LLM を活用する手法の一つであり，入力文と一緒にタスクのサンプルを入力する．先行研究では，Few-shot 文脈内学習におけるサンプルの選び方はモデルのパフォーマンスに影響すると報告されている [3, 4]．本研究では近傍事例検索を用いて要約したい判決書に似ている判決書をサンプルとして選ぶ手法を提案する．判決書は文書長が非常に長く，複数のサンプルを入力するのは非現実的であることから，本研究では 1 つのみのサンプルを入力する One-shot 文脈内学習を対象として実験する．提案手法の他に，ファインチューニングを行う手法の実験も行いそれらのパフォーマンスとコストの比較を行う．

2　データセット

実験には株式会社 LIC によって提供されたデータセットを用いる．このデータセットは学習セット，検証セット，テストセットからなり合計で 84,938 文書を含む．このデータセットの中から実験に適した判決書を得るため以下の基準で文書を選別した．

文書長：データセットには非常に長い判決書が含まれ，平均トークン長は 11,245 であり，最大トークン長は 1,087,405 に及ぶ．一方で本研究で利用する GPT-3.5-turbo-16k-0613 [5] の最大入力長は 16,384 トークンであるため，モデルの入力長制限に適した判決書のみ抽出する．入力長制限のうち 5%をプロンプトとサンプル要約と出力要約のために割り当て，残りの 95%をサンプル判決書と要約したい判決書にそれぞれ半分ずつ割り当てることにした．したがって判決書のトークン長の上限を入力長制限の 47.5%，すなわち 7,782 トークンとして抽出した．トークナイザは `cl100k_base` エンコーディングに基づく tiktoken トークナイザ[1)]を利用した．この条件で抽出すると 47,584 判決書が残り，平均トークン長は 3,637 となった．

様式：判決書には平成 2 年 1 月以降から使われている新様式とそれ以前から使われている旧様式が存在する [6]．今後公開される判決書の多くでは新様式となることが見込まれるため，旧様式を除外する．この結果，判決書数は 47,584 から 6,713 に減少した．

民事訴訟の種類：民事訴訟にはいくつかの種類があり，このデータセットには 58 種類が含まれる．

1） https://github.com/openai/tiktoken

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1 選別後の判決書数

	学習	検証	テスト
判決書数	2,055	244	288

多い順に 3 つを並べると，民事通常訴訟事件，民事控訴事件，民事抗告事件であり，これらだけでデータセットの 68.2%を占める．控訴事件と抗告事件は要約のために第一審の情報が必要になる可能性があるが，判決書ではその情報は省略されているため控訴事件と抗告事件の要約は難しい．このため，本研究では民事通常訴訟事件のみを対象とすることとした．この選別の結果 2,587 判決書が残った．

選別後の判決書数を表 1 に示す．このデータセットに含まれる判決書には法律の専門家によって書かれた判示事項が付属している．判示事項は事案の概要と裁判所の判断と判断の理由の 3 つの要素を要約したテキストである．

3 手法

本論文では **OneShot-NN** と **Regeneration** の 2 つの手法を提案する．どちらの手法も要約元の判決書に似た判決書をサンプルとして選ぶ手法である．それらに加えてファインチューニングの実験も行う．

3.1 OneShot-NN

要約元の判決書に似ている判決書を見つけるために，文書埋め込みに基づく近傍事例検索を用いる．文書同士の類似度は埋め込みのコサイン類似度で計算する．しかし，本研究で扱う判決書は平均で 4,393 文字の長い文書であるため，判決書全体を埋め込むのは現実的ではない．そこで OneShot-NN では判決書中に含まれる**事案の概要**（例：表 3）を用いて埋め込みを得る．事案の概要は判決書中で述べられている事実と争点を要約したテキストであり，判決書の冒頭部分に記述されていることが多い．事案の概要は判示事項とは異なり，裁判所の判断と判断の理由を含まないため判示事項を代替できない．
事案の概要は以下の方法で抽出する．

1. 文字列「事案の概要」を判決書中で検索する．
2. 文字列「事案の概要」を含む行とそれに続く 4 行から，正規表現「本件は.+。」にマッチする文字列を抽出する．

以上の方法で抽出した事案の概要を使って，学習セットから近傍事例を抽出する．そのために実験準備として学習セットの各判決書の事案の概要の埋め込みを計算しておく．準備しておいた学習セットの各判決書の事案の概要の埋め込みと，要約元の判決書の事案の概要の埋め込みの間のコサイン類似度を計算して，最もコサイン類似度が高い判決書を学習セットから選ぶ．このようにして選ばれた判決書とその判決書に付されている判示事項を One-shot 文脈内学習におけるサンプルとして用いる．埋め込みモデルとして Multilingual-E5-Large [7] を利用する．埋め込みの次元は 1,024 である．

事案の概要は正規表現で抽出できない場合があり，そのような場合は次の 2 つの方法でサンプルを選ぶ．**RandomDefault** では学習セットからランダムにサンプルを選ぶ．一方で，**FixedDefault** では次の方法で事前に選んだサンプル s_{best} を固定して使う．

1. ランダムに 100 判決書 S をサンプル候補として学習セットから選ぶ．
2. ランダムに 10 判決書 T を要約元判決書として学習セットから選ぶ．
3. 各 $t \in T$ を各 $s \in S$ をサンプルとして用いて要約する．出力要約文を $\text{Summary}(s,t)$ とする．
4. $s_{\text{best}} = \text{argmax}_{s \in S} \text{Score}(\text{Summmary}(s,t))$ として s_{best} を選ぶ．ただし $\text{Score}(\cdot)$ は $\text{Summary}(s,t)$ の標準化された ROUGE-1,2,3,L と BERTScore の和である．

3.2 Regeneration

Regeneration(-Regen) は OneShot-NN や他のベースライン手法に追加して行われるモジュールである．この手法では事案の概要ではなくモデルが生成した判示事項を利用して近傍事例検索を行う．判示事項は事案の概要に含まれない情報を含むため，事案の概要を用いて近傍事例検索を行う場合よりも効果的なサンプルを選ぶことができると期待する．
準備として，学習セット中の人手で書かれた判示事項の埋め込みを得る．要約を行う際には，まず何らかの方法で要約文を生成し，その要約文の埋め込みを得る．その埋め込みを利用して OneShot-NN と同様の方法で学習セットから近傍事例を抽出して，抽出された近傍事例を One-shot サンプルとして利用して再び要約文を生成する．Regeneration は何度も繰り返して適用できる手法であるが，複数回繰り返す実験を行っても有意な改善が見られなかったことから 1 回適用した場合の結果のみを報告する．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

さらに，本論文では Regeneration の派生手法として **SelectiveRegeneration(-SelectiveRegen)** を提案する．この手法では OneShot-NN において事案の概要が抽出できなかった場合のみ，Regeneration を行う．

3.3 Fine-tuning

Fine-tuning では学習セットを使って LLM をファインチューニングする．

3.4 ベースライン

ベースライン手法として 3 種類の実験を行う．**ZeroShot** は最も単純な手法で，サンプルを入力せず要約元の判決書のみを入力する．**OneShot-Random** はランダムに選んだ判決書とその判示事項をサンプルとして入力する．**OneShot-Fixed** では OneShot-NN-FixedDefault（3.1）で事前に選んだ判決書と同じものをサンプルとして入力する．

OneShot-Fixed で利用するサンプルは，事前の少量データ実験で質の高い要約文を出力することを確認しているため，OneShot-Fixed は OneShot-Random よりも高い精度となることが期待できる．

4 実験設定

要約モデルとして GPT-3.5-turbo-16k-0613 を採用する．ただし，Fine-tuning では GPT-3.5-turbo-16k-1106 を採用する．パラメータは **temperature** を 0 に設定し，他のパラメータはデフォルトとする．システムプロンプトは「あなたは法律の専門家です。あなたは判決文の要約文を書く専門家です。訴訟費用の負担については言及しないでください。敬語は絶対に使わないでください。裁判所視点で要約してください。主語は裁判所にしてください。」とし，ユーザプロンプトは「次の判決文はどのような事例を扱ったものか教えてください。50 字程度で書いてください。敬語は絶対に使わないでください。」とする．事前の実験で良い出力が得られることを確認したため，これらのプロンプトを選定した．各手法は 5 回ずつ行い，それらの平均の結果を報告する．ZeroShot ではユーザプロンプトと要約元の判決書を結合して入力する．OneShot ではユーザプロンプト，サンプル判決書，サンプル判示事項，ユーザプロンプト，要約元の判決書を結合して入力する．Fine-tuning は OpenAI API を利用して行った．モデルの学習時は ZeroShot と同様にモデルに入力し，モデルの出力の正解テキストとして判示事項を与える．推論時は ZeroShot と同様にして要約文を得る．ファインチューニングは 3 エポック行う．

5 結果

実験結果を表 2 に示す．人手で書かれた判示事項を正解データとして，モデルが生成した要約文と比較して ROUGE-1,2,3,L と BERTScore で評価した．ROUGE の値は F1 値を計算している．表中の値は 5 回実験した平均を示す．太字は最も良い値を示す．

Fine-tuning が最も良いスコアを示した．学習セット全体を利用して学習を行っているのは Fine-tuning だけであり，モデルのパラメータの更新を伴うため，最高のスコアを達成したと考えられる．以下では，Fine-tuning 以外の手法のみを比較して議論する．

OneShot-Fixed は全ての指標で OneShot-Random よりも高い値を記録しているが，OneShot-NN-FixedDefault は OneShot-NN-RandomDefault よりも低い結果を記録している．OneShot-NN-FixedDefault で選ばれたサンプルはランダムに選ばれたサンプルよりも，効果的であると期待したが結果は期待に反するものとなった．この結果は普遍的に有用な One-shot サンプルを事前に選ぶことが困難であることを示唆している．

OneShot-NN と Regeneration は異なるテキストを基にして要約元の判決書に似ている判決書を検索する．判示事項の方がより多くの情報を含むため，判示事項を利用して近傍事例検索を行う Regeneration の方がより良い結果を記録すると期待したが，結果は期待に反したものとなった．OneShot 系モデルにおいて，NN または Regeneration の何れかのみを行う手法間で比較すると，NN 適用モデルがより良い結果を示した．また，Fine-tuning 以外の手法の中で最も良い結果を示したのは SelectiveRegeneration であった．したがって，一般に判示事項生成のための近傍事例探索の手掛かりとしては事案の概要が有用であるものの，判示事項の利用ができない場合にはモデルが出力した判示事項を利用して近傍事例検索をすることが有効であることが確認できた．

ここで，コストについて比較する．OpenAI API のコストは 2023 年 1 月 9 日現在のものである．今回の実験設定でファインチューニングに必要な API 利用コストは約 1,064 ドルである．一方，本研究で実験した他の手法ではこのコストはかからない．さらにファインチューニングして構築したモデルを利用する際はトークンあたりのコストが高くな

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2　実験結果

Method \ Evaluation Metric	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-L	BERTScore
ZeroShot	37.20	14.23	7.21	25.31	71.16
ZeroShot-Regen	37.29	14.71	7.24	26.77	72.76
OneShot-Random	37.51	14.59	7.11	26.16	72.31
OneShot-Random-Regen	37.69	15.20	7.55	26.96	73.02
OneShot-Fixed	37.61	15.13	7.52	26.99	72.35
OneShot-Fixed-Regen	36.76	14.88	7.54	26.72	72.88
OneShot-NN-RandomDefault	38.37	15.43	7.75	27.33	72.80
OneShot-NN-RandomDefault-Regen	37.90	15.21	7.39	27.27	73.09
OneShot-NN-RandomDefault-SelectiveRegen	38.59	15.59	7.83	27.68	73.04
OneShot-NN-FixedDefault	38.20	15.32	7.58	27.10	72.47
OneShot-NN-FixedDefault-Regen	37.95	15.42	7.60	27.32	73.09
OneShot-NN-FixedDefault-SelectiveRegen	38.43	15.60	7.88	27.49	72.97
Fine-tuning	**47.63**	**27.86**	**19.34**	**38.26**	**77.97**

る．ZeroShot では 1 判決書につき平均で約 0.0056 ドルかかるが，Fine-tuning では 3 倍の約 0.017 ドルかかる．OneShot の設定では ZeroShot の約 2 倍の 0.011 ドル程度がかかるが，Fine-tuning よりコストは小さい．ただし OneShot-Regeneration を行う場合は，2 回モデルを推論に使うことになるため，通常の OneShot の 2 倍の 0.22 ドルが 1 判決書あたりのコストとなり Fine-tuning の推論コストを上回る．コストとパフォーマンスのバランスを考慮すると，ファインチューニングのコストを許容できる場合は Fine-tuning の手法を採用し，許容できない場合は OneShot-NN-FixedDefault-SelectiveRegen などの手法を採用するのが望ましいと考えられる．

6　関連研究

文書要約は抽出型要約と抽象型要約に大別され，法律ドメインでも両者の研究が行われている．抽出型要約の既存研究として，阪野ら [8] の研究がある．阪野らは「手がかり表現」というキーワードを用いて，判決書中の段落を「事実関係」や「論旨」などの要素に分類してから重要文を抽出する手法を提案している．また，Agarwal ら [9] はアメリカの判決書を対象として言語モデルを用いた抽出型要約の手法を提案している．抽象型要約の研究として Elaraby ら [10] の研究が挙げられる．この研究では seq2seq モデルで要約する前に，各文の役割を特殊トークンでアノテーションすることによって，生成される要約文の質が高まることが報告されている．

LLM を用いて法律関係の文書を利活用する研究も行われている．Deroy ら [11] の研究ではインドの判決書を対象とし，LLM を用いた判決書の自動要約が可能かを検証している．Zin ら [12] の研究では LLM を用いて日本語の契約書から情報を抽出する手法を提案している．この研究では抽出したい情報のクエリを入力するクエリ指向要約というアプローチをとっている．

7　結論

本論文では，近傍事例検索で One-shot 文脈内学習のサンプルを選ぶ際に判決書中に含まれる事案の概要を利用することが有用であることを示した．また，モデルが出力した要約文を利用した近傍事例検索が有用であることも示した．さらに，ファインチューニングが他の手法よりも優れた結果を示すことを確認し，各手法のコストの比較をした．

今後の展望としては，One-shot サンプルを選ぶ他の方法の検討が考えられる．例えば，サンプルを選択するモデルを強化学習によって学習する方法が考えられる．強化学習では ROUGE などの値を報酬として設定することにより，これらのスコアを改善することを目標とする学習を行うことができるため，より質が高い要約を得られる可能性がある．

出力文における幻覚 (hallucination) の解消も今後の課題として挙げられる．幻覚は自動評価で捉えるのが難しく，人手による評価が必要であると考えられる．このため，モデルの出力に対して法律の専門家のフィードバックを貰い，どのような幻覚があるかを分析する予定である．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は株式会社 LIC の支援を受けたものである.

参考文献

[1] 民事判決情報データベース化検討会について. https://www.moj.go.jp/content/001382374.pdf. Accessed: 2014-01-10.

[2] 谷川和幸. 判例情報のオープンデータ化.

[3] Subha Vadlamannati and Gözde Gül Sahin. Metric-based in-context learning: A case study in text simplification. In **INLG Oral Session 3: Leveraging Large Language Models for NLG**, 2023.

[4] Rajaswa Patil, Manasi Patwardhan, Shirish Karande, Lovekesh Vig, and Gautam Shroff. Exploring dimensions of generalizability and few-shot transfer for text-to-sql semantic parsing. In Alon Albalak, Chunting Zhou, Colin Raffel, Deepak Ramachandran, Sebastian Ruder, and Xuezhe Ma, editors, **Transfer Learning for Natural Language Processing Workshop, 03 December 2022, New Orleans, Louisiana, USA**, Vol. 203 of **Proceedings of Machine Learning Research**, pp. 103–114. PMLR, 2022.

[5] OpenAI. Openai gpt-3 api [gpt-3.5-turbo]. https://platform.openai.com/docs/guides/text-generation/chat-completions-api, 2023.

[6] 家腹尚秀. 民事判決書の在り方についての一考察. 東京大学法科大学院ローレビュー編集委員会, 2015. http://www.sllr.j.u-tokyo.ac.jp/10/papers/v10part05(iehara).pdf.

[7] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. **CoRR**, Vol. abs/2212.03533, , 2022.

[8] 阪野慎司, 松原茂樹, 吉川正俊. 手がかり表現に基づく判決文の自動要約. Nagoya, Japan, March 2005. 言語処理学会第 11 回年次大会.

[9] Abhishek Agarwal, Shanshan Xu, and Matthias Grabmair. Extractive summarization of legal decisions using multi-task learning and maximal marginal relevance. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 1857–1872, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[10] Mohamed Elaraby and Diane Litman. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, **Proceedings of the 29th International Conference on Computational Linguistics**, pp. 6187–6194, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[11] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. How ready are pre-trained abstractive models and llms for legal case judgement summarization? In Jack G. Conrad, Daniel W. Linna Jr., Jason R. Baron, Hans Henseler, Paheli Bhattacharya, Aileen Nielsen, Jyothi K. Vinjumur, Jeremy Pickens, and Amanda Jones, editors, **Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023) co-located with the 19th International Conference on Artificial Intelligence and Law (ICAIL 2023), Braga, Portugal, June 19, 2023**, Vol. 3423 of **CEUR Workshop Proceedings**, pp. 8–19. CEUR-WS.org, 2023.

[12] May Zin, Ha-Thanh Nguyen, Ken Satoh, Saku Sugawara, and Fumihito Nishino. Information extraction from lengthy legal contracts: Leveraging query-based summarization and gpt-3.5. In **JURIX 2023: The Thirty-sixth Annual Conference, Maastricht, the Netherlands, 18–20 December 2023**, Vol. 379 of **Frontiers in Artificial Intelligence and Applications**, pp. 177–186. IOS Press, 2023.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

付録

判示事項と事案の概要の例

表 3　判示事項と事案の概要の例

判示事項	原告が被告に対し，主位的に，通行地役権に基づき，予備的に，債権者代位権に基づく通路所有者の権利の代位行使として，通行妨害行為の禁止を求めた事案で，原告の通行地役権の行使は，被告車両の駐車により妨害されているとはいえないから，原告に，本件通路土地所有者である乙を代位して，妨害排除ないしは引渡しを請求することを認めるべき合理的理由はなく，原告の権利を保全するため，債権者代位権の行使を認める必要はないとした事例
事案の概要	本件は，原告が被告に対し，主位的に，通行地役権に基づき，予備的に，債権者代位権に基づく通路所有者の権利の代位行使として，通行妨害行為の禁止を求めた事案である。

各手法の出力例

表 4　各手法の出力例

手法	出力文
正解文（判示事項）	貸金業の登録を受けていないリホーム業者が資金の借り受けの仲介に基づきホーム注文者に対して有する媒介手数料請求権が自然債務であるとされた事例
ZeroShot	この判決文は、金銭貸借の媒介手数料と貸付金の返済を巡る訴訟であり，原告の訴えと被告の反訴をいずれも棄却することが判決されました。また，訴訟費用は原告と被告の間で二分されることが決定されました。
OneShot-Random	原告が、貸金業者である又は貸金業者であった被告らに対し、金銭消費貸借契約の返済として支払った金員が、利息制限法を超過するものであったと主張して、不当利得返還請求権に基づき、同超過分の返還を求め、認容された事例。
OneShot-Fixed	原告と被告の間で金銭貸借の媒介に関する争いがあり、原告は被告に対し媒介手数料の支払いを求め、被告は原告に対し不当利得の返還を求めたが、いずれの請求も棄却された事例。
OneShot-NN-RandomDefault	金銭貸借の媒介手数料として支払った金銭が不当利得として返還されるべきかを争う事案で、被告は原告が貸金業法に違反していると主張している。裁判所は、原告が無登録で金銭貸借の媒介を行っていたことを認め、媒介契約は無効であると判断した。そのため、原告の媒介手数料の請求は認められず、被告の反訴請求も棄却された。
OneShot-NN-RandomDefault-SelectiveRegen	原告が、貸金業者である又は貸金業者であった被告らに対し、金銭消費貸借契約の返済として支払った金員が、利息制限法を超過するものであったと主張して、不当利得返還請求権に基づき、同超過分の返還を求め、認容された事例。
Fine-tuning	貸金業法に定める貸金業の登録を受けていない者との間で締結した金銭貸借の媒介契約に基づく媒介手数料の支払い請求権は自然債務であるとされた事例

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).