RAG に基づく韓国語法令・判例に対する質問応答

徐基皓[1] 宇津呂武仁[2]
[1]韓国 警察庁 中央警察学校 [2]筑波大学 システム情報系 知能機能工学域
seokiho_@_.police.go.kr, utsuro_@_iit.tsukuba.ac.jp

概要

本論文では，大規模言語モデルを活用して，韓国語における刑事業務に関連する法規の解釈に基づく質問応答の性能を改善する手法を提案する．本論文においては，インターネットで公開されている刑事業務に関する法令の条文および判例情報を蓄積し，RAG(Retrieval-Augmented Generation，検索拡張生成) の枠組みのもとで質問に関連する法令条文および判例を検索し，これらの関連法令・判例をふまえて，大規模言語モデルによって質問に関する推論を行うことにより，質問に対する的確な回答を生成するシステムを開発する．本システムの応用事例として，韓国の刑法分野において，捜査における意思決定および法的解釈の局面での支援用途が挙げられる．

1 はじめに

近年，法務分野における大規模言語モデル (LLM: Large Language Models) [8] の活用が注目されており，特に法令解釈や判例分析において，その可能性が期待されている．韓国の刑事司法制度では，法改正や判例の蓄積により参照すべき法的情報が増加の一途を辿っていることから，捜査官が迅速かつ正確な意思決定を行うことは容易ではない．ここで，従来のキーワードを用いた検索システムでは，法的文書の文脈や意味的な関連性を十分に考慮できないことから，特に新任捜査官に対して，キーワードを用いた検索システムの効果的な活用法を教育・訓練する点においては，効率的な知識習得が困難であった．このような背景をふまえて，本論文では，最新の LLM 技術を活用して，刑事業務関連の法令解釈・判例参照に基づく質問応答のモデル化を目的とする．具体的には，韓国の法制情報サイト[1)]に掲載されている法令・判例情報を対象として RAG(Retrieval-Augmented Generation，検索拡張生成) [5] の枠組みを適用する．RAG の枠組みにおいては，法令・判例情報を埋め込みベクトルに変換し，FAISS[2)] によって検索用データベースに蓄積する．そして，与えられた質問に対して，質問に関連する法令条文および判例を検索し，これらの関連法令・判例をふまえて，大規模言語モデルによって質問に関する推論を行うことにより，質問に対する的確な回答を生成するシステムを開発する．RAG の枠組みにより，LLM の幻覚問題を緩和しつつ，捜査における意思決定の質の向上と効率化を実現することが期待される．

2 関連研究

本論文に関連して，文献 [3] においては，法的契約における重要箇所抽出タスクにおいて専門家による注釈が付与されたデータセットを公開し，タスクにおける各種モデルの性能を評価している．文献 [7] においては，法的テキストを対象として，多様な粒度の法的話題にトピック分類するモデルを提案した．文献 [6] においては，判例データを用いて，多言語の法的判断予測ベンチマークを提案している．文献 [4] においては，法ドメインにおける対話事例を対象とする情報抽出タスクについて研究を行っている．文献 [2] においては，弁護士による法律相談や文書作成における AI 活用法について分析を行っている．文献 [9] においては，訴訟における法的解釈や弁護士支援のためのツールとして，大規模言語モデルの有用性を評価した結果について述べている．

3 法令・判例データ

韓国においては，国家法令情報システム[3)] において，条文や行政規則のメタデータを提供している．本論文では，本システムの API を用い

1) http://www.law.go.kr
2) https://faiss.ai
3) https://open.law.go.kr

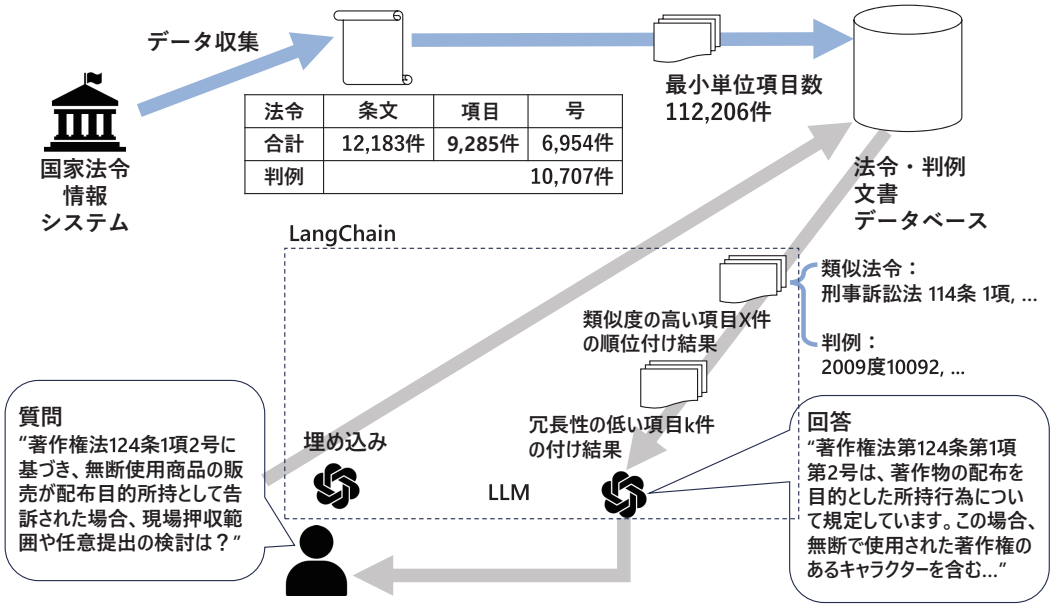

図 1　RAG に基づく韓国語法令・判例に対する質問応答の枠組み

て，法令 (条，項，号) の情報および判例情報を収集して利用する．その結果，図 1 および表 1 に示すように，合計 12,183 件の条文，9,285 件の項目，6,954 件の号，および，1975 年から 2023 年までの 10,707 件の刑事事件に関する判例情報を収集した．次に，これらの条文・判例について，条文，項目，号ごとに分割して最小単位で検索用データベースに保存した結果，112,206 件の最小項目数として保存された．また，刑事業務に関連する法令 20 件では，二つ以上の法令が適用される判例や，法令の改正や名称変更により 20 件の法令に該当しない「その他」として分類された判例が 2,419 件存在した．また，「検察官と司法警察官の相互協力及び一般的捜査準則に関する規定」のように，新たに制定された法令には該当する判例が存在しなかった．

3.1　法令データ

例えば，「刑法第 43 条 (刑の宣告と資格喪失, 資格停止)① 死刑，無期懲役または無期禁錮の判決を受けた者は，次に掲げる資格を喪失する．1. 公務員となる資格」の条文は，1 つの「条文」オブジェクトとして生成され，タイトルや内容のほか, 法令名 (例:「刑法」), 条文番号 (例:「43」), 施行日，固有識別子 (例:「刑法_43」) などのメタデータを含む．また，データベースには条文全体が独立したドキュメントとして保存される．さらに，条文内の各項目 (項，号) は個別のオブジェクトに分割される．例えば，「①死刑，無期懲役または無期禁錮の判決を受けた者は ~」は「項」オブジェクトとして保存され，法令名，条文番号，項番号 (例:「①」)，固有識別子 (例:「刑法_43_①」) を持つ．同様に，「1. 公務員となる資格」は「号」オブジェクトとして保存され，号番号 (例:「1」) や固有識別子 (例:「刑法_43_①_1」) などのメタデータが付加される．このように, 刑法第 43 条は，条文，項，号の階層構造に基づいて複数のオブジェクトに分割して保存され，それぞれのオブジェクトは固有識別子とメタデータによって検索および活用可能となる．なお，「条」は単独で存在する場合があるが，「項」は条が存在しない限り存在できない．また，条文は項の区分がなくとも，単一の内容として条のみで存在することができる．「号」は項がなくても条が存在すれば成立する．

3.2　判例データ

事件に関するデータ構造は，事件の内容とメタデータを統合した形式である．例えば，「事件番号: 2023 度 2102」「事件名: 麻薬類管理に関する法律違反 (向精神薬)」「判決要旨: 公訴事実の記載は，犯罪の日時，場所および方法を明示して事実を特定 (中略)」といった形式で構成される．この構造には，事件番号，事件名，裁判所名，判決日，参照条文などの情報が含まれ，これらがメタデータとして一元的に管理される．また，事件の内容には事件番号，参照条文，判例の詳細な記述が含まれ，これにより事件情報の効

表 1　法令ごとの条文数・項目数・号数・判例数・総事件数

法令名	条文数	項目数	号数	判例数	総事件数
刑法	459	312	18	5,276	
刑事訴訟法	641	866	150	2,059	
検察官と司法警察官の相互協力及び一般的捜査準則に関する規定	87	174	98	0	
電気通信金融詐欺被害防止及び被害金返還に関する特別法	30	69	83	3	
国民体育振興法	107	251	156	9	
韓国馬事会法	86	121	122	2	
保険詐欺防止特別法	19	18	6	1	
不正請託及び金品等受領の禁止に関する法律	31	72	82	19	
道路交通法	223	474	427	274	
交通事故処理特例法	6	8	15	45	
不正小切手取締法	7	8	3	69	
弁護士法	189	305	150	88	10,283
与信専門金融業法	129	267	242	7	
不動産登記特別措置法	12	21	10	4	
情報通信網利用促進及び情報保護等に関する法律	155	298	326	17	
性暴力犯罪の処罰等に関する特例法	68	192	85	165	
特定経済犯罪加重処罰等に関する法律	14	35	12	10	
暴力行為等処罰に関する法律	10	21	14	215	
児童・青少年の性保護に関する法律	92	226	203	20	
児童虐待犯罪の処罰等に関する特例法	78	182	85	5	
その他	9,740	5,365	4,667	2,419	
合計	12,183	9,285	6,954	10,707	10,283

率的な検索および活用が可能となる．

4　RAG による質問応答

本論文では，RAG における LLM のプロンプトとしては，zero-shot および few-shot (本論文では 5-shot で評価を行った) を用い，LLM における埋め込みとしては `text-embedding-ada-002`[4] を，LLM としては，GPT モデル `gpt-4-turbo`[5] を，それぞれ使用する．また，LangChain[6] をプラットフォームとして RAG を実装する．

4.1　多重クエリ手法

多重クエリ (multi query[7]) 手法は，LLM において，一つのクエリを複数の類似したクエリに拡張することによって，文書検索の性能を向上させる手法の一つである．この手法では，元のクエリの意図を反映しつつ，下のクエリとは異なる多様な観点から N 個の代替クエリを生成することによって，より広範囲にわたる回答を得ることが可能となる．この手法によって，RAG モデルにおいて，質問や文脈を手がかりとして最適な多重クエリが生成され，関連情報を適切に取得することが実現される．本論文では，LangChain 上の多重クエリ機能を用いる．

具体例

初期クエリが，「著作権法第 124 条第 1 項第 2 号に基づき，無断使用商品の販売が配布目的所持として告訴された場合，現場押収範囲や任意提出の検討は?」の場合，この質問をより具体的に拡張するためのプロンプトとして，「多重クエリ方式で質問拡張」と指定することにより，初期クエリに関連する詳細な質問への展開が行われる．例えば，

- 『犯罪成立の構成要件とは何か?』
- 『構成要件によって可罰性はどのように判断できるのか?』

等，「様々な詳細な質問を初期クエリに追加して，より具体的な回答を提供してください.」という内容がプロンプトに含まれる．このようなプロンプトを入力することにより，システムは

4)　https://platform.openai.com/docs/guides/embeddings#embedding-models
5)　https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4
6)　https://www.langchain.com/
7)　https://blog.langchain.dev/query-transformations/

表 2　評価結果

手法	回答種類	正解率
5-shot, 多重クエリ有	曖昧	64.6% (31/48)
	明確	60.3% (108/179)
	計	**61.2**% (139/227)
5-shot, 多重クエリ無	曖昧	62.0% (31/50)
	明確	58.2% (103/177)
	計	59.0% (134/227)
zero-shot, 多重クエリ有	曖昧	60.0% (33/55)
	明確	53.5% (92/172)
	計	55.1% (125/227)
zero-shot, 多重クエリ無	曖昧	55.2% (32/58)
	明確	50.3% (85/169)
	計	51.5% (117/227)

初期クエリにもとづき追加クエリを生成し，生成された追加クエリもふまえて関連文書を検索する．

4.2　MMR を用いた類似文書の抽出

さらに，本論文では，質問に対する関連文書を抽出するために，LangChain 上の MMR (Maximal Marginal Relevance) [1][8] を用いることにより，複数の関連性が高い文書を検索する手法を採用する．以下，MMR の定式化を示す．

$$\mathrm{MMR} = \underset{D_i \in D \setminus S}{\operatorname{argmax}} \Big[\lambda \cdot \mathrm{Sim}(D_i,\ Q) - (1-\lambda) \cdot \max_{D_j \in S} \mathrm{Sim}(D_i,\ D_j) \Big]$$

- λ: 類似度と多様性の間の重み ($0 \le \lambda \le 1$)
- D: 全ての検索候補文書の集合
- S: 検索済み文書の集合
- $\mathrm{Sim}(D_i,\ Q)$: 文書 D_i とクエリ Q の類似度
- $\mathrm{Sim}(D_i,\ D_j)$: 文書 D_i と D_j の類似度

この手法により，冗長性を避け，より多様性のある情報を含む文書を選択することができる．類似度尺度としては，SBERT[9] および BERTScore[10] の和を用いる．主要なパラメータである λ (0.1 刻み)，X(検索結果において冗長な結果を削除する候補件数，$X = 10, 12, 15$)，k(検索結果において冗長な結果を削除した結果の件数，$k = 5, 7, 10$) について，評価用質問応答事例 227 件に対して二分割交差検定を通してパラメータ調整・評価を行った．

8)　https://python.langchain.com/docs/how_to/example_selectors_mmr/
9)　https://sbert.net
10)　https://pypi.org/project/bert-score

5　評価

評価用質問応答事例 227 件に対して，提案手法により回答を生成した結果に対して，第一著者が人手で評価を行い，回答が曖昧であるか明確であるかの分類を行った後，回答が正解か否かの判定を行い正解率を算出した結果を表 2 に示す．評価において，5-shot および zero-shot，および，多重クエリの有無の合計四通りの比較を行った結果においては，5-shot・多重クエリ有の場合に最も高い正解率を示した．また，5-shot・zero-shot ともに，多重クエリ有の改善効果が確認され，5-shot において 2.2 ポイント，zero-shot において 3.6 ポイントの改善が見られた．

適用法令の類似度分析においても，5-shot 学習の優位性が確認された．類似度 0.9 以上の高精度な法令適用率は，5-shot 学習において多重クエリ適用時は 19.4% (44 件)，未適用時は 20.7% (47 件) であった．これに対し，zero-shot 学習では，それぞれ，14.1% (32 件)，および，13.7% (31 件) と，約 6 ポイント低い水準にとどまった．

回答の明確さに関する分析では，すべての方式において，曖昧と判断された回答の正答率 (55.2~64.6%) が，明確と判断された回答の正答率 (50.3~60.3%) を上回るという興味深い傾向が観察された．特に 5-shot 学習では，曖昧な回答の割合が 21.1~22.0%と比較的低く，zero-shot 学習における 24.2~25.6%よりも多少は安定した回答生成が可能であることが示された．

6　おわりに

本論文では，RAG を活用して，韓国語における刑事業務に関連する法規の解釈に基づく質問応答を行う手法を提案した．特に，インターネットで公開されている刑事業務に関する法令の条文および判例情報を蓄積し，RAG の枠組みのもとで質問に関連する法令条文および判例を検索し，これらの関連法令・判例をふまえて，大規模言語モデルによって質問に関する推論を行うことにより，質問に対する的確な回答を生成するシステムを開発した．今後の課題としては，大規模かつ高品質な質問応答組事例の収集，および，法令の改正時期と対応する判例との対応付けの管理が挙げられる．

参考文献

[1] Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In **Prof. 21st SIGIR**, page 335–336, 1998.

[2] Jonathan H. Choi, Amy Monahan, and Daniel B. Schwarcz. Lawyering in the age of artificial intelligence. **109 Minnesota Law Review (Forthcoming 2024), Minnesota Legal Studies Research Paper**, (23-31):1–65, 2023.

[3] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: an expert-annotated NLP dataset for legal contract review. In **Prof. 35th NeurIPS**, 2021.

[4] Jenny Hong, Derek Chong, and Christopher Manning. Learning from limited labels for long legal dialogue. In **Proc. Natural Legal Language Processing Workshop 2021**, pages 190–204, 2021.

[5] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In **Prof. 34th NIPS**, pages 9459 – 9474, 2021.

[6] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In **Proc. Natural Legal Language Processing Workshop 2021**, page 19–35, 2021.

[7] Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. Multi-granular legal topic classification on Greek legislation. In **Proc. Natural Legal Language Processing Workshop 2021**, page 63–75, 2021.

[8] Bhawna Singh. Introduction to large language models. In **Building Applications with Large Language Models: Techniques, Implementation, and Applications**, pages 1–25. Apress, 2024.

[9] Arianna Trozze, Toby Davies, and Bennett Kleinberg. Large language models in cryptocurrency securities cases: can a GPT model meaningfully assist lawyers? **Artifcial Intelligence and Law**, 2024.