Answer retrieval in legal community question answering

Askari, A.; Yang, Z.; Ren, Z.; Verberne, S.; Goharian, N.; Tonellotto, N.; ... ; Ounis, I.

Answer Retrieval in Legal Community Question Answering

Arian Askari(✉), Zihui Yang, Zhaochun Ren, and Suzan Verberne

Leiden University, Leiden, The Netherlands
{a.askari,z.ren,s.verberne}@liacs.leidenuniv.nl,
zoeyangzihui@gmail.com

Abstract. The task of answer retrieval in the legal domain aims to help users to seek relevant legal advice from massive amounts of professional responses. Two main challenges hinder applying existing answer retrieval approaches in other domains to the legal domain: (1) a huge knowledge gap between lawyers and non-professionals; and (2) a mix of informal and formal content on legal QA websites. To tackle these challenges, we propose CE_{FS}, a novel cross-encoder (CE) re-ranker based on the **f**ine-grained **s**tructured inputs. CE_{FS} uses additional structured information in the CQA data to improve the effectiveness of cross-encoder re-rankers. Furthermore, we propose *LegalQA*: a real-world benchmark dataset for evaluating answer retrieval in the legal domain. Experiments conducted on LegalQA show that our proposed method significantly outperforms strong cross-encoder re-rankers fine-tuned on MS MARCO. Our novel finding is that adding the question tags of each question besides the question description and title into the input of cross-encoder re-rankers structurally boosts the rankers' effectiveness. While we study our proposed method in the legal domain, we believe that our method can be applied in similar applications in other domains.

Keywords: Legal Answer Retrieval · Legal IR · Data collection · Fine-grained structured cross-encoder

1 Introduction

As an established problem in information retrieval, answer[1] retrieval [6] has been studied in a variety of domains, including healthcare [8], social media [6,29], and programming [31]. Community question answering (CQA) platforms provide common sources for answer retrieval [5,22,23], e.g., Stackoverflow[2] for finding answers of programming questions [5]. In the legal domain, users[3]

[1] We interchangeably use the word answer and response to refer to the content written by the professional lawyer.

[2] https://stackoverflow.com/.

[3] We refer to the person who posts a question as a user or questioner throughout this paper.

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14610, pp. 477–485, 2024.
https://doi.org/10.1007/978-3-031-56063-7_40

seek legal advice from professionals, and timely access to accurate answers is important. Legal language, often considered a sub-language due to its distinct characteristics, emphasizes the importance of answer retrieval in the legal domain [10,25,27]. However, answer retrieval has not been addressed in the legal domain yet.

In this paper, we propose the task of legal answer retrieval. We start by analyzing the effectiveness of the widely used two-stage ranking pipeline [3], which consists of an efficient retriever, e.g., BM25 [21], to retrieve a shortlist of documents from the collection, and a re-ranker [18] to increase the effectiveness of the initial ranking. On this premise, we propose cross-encoder$_{CAT}$ (CE$_{CAT}$), which uses cross-encoders for optimizing the re-ranking stage by concatenating the query and the candidate document in the input [3]. Given a new question and the corresponding initial ranked list produced by BM25, the CE$_{CAT}$ aims to effectively reorder the candidate documents from the initial ranked list to locate the most relevant responses on top of the ranked list. We fine-tune a cross-encoder re-ranker, called CE$_{FS}$, based on our novel **f**ine-grained **s**tructured inputs to learn the relevance between a pair of a question and its best answer based on structured information from the data; such a model would be able to effectively re-rank prior existing relevant responses where the best answer is not provided [30].

Our proposed method is inspired by recent studies that have shown task-level input modification could improve the effectiveness of cross-encoder re-rankers. For instance, BERT-FP [11] has shown that adding splitter tokens between each utterance of a dialogue can improve the effectiveness of BERT-based re-rankers. We are also motivated by the fact that, in legal CQA, question tags consist of important legal terms related to the legal question and using them can potentially bridge the knowledge gap between lawyer content and questioner content. It is noteworthy to mention that question tags are dynamically generated by the online platform that we have used in this paper.

For the first attempt, we release a new benchmark dataset, namely LegalQA, in this paper. Each question in LegalQA has a corresponding best answer selected by the questioner[4]. It consists of 9,846 questions and 33,670 answers responses by identified lawyers, organized in train, validation, and test splits. Our experiments on LegalQA show that CE$_{FS}$ significantly outperforms regular and strong fine-tuned cross-encoder re-rankers such as MiniLM-MSMARCO [24] which is a highly effective cross-encoder re-ranker trained on MS MARCO [17].

Our contributions are as follows: (1) We formulate the task of answer retrieval in the legal domain and release LegalQA. As far as we know, this is the first benchmark test collection for legal answer retrieval. (2) We evaluate the applicability of both probabilistic and existing effective fine-tuned cross-encoder re-rankers on legal answer retrieval. (3) We propose CE$_{FS}$ taking into account the different elements of a legal question with a fine-grained structured input that significantly outperforms regular fine-tuning cross-encoder re-rankers and the strong cross-encoder re-ranker trained on MS MARCO.

[4] https://github.com/arian-askari/AnswerRetrieval-Legal.

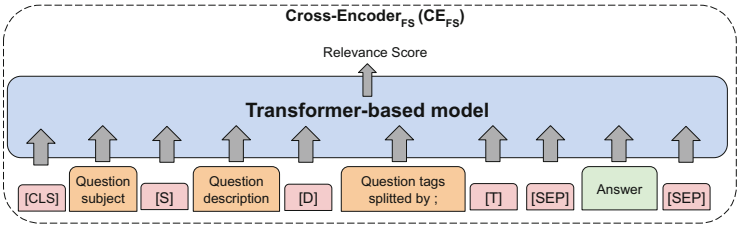

Fig. 1. Fine-grained structured input of CE$_{FS}$

2 Related Work

The objective of answer retrieval is to find the most relevant response given a question, which aligns with the core retrieval objective [5,6,22,23,29–31]. Therefore, the effective CE$_{CAT}$ re-rankers are suitable to be invested in this task as they have shown high effectiveness in addressing various core retrieval tasks [1,12,18,20]. Several studies have shown the impact of modifying the regular input of CE$_{CAT}$ could improve their effectiveness such as [2,3,7,11]. However, there are no studies that investigated the usage of question tags to improve cross-encoder re-rankers in community question-answering systems. The most relevant work to this study is the recent work by [15] that investigates the effectiveness of existing methods on legal CQA data. However, the responses in that work are written by legal forum users rather than identified lawyers, and they do not improve the effectiveness of existing methods for this particular domain. They also included the fine-tuned MiniLM on MS Marco [24] in their study and showed its poor performance in the legal domain. Furthermore, Martinez-Gil et al. [16] explore potential future directions for legal answering systems in a survey on legal systems designed for lawyers or law students including statute law retrieval and task legal textual entailment tasks. In contrast, our focus is on answer retrieval in a legal community question-answering system with a combination of legal language by lawyers and everyday language by questioners who ask questions.

3 Methods

We present our method, CE$_{FS}$, as an effective cross-encoder based re-ranker for the question answering retrieval in the legal domain. Our retrieval pipeline consists of first-stage retrieval and re-ranking. For a user's questions, the first-stage retrieval returns a set of initial responses from the dataset. These results are then re-ranked in the second step to improve the effectiveness of retrieval.

Collected Dataset for the Task. We create the LegalQA dataset by using the question URLs shared by [4] to investigate expert finding in the legal domain. In the context of the expert finding task, the input consists solely of specific

question tags, and the output is a list of experts about that specific question tag. In contrast, in answer retrieval, the standard input is the question's content, and the desired output is the most suitable answer to that question. We therefore create a new dataset, LegalQA based on the expert finding data. The LegalQA dataset is derived from a subset of the 'bankruptcy' related forum on Avvo[5] in California, spanning from January 2008 to July 2021, and comprises 9,846 questions by regular users and 33,670 answers by certified lawyers. Notably, lawyers' profiles on Avvo are associated with their real names, distinguishing them from regular users. The questions are categorized, and each category, such as 'bankruptcy,' includes questions with different category tags, for instance, 'bankruptcy homestead exemption'. Following [4], we determine a question's best answer based on whether it is selected as the most helpful by the questioner or if it receives "lawyer agree" votes from at least three certified lawyers, which is distinct from the "helpful" upvotes provided by other users. To facilitate model training and evaluation, we split the dataset into three subsets: 70% for training, 10% for validation, and 20% for testing, based on the chronological order of questions. To ensure the dataset's integrity and eliminate duplicates, we perform a lexical similarity analysis using Levenshtein distance [14] on all the questions[6]. This analysis identified 50 pairs of questions with more than a 90% overlap. In cases of high overlap between question pairs, we retain the longer question and discard the shorter one. Additionally, we reassign the responses from the removed question to the retained one.

Baselines. For the first-stage retrieval, we employ BM25 as the baseline method. Additionally, we compare BM25 to a lexical-based probabilistic first-stage retriever named LMD [19] to assess the effectiveness of BM25 compared to another lexical-based retrieval. We leave out investigating the effectiveness of Transformer-based first-stage retrievers, e.g., dense retrievers, since we focus on improving the effectiveness of the re-ranking stage of answer retrieval in the legal domain. We leave further analysis on Transformer-based first-stage retrievers to future work. For the re-ranking, we use cross-encoder re-rankers (CE_{CAT}) in different settings: (1) a fine-tuned cross-encoder on MS MARCO; (2) a fine-tuned cross-encoder on the LegalQA training set; (3) a pre-trained cross-encoder as the zero-shot baseline.

BM25 and LMD. We use BM25 as first stage retriever, a commonly used ranking function that efficiently retrieves a set of documents from the full document collection based on word overlap [9,21].

CE_{CAT}. We use CE_{CAT} as a strong re-ranker. The query q1:m and answer content a1:n sequences are concatenated with the [SEP] token, and the [CLS] token representation computed by CE is scored with a single linear layer W in the CECAT ranking model: $CE_{CAT}(q_{1:m}, a_{1:n}) = CE([CLS]\, q\, [SEP]\, a\, [SEP]) * W$. We evaluate fine-tuned both MS MARCO-trained CE_{CAT}, MiniLM-MSMARCO

[5] https://www.avvo.com/topics/bankruptcy.

[6] We use the implementation available at https://github.com/seatgeek/thefuzz.

Table 1. Effectiveness results. † denotes a statistically significant improvement of $\mathbf{CE_{FS}}$ over the second most effective re-ranker, CE_{CAT} fine-tuned on the LegalQA training set. Statistical significance was measured with a paired t-test ($p < 0.001$) with Bonferroni correction for multiple testing. All re-rankers used top-1000 retrieved answers by BM25 as the initial ranking.

Training source	Model	MAP@1k	R@1k	R@100	R@10	R@2	R@1
First-stage retrievers							
—	BM25	.120	**.542**	.354	.192	.113	.069
—	LMD	.080	.540	.321	.153	.840	.050
Cross-encoder re-rankers							
MS MARCO	CE_{CAT}	.109	**.542**	.341	.173	.101	.087
LegalQA (**ours**)	CE_{CAT}	.236	**.542**	.495	.381	.204	.181
LegalQA (**ours**)	$\mathbf{CE_{FS}}$ **(ours)**	**.270**†	**.542**	**.524**†	**.428**†	**.261**†	**.209**†

[24], and LegalQA-trained CE_{CAT} following by the above design. We use MiniLM [26] as the cross-encoder model thorough all of the experiments.

Proposed method: Cross-encoderFS (CEFS).

We propose CE_{FS} in order to capture the relevance within the question and best answer based on a fine-grained structural-based input representation tailored for cross-encoder re-rankers. In Fig. 1, we present the input representation of CE_{FS}, which is formally explained as follows:

$$CE_{FS}(q_{1:m}, a_{1:n}) = CE([CLS]\, q_{Subject}\, [S]\, q_{Description}\, [D]\, q_{Tags}\, [T]\, [SEP]\, a\, [SEP]) * W \quad (1)$$

Here, the [S], [D], and [T] tokens serve as separators, i.e., splitter tokens, for different parts of the question, namely Subject, Description, and Category tags, respectively. The input representation of CE_{FS} is designed to take into account different aspects of the retrieval context. The novelties brought by CE_{FS} can be summarized in two key aspects:

- *Structured Input.* CE_{FS} employs structured input by dividing the question into distinct sections - the subject, description, and tags. These sections are separated by splitter tokens. Such structuring not only facilitates a more comprehensive representation of the query but also emphasizes the importance of individual sections to the re-ranker since the cross-encoder in CE_{FS} can comprehend different aspects of the information and assign varying levels of importance to each section of the question.
- *Question Tags.* CE_{FS} takes the question tags into account in a straightforward yet effective manner by incorporating them into the cross-encoder re-ranker's input. Each question tag is separated by a semicolon. The motivation behind these additions is to equip the re-ranker with more comprehensive knowledge about the query, enabling it to grasp both an overview of the legal question's topic and its detailed category tags.

4 Experiments and Results

Experimental Setup. We use the Huggingface library [28] for the cross-encoder re-ranking training and inference. We add the splitter tokens into the tokenizer of the cross-encoder. Following prior work [12] we use Cross-Entropy loss [32], training batch size of 32, and Adam [13] optimizer with a learning rate of $7*10^{-6}$ for all cross-encoder layers, regardless of the number of layers trained.

Table 2. Results of the ablation study on CE_{FS}.

Model	MAP@1k	R@100	R@10	R@1
CE_{FS} w/o [T] splitter and query tags	.252	.510	.405	.163
CE_{FS} w/o [S] splitter and question subject	.239	.498	.379	.154
CE_{FS} w/o [D] splitter and question description	.191	.461	.315	.119
CE_{FS}	**.270**	**.524**	**.428**	**.209**

Ranking Quality. Table 1 illustrates the effectiveness of lexical-based first-stage retrievers and cross-encoder re-rankers. For re-ranking, our proposed method, CE_{FS}, demonstrates significant improvements over both CE_{CAT} fine-tuned on the LegalQA dataset and CE_{CAT} trained on MS MARCO, referred to as MiniLM-MSMARCO [24]. E.g., in terms of MAP@1k, CE_{FS} achieves 0.270 vs. 0.236 for CE_{CAT} and 0.109 for MiniLM-MSMARCO. The higher effectiveness of CE_{FS} over the CE_{CAT} confirms the effectiveness of the proposed method, and the low performance of MiniLM-MSMARCO on LegalQA confirms the challenges of the legal domain since MiniLM-MSMARCO achieves a three times higher effectiveness on MS MARCO dataset in terms of MAP@1k. Among the first-stage retrievers, BM25 outperforms LMD in terms of initial ranking effectiveness. The relatively low effectiveness of BM25 reveals a noticeable difference in lexical word overlap between the question and the best answer. This difference serves as an indicator of a knowledge gap between the questioner and the lawyer, resulting in a higher occurrence of lexical mismatches between the question and the relevant answer. Achieving a higher overall effectiveness in this task is a potential area for improvement in future works.

Ablation Study. We do an ablation study on the CE_{FS} to analyze to what extent each section of the fine-grained structured input of CE_{FS} has an impact on the effectiveness of CE_{FS}. As shown in Table 2, the effectiveness of CE_{FS} is highest when we use all of the splitter and corresponding contents. We see that query tags and [T] splittor have less impact on the effectiveness and question description and [D] has the highest impact. Question subject and [S] have the second-highest impact. This suggests that although the query tags have a role in improved effectiveness, question subject and description have still a large impact on the effectiveness.

5 Conclusion

For investigating answer retrieval in legal community question answering, we created a dedicated dataset, called LegelQA, which we divide into training, validation, and test sets. We use this dataset to train neural retrieval models, introducing a novel and highly effective re-ranker called CE_{FS}, and take a fine-grained structured approach to leverage the information available in the legal domain, demonstrating significantly higher effectiveness over common strong cross-encoder re-rankers. We investigate the impact of each part of the fine-grained structured input within our method and highlight the significant role played by question tags in improving retrieval effectiveness besides the most important part of the question which is question description. While our method is initially proposed for answer retrieval within the legal domain, we foresee its potential application in answer retrieval for community question-answering systems across various domains where question tags are provided alongside each question post. For future work, our data, LegalQA, facilitates other tasks such as legal response generation.

Acknowledgments. This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval (H2020-EU.1.3.1., ID: 860721).

References

1. Abolghasemi, A., Verberne, S., Azzopardi, L.: Improving BERT-based query-by-document retrieval with multi-task optimization. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13186, pp. 3–12. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_1
2. Askari, A., Abolghasemi, A., Aliannejadi, M., Kanoulas, E., Verberne, S.: Closer: conversational legal longformer with expertise-aware passage response ranker for long contexts. In: The 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023). ACM (2023)
3. Askari, A., Abolghasemi, A., Pasi, G., Kraaij, W., Verberne, S.: Injecting the BM25 score as text improves BERT-based re-rankers. In: Advances in Information Retrieval, pp. 66–83. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_5
4. Askari, A., Verberne, S., Pasi, G.: Expert finding in legal community question answering. In: Hagen, M., et al. (eds.) Advances in Information Retrieval, pp. 22–30. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_3
5. Atkinson, J., Figueroa, A., Andrade, C.: Evolutionary optimization for ranking how-to questions based on user-generated contents. Expert Syst. Appl. **40**(17), 7060–7068 (2013)
6. Bian, J., Liu, Y., Agichtein, E., Zha, H.: Finding the right facts in the crowd: factoid question answering over social media. In: Proceedings of the 17th International Conference on World Wide Web, pp. 467–476 (2008)

7. Boualili, L., Moreno, J.G., Boughanem, M.: MarkedBERT: integrating traditional IR cues in pre-trained language models for passage retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1977–1980 (2020)
8. Budler, L.C., Gosak, L., Stiglic, G.: Review of artificial intelligence-based question-answering systems in healthcare. Wiley Interdisc. Rev.: Data Min. Knowl. Disc. **13**(2), e1487 (2023)
9. Chen, T., Zhang, M., Lu, J., Bendersky, M., Najork, M.: Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In: Hagen, M., et al. (eds.) ECIR 2022. LNCS, vol. 13185, pp. 95–110. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-99736-6_7
10. Haigh, R.: Legal English. Routledge (2018)
11. Han, J., Hong, T., Kim, B., Ko, Y., Seo, J.: Fine-grained post-training for improving retrieval-based dialogue systems. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1549–1558 (2021)
12. Hofstätter, S., Althammer, S., Schröder, M., Sertkan, M., Hanbury, A.: Improving efficient neural ranking models with cross-architecture knowledge distillation. arXiv preprint arXiv:2010.02666 (2020)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals. In: Soviet Physics-Doklady, vol. 10, pp. 707–710 (1966)
15. Mansouri, B., Campos, R.: FALQU: finding answers to legal questions. arXiv preprint arXiv:2304.05611 (2023)
16. Martinez-Gil, J.: A survey on legal question-answering systems. Comput. Sci. Rev. **48**, 100552 (2023)
17. Nguyen, T., et al.: Ms MARCO: a human generated machine reading comprehension dataset. In: CoCo@ NIPs (2016)
18. Nogueira, R., Cho, K.: Passage re-ranking with BERT. arXiv preprint arXiv:1901.04085 (2019)
19. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: ACM SIGIR Forum, vol. 51, pp. 202–208. ACM New York, NY, USA (2017)
20. Rau, D., Kamps, J.: The role of complex NLP in transformers for text ranking. In: Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 153–160 (2022)
21. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: SIGIR'94, pp. 232–241. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_24
22. Roy, P.K., Ahmad, Z., Singh, J.P., Alryalat, M.A.A., Rana, N.P., Dwivedi, Y.K.: Finding and ranking high-quality answers in community question answering sites. Glob. J. Flex. Syst. Manag. **19**, 53–68 (2018)
23. Roy, P.K., Saumya, S., Singh, J.P., Banerjee, S., Gutub, A.: Analysis of community question-answering issues via machine learning and deep learning: state-of-the-art review. CAAI Trans. Intell. Technol. **8**(1), 95–117 (2023)
24. Sentence-BERT: cross-encoder for ms MARCO: ms-marco-minilm-l-12-v2 (2023). https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2
25. Tiersma, P.M.: Legal language. University of Chicago Press (1999)
26. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv. Neural. Inf. Process. Syst. **33**, 5776–5788 (2020)

27. Williams, C.: Tradition and change in legal English: verbal constructions in prescriptive texts, vol. 20. Peter Lang (2007)
28. Wolf, T., et al.: Huggingface's transformers: state-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
29. Xiong, W., et al.: TWEETQA: a social media focused question answering dataset. arXiv preprint arXiv:1907.06292 (2019)
30. Yang, W., et al.: End-to-end open-domain question answering with BERTserini. arXiv preprint arXiv:1902.01718 (2019)
31. Yen, S.J., Wu, Y.C., Yang, J.C., Lee, Y.S., Lee, C.J., Liu, J.J.: A support vector machine-based context-ranking model for question answering. Inf. Sci. **224**, 77–87 (2013)
32. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in Neural Information Processing Systems, vol. 31 (2018)