KCC: Korean Civil Case Dataset for Legal Information Retrieval

Anonymous ACL submission

Abstract

Analyzing relevant or similar precedent cases is crucial in the field of law. This study presents a novel legal information retrieval dataset, KCC, for Korean civil judgments, consisting of 2,942 civil cases, treated by Korean courts in between March 3rd in 1947 and December 31st, 2022. In the proposed dataset, we introduce and annotate a 4-level case similarity criteria, which is verified by legal experts, resulting in both high-level legal reasoning as well as factual circumstances can be considered in the legal IR tasks. Experiments on the proposed dataset using popular legal IR methods demonstrate promising performance in legal IR tasks. We believe the proposed dataset can be used as valuable resource for developing legal IR models, which can assist legal professionals.

1 Introduction

In the field of law, analyzing precedents is crucial in understanding legal cases as well as reaching to legal decision. Hence, legal experts including lawyers and judges invest a lot of time in searching through extensive legal documents to find and comprehend existing cases that are similar to their own. To reduce expensive time and cost, NLP technologies have been actively applied to retrieve similar precedents with reference cases (Bench-Capon et al., 2012; Zhong et al., 2020). Such legal information retrieval (IR) technologies (Fraenkel, 1969; Maxwell and Schafer, 2008; van Opijnen and Santos, 2017; Moens, 2001) can help legal experts to easily find cases that support their arguments and get insights into factual elements that are significantly considered by the court, through comparisons with similar cases. Additionally, reviewing court judgments on similar cases using NLP technologies can be beneficial for individuals who are unfamiliar with the law, as it can help them understand their current legal situation.

To improve legal IR performance, ensuring information about the relevant cases is crucial. This is especially important in the statutory law systems. In the case law systems, there can be many cited cases that can be used as relevant information (Rabelo et al., 2022), but the statutory law systems usually lack such information[1]. For overcoming this problem, researchers have started to build legal case datasets annotated with relevance. For example, Chinese Legal Case Retrieval Dataset (Ma et al., 2021) consists of 107 query cases and 10,700 candidate cases selected from a corpus of over 43,000 Chinese criminal judgments; each candidate case in LeCaRD includes a case name, fact description, judgment, and relevance label.

Since the Republic of Korea is a nation characterized by a statutory legal system, the availability of publicly accessible legal cases is limited, and there exists no dataset annotated by case relevance in the field of legal IR. Hence, conducting a Legal IR research under these circumstances poses significant challenges. Note that utilizing existing legal datasets from other countries is challenging, since judgment systems are significantly different across the nations. Consequently, the development of a dedicated dataset tailored to Korean legal IR becomes an imperative necessity for the advancement of a Korean Legal IR system.

Therefore, this paper aims to provide a novel legal IR dataset, KCC, for Korean civil judgments, providing a valuable resource for future studies and developments for the Korean legal IR systems. We focus on civil cases, because these account for 70% of all the cases in the Republic of Korea in 2022 (Supreme Court Administration Office, 2023). To this end, we first collected a total of 38,372 korean civil cases, each of which includes detailed information such as case number, sentence date, judgment note, etc. We then created pairs of query

[1] https://en.m.wikipedia.org/wiki/Statutory$_{law}$

and its candidate cases, and annotated similarities for each pair. In this way, we introduce a 4-level similarity criteria, summarized in Table 1. We manually annotated the similarity for each pair, which is verified by legal experts.

Based on the proposed data, KCC, we perform a legal IR task using widely-used methods in legal IR. Experimental results show that utilizing LLMs through prompt engineering could potentially outperform other deep learning methods. We believe the performance reports of widely-used IR methods can be a reference for future research and developing new methods in legal IR. We expect our proposed dataset can be used as training and validation resource toward developing legal IR systems that can assist legal professionals by supporting legal decision making processes. The proposed dataset, KCC, is planned to be released if accepted.

2 KCC: Korean civil case dataset

2.1 Data Source

We first obtained 35,217 Korean civil cases, treated by Korean courts from March 3rd in 1947 to October 17th in 2017, from AIhub[2], a popular platform providing a wide range of datasets across domains. We further collected 3,155 civil cases directly from the Ministry of Government Legislation, the Republic of Korea, for the period between October 18th, 2017 and December 31st, 2022. Finally, our dataset includes a total of 38,372 civil cases.

Since the types of litigation were expressed using different terms in each precedent, we merged similar litigation types into a single group. For example, ‘Transfer of Ownership Registration’, ‘Cancellation of Transfer of Ownership Registration’, and ‘Cancellation of Preservation Registration of Ownership’ are all consolidated into the ‘Transfer of Ownership Registration’ group.

The detailed information of the collected data is summarized in Table 4 in Appendix. For each pair consisting of a query case and a candidate case, our dataset includes the precedent number, judgment note, judgment abstract, and judgment text for the both cases, and the similarity between two cases. Judgment note, judgment abstract, and judgment text were extracted from the original precedent texts. Judgment note handles the legal key issues of the case law, providing a concise overview of the legal aspects involved. The judgment abstract, on the other hand, is a distilled version of the judgment text, focusing primarily on the legal reasoning while excluding the factual details. The use of judgment abstract and note allows for future structured similarity evaluations, enhancing the utility of the proposed dataset in legal analysis.

Label	Criterion
label 3	The legal judgment of the query case was applied identically to the candidate case.
label 2	The legal judgment of the query case is also described in the candidate case, but it is not used as a key legal reasoning to reach a conclusion, or the conclusion of the case is different.
label 1	The query case and candidate case share the same keywords.
label 0	Both the keywords and the legal judgment of the query case and candidate case are different.

Table 1: Case similarity criteria proposed in KCC.

2.2 Case Similarity Criteria

For calculating the similarity between cases, prior studies on legal datasets mostly followed the guidance of official government documents or legal experts. Since there is no official Korean civil case similarity criteria, we introduce a Korean case similarity criteria under the guidance of legal experts in Korea. To determine the similarity between two precedents, both factual circumstances and legal judgments should be considered. Factual circumstances can be relatively easily analyzed by extracting keywords and establishing the relationship between the plaintiff and the defendant. However, legal judgments require a higher level of legal reasoning as they are determined by legal experts, and can be applied to cases with different factual circumstances. Note that the Supreme Court in the Republic of Korea focuses only on legal judgments without assessing factual circumstances; hence considering legal judgments should be prioritized in searching similar relevant precedents on a Supreme Court level in the Republic of Korea.

Table 1 summarizes the 4-level criteria of case similarity, which considers both factual circumstances and legal judgments, which is reviewed by legal experts in Korea. If the 4 level criteria is converted to a binary classification, labels 2 and 3 can be substituted as ‘similar’, and the labels 0 and 1 as ‘dissimilar’. To exemplify the proposed criteria, refer to an example in Table 5 in Appendix.

[2] https://www.aihub.or.kr

2.3 Query Selection and Candidate Pooling

2.3.1 Query Selection Strategy

Similar to the LeCaRD's query sampling strategy (Ma et al., 2021) where query cases are selected based on the frequency and distribution of charges, we first sorted the types of litigation based on their frequency. Among the more than 500 objectives in lawsuits, the top 20 lawsuit objectives in terms of frequency account for approximately 54% of all civil cases. See Table 6 in Appendix that summarizes the top 20 objectives in terms of frequency. Therefore, we select the 20 representative query case for each top 20 lawsuit objective based on the following condition. The query case should be (i) the main legal judgment should be only one, (ii) the legal judgment should be clear, and (iii) the legal judgment should be commonly used in actual civil cases. Legal professionals were involved in the query selection process.

2.3.2 Candidate Pooling

For each query case, candidate cases were selected based on (i) a Transformer-based pretrained language model (PLM) and (ii) factual keyword searching. Using a PLM[3], the contextual representation of legal cases representing factual circumstances and legal judgments can be captured. We then calculated the cosine similarity between the generated embeddings for the query and candidate cases, which is added to the candidate pool. Since legally meaningful cases can be found within the top 50 cases in terms of similarity, we limited the candidate pool to have at most 50 cases. We also applied keyword searching, which is one of the widely-used traditional IR methods, which can identify candidate cases where selected (similar) keywords are used. The keywords of the query case were chosen under the supervision of legal experts and limited to 3 to 4 words. Note that the keywords for query cases were carefully reviewed by legal experts to ensure their accuracy and factual relevance to the query case.

2.4 Annotation and Analysis

2.4.1 Annotation Process

The annotation of the similarity between a query and a candidate case was conducted by three paralegals, all of whom hold bachelor's degree in law and have completed over 40 credits in law-related subjects in a university. To verify the label results annotated by the three annotators, two legal experts, who hold J.D. and are qualified lawyers, also participated in the process.

The 20 query cases are divided into three groups, each of which has 7, 7, and 6 query cases, and each group is allocated to each annotator to perform annotations. To assess the agreement among annotators, and between annotators and experts, a random sample of 140 case pairs was selected, and annotated by all the annotators and experts. The Krippendorff's alpha score, which measures agreement and reliability between multiple annotations, was calculated. The 3 paralegals show an agreement score of 0.89, and 2 experts exhibit an agreement score of 0.93. The agreement score between the annotators and experts is 0.90, which suggests that the annotators tend to reach a good consensus, and demonstrate a strong level of agreement with legal experts. In addition to Krippendorff's alpha, we calculated the Cohen's Kappa to measure the inter-annotator agreement. The Cohen's Kappa scores, detailed in Table 7 in Appendix, underscore the strong agreement across our annotators, reflecting a high level of precision in the annotation process.

2.4.2 Data Analysis

The detailed annotation result is summarized in Table 6 in Appendix. The total number of query-candidate case pairs is 2,942. Among the entire dataset, 201 pairs (6.8%), 174 pairs(5.9%), 350 pairs (11.9%), and 2217 pairs (75.4%) are annotated as labels 3, 2, 1, and 0, respectively. If the 4-level categorization is converted into a binary classification (i.e., similar vs. dissimilar), the labels 3 and 2 can be considered as 'similar' and labels 1 and 0 can be considered as 'dissimilar'; in our dataset, there were 375 similar pairs and 2,567 dissimilar pairs. For each query, there is a significant label imbalance between similar and dissimilar pairs; only 6.8% of total pairs correspond to label 3 while 75.4% of total pairs correspond to label 0. Note that the label imbalance can be easily observed in the real-world legal scenarios.

3 Experiments

3.1 Legal IR Models

In the field of legal IR, retrieval models based on the keyword overlap between two documents, such as TF-IDF (Salton and Buckley, 1988) and BM25 (Robertson et al., 1995), have been widely used. The emergence of pretrained language

[3] https://huggingface.co/klue/bert-base

Types	Model	Precision@5	Precision@10
Traditional Model	BM25	0.35	0.375
Neural Networks	1D-CNN	0.25	0.25
	LSTM	0.2	0.175
	BERT-PLI	0.22	0.18
	BERT (CE)	0.43	0.485
	LCube (CE)	0.19	0.19
Prompt Engineering	Legal-CoT	0.50	0.55
	Legal-Syllogism	0.62	0.67

Table 2: Performance results on popular IR models (CE = cross-encoder).

Model	Acc.	F1 (0)	F1 (1)
finetuned BERT	0.76	0.48	0.75
Legal-CoT ZS	0.64	0.47	0.69
Legal-Syllogism ZS	0.80	0.42	0.88

Table 3: Performance results on prompting (0 = similar label; 1 = dissimilar label).

models has paved the way for developing legal IR models using deep learning. For example, BERT_LF (Hu et al., 2022), SAILER (Li et al., 2023), and IOT-MATCH (Yu et al., 2022b) utilized BERT-based pretrained models on Chinese Legal Corpus (Zhong et al., 2019) as a backbone. More recently, employing prompt engineering techniques on large language models (LLMs) has been actively explored in the legal domain.

In our experimental setup, we considered the following three representative approaches: (i) traditional IR model, (ii) neural network models, and (iii) prompt engineering techniques. For the traditional IR model, we selected BM25 (Robertson et al., 1995) that uses keywords matching and document length to retrieve the most relevant results. For the neural network models, we applied the widely-used models, including 1D-CNN, LSTM, and Transformer-based models. The Transformer-based model uses the finetuned BERT (Shao et al., 2020) and LCube, a T5 model pretrained on a Korean legal corpus (Hwang et al., 2022).

In prompting engineering techniques, we employed variants of Legal Syllogism Prompting (Jiang and Yang, 2023) and Legal Chain of Thought (Yu et al., 2023). Since these methods are not tailored for legal case retrieval, we adopted each method to summarize legal cases and then leveraged case summaries to assess similarity, instead of directly applying for the legal IR task.

3.2 Results

We first examine how the popular legal IR models can retrieve the most similar legal precedents when trained on the proposed dataset, under the 4-level similarity criteria. We next investigate the effectiveness of the prompting approach under a binary classification (i.e., similar vs. dissimilar) and the zero-shot setting. In particul ar, we seek to investigate the feasibility of utilizing existing LLMs as retrieval models with minimal modifications.

Table 2 outlines the performance of popular IR models, showing that traditional models like BM25 perform robustly, often surpassing neural network approaches. This echoes the findings from LeCARD's study (Ma et al., 2021), which also noted the strength of BM25 against more complex models. Remarkably, BERT with a cross-encoder configuration demonstrates superior performance, emphasizing the advantages of domain-specific pretraining. Nevertheless, the models consistently underperform on two particular query cases, suggesting a common area of difficulty and a potential focal point for further model refinement.

Looking at Table 3 regarding prompt engineering techniques, we unveil the potentiality of prompt engineering techniques on legal IR. Notably, the 'Legal Syllogism' approach demonstrates a higher performance compared to 'Legal CoT', showcasing its superior ability in leveraging context for legal case retrieval. In particular, when employing multilingual GPT-4 as the underlying Pretrained Language Model (PLM), the prompt engineering methods outshine fine-tuned PLM even in a zero-shot setting. This may be due to the extensive training data of LLMs compared to the models pretrained on more limited corpora.

4 Conclusion

This paper showcased the first expert-verified IR dataset, KCC, for Korean civil judgments, providing a valuable resource for future studies and developments in the legal domain. By introducing a 4-level case similarity criteria in the proposed dataset, verified by legal experts, both a high-level legal reasoning as well as factual circumstances can be considered in legal IR tasks. Experimental results showed that utilizing LLMs through prompt engineering could potentially outperform other deep learning methods. However, there is a potential for improving the performance in the future by utilizing other techniques such as few-shot learning.

Ethical Considerations

This research involves the analysis of data that originally contained personally identifiable information. To uphold the ethical standards of research and ensure compliance with privacy laws, all datasets used in this study have undergone a rigorous anonymization process conducted by the Supreme Court and the Ministry of Government Legislation of the Republic of Korea. This procedure included the removal of names and any information that could potentially identify individuals.

Limitations

This study employs a dataset encompassing 20 query cases, which do not encompass the full spectrum of civil cases in the Republic of Korea. However, they were carefully selected based on frequency, representing the most common types of civil disputes. Despite the limited number of query cases, we believe that our dataset adequately covers the majority of recurrent civil case types due to this methodical selection process.

Moreover, the dataset predominantly features rulings from the Korean Supreme Court. Given the court's tendency for succinct fact description, the level of detail in case narratives can be limited. This conciseness is not a choice but a reflection of the nature of the data released by the supreme Court and the Ministry of Government Legislation, where more than 70% of publicly available data are from the Supreme Court.

References

Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G. Conrad, Enrico Francesconi, Thomas F. Gordon, Guido Governatori, Jochen L. Leidner, David D. Lewis, Ronald P. Loui, L. Thorne Mccarty, Henry Prakken, Frank Schilder, Erich Schweighofer, Paul Thompson, Alex Tyrrell, Bart Verheij, Douglas N. Walton, and Adam Z. Wyner. 2012. A history of ai and law in 50 papers: 25 years of the international conference on ai and law. *Artif. Intell. Law*, 20(3):215–319.

Aviezri S. Fraenkel. 1969. Legal information retrieval. *Advances in Computers*, 9:113–178.

Weifeng Hu, Siwen Zhao, Qiang Zhao, Hao Sun, Xifeng Hu, Rundong Guo, Yujun Li, Yan Cui, and Long Ma. 2022. Bert_lf: A similar case retrieval method based on legal facts. *Wireless Communications and Mobile Computing*, 2022.

W. Hwang, D. Lee, K. Cho, H. Lee, and M. Seo. 2022. A multi-task benchmark for korean legal language understanding and judgment prediction. *arXiv preprint arXiv:2206.05224*.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 417–421.

H. Li, Q. Ai, J. Chen, Q. Dong, Y. Wu, Y. Liu, ..., and Q. Tian. 2023. Sailer: Structure-aware pretrained language model for legal case retrieval. *arXiv preprint arXiv:2304.11370*.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: A legal case retrieval dataset for chinese law system. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval(Canada) (SIGIR '21)*, pages 2342–234.

K. Tamsin Maxwell and Burkhard Schafer. 2008. Concept and context in legal information retrieval. In *Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 63–72, NLD. IOS Press.

MF Moens. 2001. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9:29–57.

J. Rabelo, R. Goebel, and MY. et al. Kim. 2022. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *Rev Socionetwork Strat*, 16:111–133.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In *IJCAI*, pages 3501–3507.

Supreme Court Administration Office. 2023. Judicial yearbook.

Marc van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artif Intell Law*, 25:65–87.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022a. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596.

Weijie Yu, Zhongxiang Sun, Jun Xu, Zhenhua Dong, Xu Chen, Hongteng Xu, and Ji-Rong Wen. 2022b. Explainable legal case matching via inverse optimal transport-based rationale extraction. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 657–668.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does nlp benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230.

Haoxi Zhong, Zhengyan Zhang, Zhiyuan Liu, and Maosong Sun. 2019. Open chinese language pretrained model zoo. *Technical report*.

A Appendix

A.1 Experiment Settings

The training of all models was conducted using precedent notes, and for the non-PLM models, nouns and embeddings of nouns were used while KoBERT [4] was utilized as a backbone PLM. The implementation of the BERT-PLI model followed the fine-tuning process described in (Shao et al., 2020).

In prompt engineering methods, we followed a similar approach to the one described in (Yu et al., 2022a). While the OpenAI 'text-davinci-003' model was used in (Yu et al., 2022a), we used OpenAI's 'gpt-4' model in our study. After approaching with various prompts, we finally used the following prompt: {prompt: "Evaluate the similarity or dissimilarity of the legal judgments in claim 1 and claim 2 by analyzing factual circumstances, legal provisions, legal judgments, and decisions in a step-by-step legal reasoning process."} + {query precedent note} + {candidate precedent note} + {"Answer by 'similar' or 'dissimilar'."}. For Legal Syllogism, we used following prompt instead: {prompt: "Evaluate the similarity or dissimilarity of the legal judgments in claim 1 and claim 2 by Major Hypothesis(Legal Circumstance), Minor Hypothesis(Factual Description) and Conclusion."} + {query precedent note} + {candidate precedent note} + {"Answer by 'similar' or 'dissimilar'."}. For reproducibility, the temperature was set to 0.

For evaluation metrics, $precision@5$ and $precision@10$ were used. Following the evaluation method used in (Ma et al., 2021), candidate cases were sorted in a descending order based on $\delta = pred_1 - pred_0$, where $pred_1$ represents the logit of label 3 (most relevant) while $pred_0$ represents the sum of the logits for the other labels.

[4] https://huggingface.co/klue/bert-base

Key	**Description**
precedentNumber	Identification number assigned by the Ministry of Government Legislation
caseName	Objective of the lawsuit
caseNumber	Identification number assigned by the Ministry of Justice
sentenceDate	Date of the court ruling
judgmentAbstract	Abstract of the precedent
judgmentNote	Key issues in the precedent
precedentText	Whole text of precedent
label	Similarity label between query and candidate case

Table 4: Definition of each key in the proposed dataset.

Case	Legal judgment	Similarity & Reason
Query	In the case of a so-called contract for work, such as when the contractor directs a specific action or contracts out a specific business, even if he/she is a contractor, the contractor is liable for accidents caused by the negligence of the contractee under the provisions of the employer's liability under Article 756 of the Civil Act.	-
Candidate 1	To impose the employer's reliability for an action of contractee on a contractor, there must be a relationship of direction and supervision between them.	Similarity: 2 Reason: The legal principle applied in both cases are the same, The candidate case reached a different conclusion compared to the query case because there was no direction and supervision relationship between the parties.
Candidate 2	If a company contracted for the construction of a new building and subcontracted part of it to a third party, and dispatched its employees to the above construction site to direct and supervise the construction, the company is liable for compensation as an employer under Article 756 of the Civil Act for damages caused by a third party employee to another person.	Similarity: 3 Reason: : If the contractor directed and supervised the work, he/she are liable as an employer for damages caused by an contractee's employee.
Candidate 3	If persons in a partnership delegate the execution of a task that should be jointly handled to one of the partners and have him/her handle it, enabling them to carry out the task, the other partner shall be deemed to be in the position of an employer as well as a partner of the executor, and shall be liable for damages as an employer for accidents occurring in the course of the execution of the task.	Similarity: 3 Reason: The relationship between the parties in the two cases was different(query case: contract for work, candidate case: contract for partnership), but regardless of the type of relationship, it was determined that if one person directed or supervised the other person to perform a task, the person assuming the role of director or supervisor would be held liable as the employer.

Table 5: An example of a query case and its candidate cases, translated from Korean to English.

Case	**Objective of lawsuit**	**label 3**	**label 2**	**label 1**	**label 0**	**Total**
154548	Transfer of Owner Registration	8	2	8	269	287
106730	Reimbursement Claim	19	7	8	16	50
75735	Claim of Damage Compensation	36	20	19	24	99
170963	Unjust Enrichment Restitution	5	7	5	33	50
99762	Rental Fee	21	8	24	47	100
118470	Insurance Payment	13	19	18	100	150
136501	Dividend Dispute	8	8	6	128	150
82517	Building Surrender	15	1	22	388	426
81918	Verification of Debt Existence	7	15	7	21	50
155133	Building Demolition	1	1	5	112	119
76930	Invalidity Confirmation of Termination	5	5	14	76	100
145960	Rescission of Unauthorized Acts	21	11	55	13	100
106724	Promissory Note Payment	1	21	9	191	222
151153	Delivery of Land	4	8	9	213	234
124097	Salary/Wage	10	10	19	9	48
108680	Ownership Verification	10	14	12	14	50
79060	Transfer Payment	4	6	11	444	465
110940	Goods Payment	1	1	26	53	81
152477	Transfer Order of Monetary Claim	6	5	31	26	68
83349	Guarantee Obligation Payment	6	5	42	40	93
	Sum	201	174	350	2217	2942

Table 6: The annotation results for the 20 query cases.

	Exp1	**Exp2**	**Para1**	**Para2**	**Para3**
Exp1	1.0	0.84	0.71	0.76	0.66
Exp2		1.0	0.70	0.75	0.72
Para1			1.0	0.76	0.72
Para2				1.0	0.68
Para3					1.0

Table 7: Inter-annotator agreement scores by Kohen’s Kappa (**Exp**: experts, **Para**: paralegals).