Overview of Japanese Statute Law Retrieval and Entailment Task at COLIEE-2018

Masaharu Yoshioka[1,2], Yoshinobu Kano[3], Naoki Kiyota[3], and Ken Satoh[4]

[1] Graduate School of Information Science and Technology, Hokkaido University, N14 W9, Kita-ku, Sapporo-shi, Hokkaido, Japan
yoshioka@ist.hokudai.ac.jp
[2] Global Station for Big Date and Cybersecurity, Global Institution for Collaborative Research and Education, Hokkaido University, Kita-ku, Sapporo-shi, Hokkaido, Japan
[3] Faculty of Informatics, Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu-shi, Shizuoka 432-8011 Japan
kano@inf.shizuoka.ac.jp, nkiyota@kanolab.net
[4] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
ksatoh@nii.ac.jp

Abstract. We present an overview of two tasks (Tasks 3 and 4) of COLIEE-2018 using the Japanese statute law (civil law) and its English translation. Task 3 is the task of retrieving articles to decide the appropriateness of the legal question and Task 4 is the task of entailing whether the legal question is correct or not. There are 17 run submissions from eight teams (including two run submissions from the organizers) for Task 3 and seven run submissions from three teams for Task 4. We present a summary of the evaluation results of both tasks.

Keywords: Legal Information Retrieval · Legal Information Entailment · Query Analysis · Natural Language Processing.

1 Introduction

Competition on legal information extraction/entailment (COLIEE) is a series of evaluation campaigns to discuss the state of the art for information retrieval and entailment using legal texts [7, 6, 4]. In the previous COLIEE (COLIEE 2015-2017), there were two tasks (information retrieval (IR) and entailment) of using the Japanese statute law (civil law). In COLIEE-2018, we conduct two new tasks (IR and entailment) of using a Canadian case law and two tasks of using the Japanese statute law that had the same settings as in the previous campaigns.

For the IR task (Task 3), based on the discussion about the analysis of IR tasks of the previous COLIEE [12], we modified the evaluation measure of the final results and also asked the participants to submit ranked relevant article results to discuss the detailed difficulty of the questions.

For the entailment task (Task 4), we performed categorized analyses to show different issues of the problems and characteristics of the submissions, in addition to the accuracy evaluation, similar to tasks of the previous COLIEE.

This paper summarizes an overview of Tasks 3 and 4 of COLIEE-2018. There were 17 run submissions from eight teams (including two run submissions from the organizers) for Task 3 and seven run submissions from three teams for Task 4. We present a summary of the evaluation results of the two tasks. Sections 2 and 3 discuss the evaluation for Tasks 3 and 4, respectively. Section 4 summarizes the paper with a discussion about the future direction of these tasks based on the evaluation results.

2 Task 3: Japanese Statute Law Retrieval Task

2.1 Task description

Task 3 is the task of retrieving articles to decide on the appropriateness of the legal question. The training and test data of the legal questions were collected from the Japanese bar exam. All the questions and the Japanese civil law articles (1056 articles in total) were provided in two languages, Japanese and English. The English version of the law articles and questions was provided by the organizers. The participants were asked to submit relevant articles for the questions using Japanese or English data. Each participant can submit at most three runs for Task 3.

2.2 Data set

In this evaluation campaign, for this task, questions related to Japanese civil law were selected from the Japanese bar exam. The organizers provided a data set used for previous campaigns [7, 6, 4] as training data (651 questions) and new questions selected from the 2017 bar exam as test data (69 questions).

As the system mostly returns only one article for each question, the number of relevant article(s) for the question affects the system performance. The number of questions classified by the number of relevant articles is listed in Table 1.

Table 1. Number of questions classified by number of relevant articles

number of relevant article(s)	1	2	3	total
number of questions	51	16	2	69

2.3 Submitted runs

The following eight teams (in alphabetical order except for the organizers' team as baseline) submitted their results (17 runs in total). Three teams (HUKB, JNLP, and UA) had experience in submitting results in the previous campaign and four teams (Smartlaw, SPABS, UB, and UE) were new to the campaign.

HUKB (two runs) [13] used structural analysis results (condition, decision) of the article and questions and used Indri [10] to calculate similarity measures among different parts. The SVM-rank [2] software was used to aggregate such similarity measures. HUKB1 decided on the number of returned articles based on the analysis of retrieval difficulty for IR. HUKB2 returned only one article for each question.

JNLP (two runs) [11] used the structural analysis results (requisite and effectuation) of articles and the TF-IDF-based vector space model for calculating the similarity among them. JNLP1 used the similarity between query and articles only for article ranking. JNLP2 calculated the final similarity value as a linear combination of similarity used for JNLP1 and the similarity between the query and the article effectuation part. Both runs returned two articles for all questions based on the analysis of training data.

Smartlaw (three runs) [1] calculated the similarity of a question and an article by checking the similarity between (1–4) gram sets extracted from the question and the article. Based on the experimental analysis, they submitted three runs whose settings for constructing (1–4) gram sets were different; Smartlaw, Smartlaw_2gram, and Smartlaw_3gram used bigram+trigram, and bigram and trigram, respectively.

SPABS (three runs) used recurrent neural network (RNN) models to calculate the similarity between question and articles. For training word embedding, they used English legal documents with Word2Vec. SPABS_bm25 is their baseline result using BM25.

UA (one run) [5] used the same system of COLIEE-2017 for Task 3. This system uses the TF-IDF model of Lucene. [5]

UB (three runs) used the Terrier 4.2 [6] with the PL2 term-weighting model as the IR platform. UB3 used TagCrowd [7] to select important keywords from each question and used them as a query of the IR platform. UB2 used query expansion after UB3 retrieval, and UB1 used word embeddings.

UE (one run) used the rule-based method to retrieve relevant documents.

ORG (two runs) used Indri [10] with simple settings (the question was used as query and all articles with title were indexed as a document) [12].

The teams who participated in the previous COLIEE proposed an extension or equivalent system for Task 3, and new teams proposed methods that were different from previous ones.

2.4 Evaluation of the Submitted Runs

Table 2 shows the evaluation results of submitted runs including the organizers runs. The official evaluation measures used in this task were F2 measure, precision (Prec.), and recall (Rec.). The terms "ret.", and "rel" represent the number

[5] https://lucene.apache.org/
[6] http://terrier.org/
[7] https://tagcrowd.com/

of returned articles and the number of returned relevant articles, respectively. The columns after the mean average precision (MAP) are explained below.

$$precision = \frac{number\ of\ retrieved\ relevant\ articles}{number\ of\ returned\ articles} \tag{1}$$

$$recall = \frac{number\ of\ retrieved\ relevant\ articles}{number\ of\ relevant\ articles} \tag{2}$$

$$f2 = \frac{5 \times precision \times recall}{4 \times precision + recall} \tag{3}$$

There are two differences in the evaluation measure used for the task compared with the former campaigns.

1. F2 measure was used instead of F measure (harmonic means of precision and recall).
   F2 measure is a variation of F-measure that weights recall higher than precision. If we assume that the IR task is a preprocess to provide relevant article(s) to the entailment system, a set of candidate article(s) including relevant article(s) to the entailment system needs to be provided. [8]
2. Macro average is used instead of micro average.
   In the former campaigns, the micro average (the average of evaluation measures is calculated based on the aggregated numbers of relevant articles, returned articles, and returned relevant articles for all questions) was used for evaluation. However, this measure is not very appropriate for different numbers of relevant articles. For example, for analyzing a recall, questions with multiple relevant articles are more important than one with one relevant article. In addition, when the system returns many articles for one query because of the uncertainty of the returned results, this seriously deteriorates the precision of the micro average. However, using the macro average (each evaluation measure is calculated based on the number of relevant articles, returned articles, and returned relevant articles for each question; after calculating the evaluation measure for each question, the average of such a measure over all questions is calculated), we can reduce the effect of such different characteristics among all retrieved results.

In the previous campaigns, because most of the teams submitted only one or two articles for each question, we could only evaluate the topic difficulties based on the number of systems that can return such articles as the relevant one. However, it is almost impossible to estimate the reason for the problem. For example, some questions have difficulties ranking the relevant articles higher because of vocabulary mismatch, and some questions have difficulties selecting the appropriate one from similar articles (relevant articles are ranked higher but not first-ranked). Therefore, we decided to ask the participants to submit a long ranking list (100 articles) in addition to the selected relevant article candidate list.

[8] This assumption is based on the concept that an entailment system can identify the most relevant part of texts from the provided texts.

This list provides information that could discuss the type of difficulties in retrieving relevant articles. For the long list, the MAP, recall at k (R_k: recall calculated by using the top k ranked documents as returned documents) are used for evaluation measure.

Table 2 shows information about the evaluation measure for the long rank list. However, because UE did not submit this long list, values are given as "-".

Table 2. Evaluation results of submitted runs (Task 3) and the corresponding organizers' run

runid	lang	ret.	rel.	F2	Prec.	Rec.	MAP	R_5	R_{10}	R_{30}
UB3	E	69	54	**0.6964**	**0.7826**	0.6860	**0.7988**	0.7978	0.8539	**0.9551**
UA	E	69	50	0.6602	0.7246	0.6522	0.7451	0.7303	0.7528	0.8539
ORGE1	E	69	49	0.6368	0.7101	0.6280	0.7381	0.7528	0.8090	0.8989
UB2	E	69	47	0.6232	0.6812	0.6159	0.7542	**0.7978**	**0.8652**	**0.9551**
JNLP1	E	138	57	0.6118	0.4130	**0.7126**	0.7398	0.7640	0.8202	0.9213
Smartlaw	E	138	57	0.6042	0.4130	0.7005	0.7036	0.7079	0.7640	0.8315
JNLP2	E	138	56	0.5997	0.4058	0.6981	0.7296	0.7528	0.8090	0.9101
SPABS_bm25	E	138	55	0.5821	0.3986	0.6739	0.7070	0.7753	0.8202	0.9101
UE	E	69	34	0.4516	0.4928	0.4469	-	-	-	-
Smartlaw_3gram	E	69	34	0.4387	0.4928	0.4324	0.4700	0.4494	0.4607	0.5056
UB1	E	69	31	0.4171	0.4493	0.4130	0.5355	0.5730	0.7191	0.8202
Smartlaw_2gram	E	141	34	0.3421	0.3023	0.4275	0.4594	0.4382	0.4831	0.5169
SPABS_rnnen	E	138	19	0.2150	0.1377	0.2536	0.2638	0.3371	0.4494	0.5730
SPABS_rnnsq	E	138	17	0.1957	0.1232	0.2319	0.2662	0.3483	0.4494	0.6067
HUKB2	J	69	53	0.6859	0.7681	0.6763	0.7805	0.7865	0.8427	0.9326
HUKB1	J	74	53	0.6826	0.7536	0.6763	0.7805	0.7865	0.8427	0.9326
ORGJ1	J	69	51	0.6633	0.7391	0.6546	0.7703	0.7753	0.8427	0.9326

Based on the comparison between ORGJ1 and ORGE1, we confirmed that there was no large difference between the English and the Japanese data.

As the average of the relevant articles per query was 1.29 (89/69), the performance of the systems that returned two articles for each question was worse than those that returned one article only. The best-performing system was UB3, which used a tag cloud algorithm to select appropriate keywords to construct the query and the Terrier IR platform to retrieve the final results. Teams that participated in the previous campaigns had almost similar scores except for JNLP, which returned two articles for each question. The performance of new teams, except for UB, was worse than the baseline system.

We discuss the difficulty of the questions based on the averaged evaluation measure among team's top run results for each language (eight results; HUKB2, JNLP1, SPABS_bm25, UB3, UA, Smartlaw, ORGJ, and ORGE). For the questions that have one relevant article, 28 out of 51 questions had an average MAP of 1.0. This means that these questions were easy and no system made the mistake of ranking relevant articles as the first article. For these questions, the

system that returned two articles for each question had a poor precision score (0.5) even though the systems ranked the relevant article as first-ranked article. We will not discuss here the easy questions in detail, and instead only focus on the difficult questions.

Figure 1 shows averages of MAP, R_5, R_{10} for the 23 questions with a single relevant article. In most cases, all of the system could find the relevant articles as higher ranked articles (14 questions had $R_5 = 1$ and two questions had $R_5 = 0.875$, which means that only one system cannot rank the articles in the top five.) There were a few questions for which it was difficult to rank relevant articles higher. The most difficult problem for which nobody managed to rank relevant articles in the top 10 is H29-1-0.

> H29-1-O: A is a nineteen-year-old male and subject to the parental authority of his parents. In the case where A concluded a contract not in writing to the effect that A would make B a gift of 1,000,000 yen from the property which A had obtained by inheritance, even if A revoked the gift on the ground that it was a gift not in writing, if A did not obtain the consent of the person who had parental authority in relation to him with respect to the revocation of the gift, A may rescind the revocation of the gift.

The relevant article (Article 5) uses the keywords "statutory agent" and "parental authority" may serve as "statutory agent." However, the system tends to retrieve articles that discuss "parental authority."

For these questions, the RNN-based approach SPABS_rnnsq selected this article as a third relevant article. Although the overall performance of the RNN-based system is not very effective, this type of system may be good at finding relevant articles with a large vocabulary mismatch.

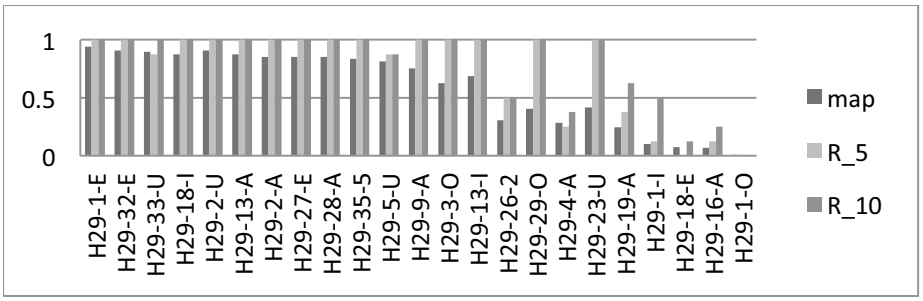

Fig. 1. Averages of MAP, R_5, R_{10} for difficult questions with a single relevant article

Figure 2 shows the averages of precision, recall, MAP, R_5, R_{10} for questions with multiple relevant articles (two questions, H29-28-E and H29-35-I, have three relevant articles and 16 other questions have two relevant articles). There are a few questions where both the first- and second-ranked articles are relevant

(MAP = 1). In other cases, there are many questions for which the content is similar to one of the relevant articles. but the other relevant articles are not so similar to the questions. H29-9-U is one such example of this type of question.

> H29-9-U: If the principal movable owned by A and the accessory movable owned by B are so joined to each other that they can no longer be separated without damaging the same, ownership of the composite Thing shall vest in A and B may demand compensation against A.

Relevant articles for this question are Articles 243 and 248. Article 243 shares the phrase "are so joined to each other that they can no longer be separated without damaging the same, ownership of the composite Thing shall vest in" and it is easy to rank it first. In contrast, article 248 only contains the phrase "may demand compensation." As a result, the content-based similarity is comparatively lower than for the first-ranked article and other articles similar to 243. One of the clues that can be used to identify that Article 248 is relevant is that it has a reference part to Article 243.

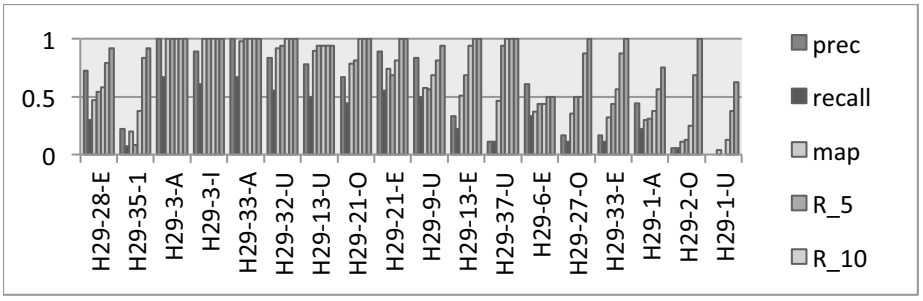

Fig. 2. Averages of precision, recall, MAP, R_5, R_10 for noneasy questions with a single relevant article

2.5 Discussion

As we have conducted a series of campaigns to retrieve relevant articles to entail the questions of the Japanese bar exam, most of the systems succeed in retrieving relevant articles of the simple questions that have only one relevant article and higher vocabulary (phrase) overlap between the question and the relevant article. However, the retrieval performance of the questions with vocabulary mismatch is insufficient. A semantic matching technique including the RNN approach may be one way to tackle this type of problem. However, to avoid the side effect of degrading the retrieval performance of easy questions, it is better to have a preprocess to select whether it is necessary to use such a semantic matching technique.

For the questions with multiple answers, there are many questions for which content-based similarity is inadequate in identifying the second or third supplemental relevant articles. Information about the relationships between articles may be a candidate information resource that is currently underused; further discussion is necessary to tackle this type of problem.

3 Task4: Entailment Task

3.1 Task description

Task 4 is a task to determine entailment relationships between a given problem sentence and an article sentence. Participants should answer yes or no regarding the given problem sentence. Pure entailment tasks were held until COLIEE-2016, where t1 (relevant article sentences) and t2 (problem sentences) were given. However, because of the limited number of available problems, COLIEE-2017 and -2018 did not hold this type of task. In Task 4 of COLIEE-2018, t1 (the relevant articles) was not given, and participants should find the relevant articles by themselves. Training and test data of the legal questions were collected from the Japanese bar exam, as for Task 3. All questions and Japanese civil law articles (1044 articles in total) were provided in two languages, Japanese and English. The English version of the law articles and questions was provided by the organizers. The participants were asked to submit yes/no answers for each question using Japanese or English data.

3.2 Data set

Our data set was derived from the civil law short answer (multiple choice) part of the Japanese legal bar exam. The organizers provided a data set of 644 questions used in the previous COLIEE tasks [7, 6, 4] as training data, and 69 new questions were selected from the 2017 legal bar exam as test data.

3.3 Submitted runs

The following three teams submitted their results. As one team submitted five runs, there were seven runs in total. Two teams (KIS and UA) had experience in submitting results in previous tasks and one team (UE) was new to the task.

KIS (5 runs) [8] analyzed Japanese sentences linguistically, and used predicate argument structures to determine similarities. [9] used frame information to calculate similarity between predicates. Their final results were an ensemble of these different modules by SVM.

UA (one run) [3] used almost the same system as in COLIEE-2017 for Task 4. Their system used condition/conclusion/exception detection rules, and negation dictionaries are created manually.

UE (one run) combined deep neural network with additional features, and word2vec to retrieve the corresponding civil law articles.

3.4 Evaluation of the submitted runs

Table 3 shows the evaluation results of submitted runs. The official evaluation measure used in this task was accuracy. LANG shows the language of the data, J for Japanese and E for English. Correct answers are the numbers of correct system outputs among 69 questions. The baseline system was simply No answers to all questions.

Table 3. Evaluation results of submitted runs (Task 4) and baseline result

Team	LANG	Correct Answers	Accuracy
BaseLine	N/A	35 (All No)	0.5072
UA	?	44	0.6377
KIS_Frame	J	39	0.5652
KIS_mo3	J	38	0.5507
KIS_dict	J	37	0.5362
KIS_SVM	J	36	0.5217
KIS_Frame2	J	35	0.5072
UE	E	33	0.4783

The best system was UA, with an accuracy of 0.6377. The baseline was almost 0.5, because this task was a binary classification, with 35/69 questions being No.

The effect of language differences was unclear. In our statute law tasks, the Japanese legal bar exam was the original data, which was manually translated into English. Team UA used the translation system and the Korean parser internally. The translation process might have absorbed ambiguities and paraphrases.

Because an entailment task is essentially a complex composition of different subtasks, we manually categorized our test data into categories, depending on what sort of technical issues had to be resolved. Table 4 shows our categorization results. As this was a compositional task, overlap between categories was allowed. Our categorization was based on the original Japanese version of the legal bar exam.

3.5 Discussion

Our categorization shown in the previous section suggested several issues and analyses. The largest number among these categories was for the conditions. UA, the best team, was better in this condition category. Their condition detection should have performed successfully. KIS_Frame2, which used the frame information, was good in case roles, person relations, and person roles. Their frame relation would have a certain effect in these deep semantic issues.

Because the distribution of Yes/No answers is quite diverse between submissions, an ensemble could perform better results if we could capture meaningful information for each submission.

Table 4. Technical category statistics of questions, and correct answers of submitted runs for each category. # stands for number of corresponding questions, team names stand for their number of correct answers for the corresponding category, acc stands for accuracy of its left-hand-side correct answers*.

Category	#	UA	acc	UE	acc	KIS1	acc	KIS2	acc	KIS3	acc	KIS4	acc	KIS5	acc
Itemized	3	1	0.33	2	0.67	1	0.33	1	0.33	1	0.33	1	0.33	2	0.67
Numerical priority	3	2	0.67	2	0.67	1	0.33	1	0.33	2	0.67	1	0.33	2	0.67
Entailment	5	2	0.40	2	0.40	1	0.20	1	0.20	4	0.80	2	0.40	2	0.40
Dependency	5	3	0.60	1	0.20	2	0.40	2	0.40	3	0.60	0	0.00	4	0.80
Article search	5	3	0.60	2	0.40	3	0.60	3	0.60	1	0.20	1	0.20	4	0.80
Paraphrase	5	2	0.40	4	0.80	3	0.60	3	0.60	2	0.40	3	0.60	3	0.60
Negation	7	5	0.71	3	0.43	5	0.71	5	0.71	2	0.29	1	0.14	7	1.00
Legal terms	7	4	0.57	2	0.29	2	0.29	2	0.29	3	0.43	4	0.57	3	0.43
Normal terms	9	5	0.56	5	0.56	4	0.44	4	0.44	5	0.56	6	0.67	4	0.44
Predicate argument	9	8	0.89	3	0.33	5	0.56	5	0.56	5	0.56	4	0.44	5	0.56
Verb paraphrase	13	7	0.54	6	0.46	7	0.54	7	0.54	7	0.54	7	0.54	4	0.31
Case role	15	8	0.53	6	0.40	9	0.60	9	0.60	6	0.40	11	0.73	6	0.40
Ambiguity	17	9	0.53	7	0.41	8	0.47	8	0.47	8	0.47	10	0.59	9	0.53
Anaphora	20	13	0.65	5	0.25	12	0.60	11	0.55	8	0.40	8	0.40	13	0.65
Morpheme	25	18	0.72	16	0.64	20	0.80	19	0.76	10	0.40	16	0.64	16	0.64
Person relationship	26	14	0.54	11	0.42	13	0.50	13	0.50	13	0.50	18	0.69	10	0.38
Person role	27	16	0.59	12	0.44	14	0.52	14	0.52	14	0.52	18	0.67	13	0.48
Conditions	31	19	0.61	9	0.29	13	0.42	12	0.39	16	0.52	11	0.35	16	0.52

* KIS1, KIS2, KIS3, KIS4, and KIS5 corresponds to KIS_mo3, KIS_dict, KIS_SVM, KIS_Frame2, and KIS_Frame respectively.

We did not have any end-to-end machine learning system this year, but we had many such systems previously. It would still be difficult to solve these entailment-based question-answering problems. The number of available questions was still too small to perform an effectively supervised machine learning. Evaluations of the training data (past years' problems) showed quite different scores between the different years. As our category analysis shows, many different issues remain to solve these questions; each category would require hundreds or thousands of training data sets at least. Furthermore, such deep semantic issues are still unresolved in general. A better linguistic analyzer would be needed as a base, not just end-to-end systems. In other words, our problems are very good material for challenging and evaluating deep semantics that are still unresolved.

4 Summary

This paper summarizes an overview of Tasks 3 and 4 of COLIEE-2018. For Task 3, we found that there were three types of problem in the test data; i.e., easy questions, difficult questions with vocabulary mismatch, and questions with multiple answers. Most submission systems are good at retrieving relevant answers for easy questions, but it is still difficult to retrieve relevant articles with other question types. It may be necessary to focus on such question types to improve the overall performance of the IR system. For Task 4, the overall performance of the submissions was still not sufficient for their systems to be used in the real application. However, a detailed analysis could capture the characteristics of the submitted systems. We found that to discuss and develop deep semantic analysis issues in the real application, and natural language processing in general is still a challenging task.

Acknowledgement

We would like to thank other COLIEE organizers for their supports to construct training data sets and organize this campaign. This work was partially supported by JSPS KAKENHI Grant Number 16H01756, 17H06103, 18H0333808, and JST CREST.

References

1. Chen, Y., Zhou, Y., Lu, Z., Sun, H., Yang, W.: Legal information retrieval by association rules. In: International Workshop on Juris-informatics (JURISIN 2018) (2018)
2. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 133–142. KDD '02, ACM, New York, NY, USA (2002). https://doi.org/10.1145/775047.775067, http://doi.acm.org/10.1145/775047.775067

3. Juliano Rabelo, Mi-Young Kim, H.B., Goebel, R.: Legal information extraction and entailment for statute law and case law. In: International Workshop on Juris-informatics (JURISIN 2018) (2018)
4. Kano, Y., Kim, M.Y., Goebel, R., Satoh, K.: Overview of coliee 2017. In: Satoh, K., Kim, M.Y., Kano, Y., Goebel, R., Oliveira, T. (eds.) COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment. EPiC Series in Computing, vol. 47, pp. 1–8. EasyChair (2017). https://doi.org/10.29007/fm8f, https://easychair.org/publications/paper/Fglr
5. Kim, M., Goebel, R.: Two-step cascaded textual entailment for legal bar exam question answering. In: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017. pp. 283–290 (2017). https://doi.org/10.1145/3086512.3086550, http://doi.acm.org/10.1145/3086512.3086550
6. Kim, M.Y., Goebel, R., Kano, Y., Satoh, K.: Coliee-2016: evaluation of the competition on legal information extraction and entailment. In: International Workshop on Juris-informatics (JURISIN 2016) (2016)
7. Kim, M.Y., Goebel, R., Satoh, K.: Coliee-2015: evaluation of legal question answering. In: Ninth International Workshop on Juris-informatics (JURISIN 2015) (2015)
8. Reina Hoshino, Ryosuke Taniguchi, N.K., Kano, Y.: Question answering system for legal bar examination using predicate argument structure. In: International Workshop on Juris-informatics (JURISIN 2018) (2018)
9. Ryosuke Taniguchi, R.H., Kano, Y.: Legal question answering system using framenet. In: International Workshop on Juris-informatics (JURISIN 2018) (2018)
10. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis. pp. 2–6 (2005)
11. Vu Tran, S.T.N., Nguyen, M.L.: Jnlp group: Legal information retrieval with summary and logical structure analysis. In: International Workshop on Juris-informatics (JURISIN 2018) (2018)
12. Yoshioka, M.: Analysis of coliee information retrieval task data. In: Arai, S., Kojima, K., Mineshima, K., Bekki, D., Satoh, K., Ohta, Y. (eds.) New Frontiers in Artificial Intelligence. pp. 5–19. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-93794-6_1
13. Yoshioka, M., Song, Z.: Hukb at coliee2018 information retrieval task. In: International Workshop on Juris-informatics (JURISIN 2018) (2018)