EPiC Series in Computing

Volume 47, 2017, Pages 1–8

COLIEE 2017. Fourth Competition on Legal Information Extraction/Entailment

Overview of COLIEE 2017

Yoshinobu Kano[1], Mi-Young Kim[2], Randy Goebel[2], and Ken Satoh[3]

[1] Faculty of Informatics, Shizuoka University, Japan

[2] University of Alberta, Canada

[3] National Institute of Informatics, Japan

kano@inf.shizuoka.ac.jp, {miyoung2, rgoebel}@ualberta.ca, ksatoh@nii.ac.jp

Abstract

We present the evaluation of the legal question answering Competition on Legal Information Extraction/Entailment (COLIEE) 2017. The COLIEE 2017 Task consists of two sub-Tasks: legal information retrieval (Task 1), and recognizing entailment between articles and queries (Task 2). Participation was open to any group based on any approach, and the tasks attracted 10 teams. We received 9 submissions to Task 1 (for a total of 17 runs), and 8 submissions to Task 2 (for a total of 20 runs).

1 Introduction

During the last three years, we held the first, second and third competitions on legal information extraction/entailment, COLIEE 2014 , COLIEE 2015 (Kim et al., 2015), COLIEE 2016 (Kim et al., 2016), on a legal data collection, and this helped establish a major experimental effort in the legal information extraction/retrieval field. We held this fourth competition (COLIEE 2017) this year, with the motivation of continuing to help create a research community of practice for the capture and use of legal information. The COLIEE competition is about the legal question answering, which combines the two tasks of information retrieval and textual entailment.

We have previously held our COLIEE competitions in conjunction with the Japanese AI Society Juris-Informatics (JURISIN) workshop. The JURISIN workshop series was created to promote community discussion on both fundamental and practical issues on legal information processing, with the intention to embrace various backgrounds such as law, social science, information processing, logic and philosophy, and including the conventional "AI and law" area. This year's COLIEE 2017 is being held in conjunction with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017). Participating teams include HUKB(Yoshioka & Onodera, 2017), iLis7 (Heo et al., 2017), iLis9 (Jung et al., 2017), JAISTNLP (Nguyen et al., 2017), JNLP (Carvalho et al., 2017), KIS (Kano et al., 2017), NOR (Nanda et al., 2017), UA (Kim & Goebel, 2017), and NAIST (Morimoto et al., 2017).

K.Satoh, M.-Y.Kim, Y.Kano, R.Goebel and T.Oliveira (eds.), COLIEE 2017 (EPiC Series in Computing, vol. 47), pp. 1–8

2 The Legal Question Answering Task

This competition focuses on two aspects of legal information processing related to answering yes/no questions from Japanese legal bar exams (the relevant data sets have been translated from Japanese to English, and were available in both languages).

1) Task 1 of the legal question answering Task involves reading a legal bar exam question Q, and extracting a subset of Japanese Civil Code Articles S_1, S_2,..., S_n from the entire Civil Code, which are appropriate for answering the question such that

Entails(S_1, S_2, ..., S_n , Q) or Entails(S_1, S_2, ..., S_n , not Q).

Given a question Q and the entire Civil Code Articles, we have to retrieve the set of "S_1, S_2, ..., S_n" as the outcome of Task 1.

2) Task 2 of the legal question answering Task involves the identification of an entailment relationship such that

Entails(S_1, S_2, ..., S_n , Q) or Entails(S_1, S_2, ..., S_n , not Q).

Given a question Q and articles S_1, S_2, ..., S_n, we have to determine if relevant articles entail "Q" or "not Q". The answer of this track is binary: "YES"("Q") or "NO"("not Q"). This phase typically requires some information extraction (e.g., named entity identification, relation extraction, etc.), followed by any variety of methods for textual inference, to confirm entailments. Details are given in the next section.

2.1 Task 1

Our goal is to explore and evaluate legal document retrieval technologies that are both effective and reliable. The Task investigates the performance of systems that search a static set of civil law articles using previously unseen queries, and return relevant articles. We say an article is "Relevant" to a query if and only if the query sentence can be entailed from the meaning of the article. If combining the meanings of more than one article (e.g., "A," "B," and "C") can answer a query sentence, then all the articles ("A," "B,", and "C") are considered "Relevant." If a query can be answered by an article "D," and it can be also answered by another article "E" independently, we also say that both "D" and "E" are "Relevant." This Task requires the retrieval of *all* the articles that are relevant to answering a query.

Japanese civil law articles (and English translation) have been provided, and training data consists of query and relevant article pairs. The process of executing the queries over the articles and generating the experimental runs should be entirely automatic. Test data includes only queries but no relevant articles.

2.2 Task 2

The goal of Phase 2 is to construct Yes/No question answering systems for legal queries, by entailment from the relevant articles. The Task investigates the performance of systems that answer "Y" or "N" to previously unseen queries by somehow comparing the intersection of meanings between queries and relevant articles. Training data consists of triples: a query, relevant articles and a correct answer "Y" or "N." For the competition evaluation, the process of finding relevant articles, executing the queries over the relevant articles and generating the experimental runs should be entirely automatic. Test data includes only queries and corresponding "Y/N" labels for each query.

3 Legal Question Answering Data Corpus

The corpus of legal questions is drawn from Japanese Legal Bar exams, and the relevant Japanese Civil Law articles have been also provided.

1) Task 1 problem is to use an identified set of legal yes/no questions to retrieve relevant Civil Law articles. In this case, the correct answers have been determined by a collection of law students, and those answers are used to calibrate the performance of a program to solve Task 1.

2) Task 2 requires some method of information extraction from both the question and the articles, and then to confirm a simple entail relationship as described in the Section 2: either the articles confirms "yes" or "no" as an answer to the yes/no questions.

Participants can choose which task they will attempt, amongst the two tasks as follows:

Task 1: legal information retrieval task. Input is a bar exam "Yes/No" question and output should be relevant civil law articles.

Task 2: Recognizing Entailment between law articles and queries. Input is a question and the entire set of articles, and output should be "Yes" or "No.

Table 1 shows an example of query and relevant articles. Table 2 shows examples of sentence pairs holding different entailment relations.

Question	*A person who made a manifestation of intention which was induced by duress emanated from a third party may rescind such manifestation of intention on the basis of duress, only if the other party knew or was negligent of such fact.*
Related Article	*(Fraud or Duress) Article 96 (1)Manifestation of intention which is induced by any fraud or duress may be rescinded.(2)In cases any third party commits any fraud inducing any person to make a manifestation of intention to the other party, such manifestation of intention may be rescinded only if the other party knew such fact.(3)The rescission of the manifestation of intention induced by the fraud pursuant to the provision of the preceding two paragraphs may not be asserted against a third party without knowledge.*

Table 1. Example of a query and a relevant article

Question	A special provision that releases warranty can be made, but in that situation, when there are rights that the seller establishes on his/her own for a third party, the seller is not released of warranty.
Related Article	(Special Agreement Disclaiming Warranty)Article 572 Even if the seller makes a special agreement to the effect that the seller will not provide the warranties set forth from Article 560 through to the preceding Article, the seller may not be released from that responsibility with respect to any fact that the seller knew but did not disclose, and with respect to any right that the seller himself/herself created for or assigned to a third party.
Label	Yes
Question	A manager must engage in management exercising care identical to that he/she exercises for his/her own property.

Related Article	(Urgent Management of Business)Article 698 If a Manager engages in the Management of Business in order to allow a principal to escape imminent danger to the principal's person, reputation or property, the Manager shall not be liable to compensate for damages resulting from the same unless he/she has acted in bad faith or with gross negligence.
Label	No

Table 2. Examples of sentence pairs holding different entailment relations

The structure of the test corpora is derived from a general XML representation developed for use in RITEVAL, one of the tasks of the NII Testbeds and Community for Information access Research (NTCIR) project*.

The RITEVAL format was developed for the general sharing of information retrieval on a variety of domains. The format of the COLIEE competition corpora is derived from an NTCIR representation for confirmed relationships between questions and the articles and cases relevant to answering the questions, as in the following example:

<pair label="Y" id="H18-1-2">
<t1>
(Seller's Warranty in cases of Superficies or Other Rights) Article 566 (1) In cases where the subject matter of the sale is encumbered with for the purpose of a superficies, an emphyteusis, an easement, a right of retention or a pledge, if the buyer does not know the same and cannot achieve the purpose of the contract on account thereof, the buyer may cancel the contract. In such cases, if the contract cannot be cancelled, the buyer may only demand compensation for damages. (2)The provisions of the preceding paragraph shall apply mutatis mutandis in cases where an easement that was referred to as being in existence for the benefit of immovable property that is the subject matter of a sale, does not exist, and in cases where a leasehold is registered with respect to the immovable property.(3)In the cases set forth in the preceding two paragraphs, the cancellation of the contract or claim for damages must be made within one year from the time when the buyer comes to know the facts.
(Seller's Warranty in cases of Mortgage or Other Rights)Article 567(1) If the buyer loses his/her ownership of immovable property that is the object of a sale because of the exercise of an existing statutory lien or mortgage, the buyer may cancel the contract.(2)If the buyer preserves his/her ownership by incurring expenditure for costs, he/she may claim reimbursement of those costs from the seller.(3)In the cases set forth in the preceding two paragraphs, the buyer may claim compensation if he/she suffered loss.
</t1>
<t2>
There is a limitation period on pursuance of warranty if there is restriction due to superficies on the subject matter, but there is no restriction on pursuance of warranty if the seller's rights were revoked due to execution of the mortgage.
</t2>
</pair>

The above is an example where query id "H18-1-2" is confirmed to be answerable from article numbers 566 and 567 (relevant to Task 1). The pair label "Y" in this example means the answer of query is "Yes," which is entailed from the relevant articles (relevant to Task 2).

* As described at http://sites.google.com/site/ntcir11riteval/

The COLIEE training data was built from the bar exam (short answer test) civil code part published from 2006-2015, and consists of 10 XML files. Each file corresponds to one year's publication. The total number of queries in the training data is 570. The test data size is 78 queries (extracted from the bar exam of 2016).

4 Evaluation Metrics

The measures for ranking competition participants are intended only to calibrate the set of competition submissions, rather than provide any deep performance measure. The data sets for Task 1 are annotated, so simple information retrieval measures (precision, recall, F-measure, accuracy) can be used to rank each submission. The intention is to start to build a community of practice regarding legal textual entailment, so that the adoption and adaptation of general methods from a variety of fields is considered, and that participants share their approaches, problems, and results.

For Task 1, evaluation measure will be precision, recall and F-measure:

- Precision = (the number of correctly retrieved articles for all queries)/(the number of retrieved articles for all queries),
- Recall = (the number of correctly retrieved articles for all queries)/(the number of relevant articles for all queries),
- F-measure = (2 x Precision x Recall)/(Precision + Recall).

For Task 2, the evaluation measure will be accuracy, with respect to whether the yes/no question was correctly confirmed:

- Accuracy = (the number of queries which were correctly confirmed as true or false)/(the number of all queries).

5 Submitted Runs and Results

Overall, 10 teams submitted their system results. Some participants submitted multiple runs for a task. We received 9 submissions to Task 1 (for a total of 17 runs), and 8 submissions to Task 2 (for a total of 20 runs).

ID	Approaches
HUKB	article structure analysis, phrase matching, rank SVM with 15 similarity scores and alignment scores, ensemble
iLis7-1	TF-IDF, LSM, LDA, Word2Vec, LSA
JAISTNLP	TF-IDF
JNLP	ranking related n-gram collections, term order probabilities, relevance disambiguation.
NOR	LDA
UA	TF-IDF, language model

Table 3. Approaches of submitted systems for IR (Task 1)

ID	Approaches
iLis9	TF-IDF, negation detection
JAISTNLP	ranking, encoding-based and attention neural network
KIS	case-role based linguistic analysis, predicate-argument structures
NAIST	Word2vec, attention neural network
NOR	CNN, LSTM
UA	Korean dependency parser, Excite Japanese/Korean machine translation, semantic dictionary, k-means clustering

Table 4. Approaches of submitted systems for textual entailment (Task 2)

Tables 3 and 4 summarize the submitted systems' techniques, including systems when corresponding proceedings exist. We present the results achieved by runs against the Information

Run	Prec.	Recall	F	L.	Run	Prec.	Recall	F	L.
HUKB-1	0.658	0.472	0.550	J	JNLP1-T	0.500	0.354	0.414	E
HUKB-2	0.586	0.490	0.534	J	KID17	0.703	0.518	0.596	E
HUKB-3	0.551	0.536	0.543	J	KIS-IE-M	0.263	0.272	0.267	J
iLis7-1	0.734	0.554	0.632	E	KIS-IE-NM	0.346	0.245	0.287	J
iLis7-2	0.654	0.500	0.567	E	NOR17	0.462	0.500	0.480	E
JAISTNLP2-1a-norerank	0.628	0.445	0.521	E	UA-LM	0.602	0.427	0.500	E
JAISTNLP2-1b-rerank	0.615	0.436	0.510	E	UA-TFIDF	0.666	0.472	0.553	E
JNLP1-R	0.686	0.536	0.602	E	VNPT	0.430	0.281	0.340	E
JNLP1-RT	0.689	0.545	0.609	E					

Table 5. IR results (Task 1) on the formal run data. Prec, F, and L stands for Precision, F-measure, and Data Language respectively. E and J stands for English and Japanese in the Language columns.

Run	Accuracy	Language	Run	Accuracy	Language
iLis7	0.564	English	KIS-YN-CM	0.538	Japanese
iLis9-1	0.576	English	KIS-YN-CS	0.589	Japanese
iLis9-2	0.538	Japanese	KIS-YN-M	0.576	Japanese
JAISTNLP2-2a-1a-norerank	0.512	English	KIS-YN-S	0.653	Japanese
JAISTNLP2-2a-1b-rerank	0.474	English	NAIST1	0.615	Japanese
JAISTNLP2-2b-1a-norerank	0.487	English	NAIST2	0.653	Japanese
JAISTNLP2-2b-1b-rerank	0.500	English	NAIST3	0.474	Japanese
JNLP1-R	0.435	English	NOR17	0.538	English
JNLP1-RT	0.487	English	UA-LM	0.717	Japanese
KIS-YN-A	0.538	Japanese	UA-TFIDF	0.692	Japanese

Table 6. Entailment results (Task 2) on the formal run data

Retrieval and Entailment tasks in Tables 5 and 6.

We performed comparisons between the top three runs in Task 2. Among 78 queries, the best run correctly answered 56 queries. When comparing the best and the second, matched answers were 45 and 41 queries for each second team. These three runs agreed in only 27 queries, where 25 were correct answers. This comparison implies that the 25 queries could be similar in surficial level, which are easy to answer correctly regardless of employed methods. It is difficult to determine why around 20 queries are non-agreed, of which at least one of these three teams answered correctly. It might be due to the difference of the methods. However, if we assume that the 25 queries are easy-to-answer, then the there are 53 that remain. This means that we can answer 26 queries correctly even in a chance rate, which is higher than the 20 non-agreed queries. Although we would need more evaluation data to obtain more stable statistic evaluation, this comparison suggests that there is still lots of opportunity to improve the competition systems in future.

There will be a live competition held in the conference as our first trial. Each team is required to run the very system used to submit the formal run results, while asked to return their answers in real time.

6 Conclusion

We have summarized the results of the COLIEE 2017 competition. Two tasks were evaluated: (1) Task 1: finding relevant articles (information retrieval) (2) Task 2: answering yes/no questions given a query (textual entailment).

There were 10 teams who participated in this competition. There were 9 submissions to Task 1 (for a total of 17 runs), and 8 submissions to Task 2 (for a total of 20 runs).

While the evaluation scores are getting higher from previous years, there are still many unresolved issues in these tasks, as suggested by the absolute evaluation scores and comparison results.

Acknowledgements

This research was partially supported by MEXT Kakenhi Japan, JST CREST Japan, and the Alberta Machine Intelligence Institute (Amii).

References

Carvalho, D. S., Tran, V., Tran, K. Van, & Nguyen, L. M. (2017). Improving Legal Information Retrieval by Distributional Composition with Term Order Probabilities. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Heo, S., Hong, K., & Rhim, Y.-Y. (2017). Legal Content Fusion for Legal Information Retrieval. In *the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Jung, B., Soh, C., Hong, K., Lim, S., & Rhim, Y.-Y. (2017). Multiple Agent Based Entailment System (MABES) for RTE. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Kano, Y., Hoshino, R., & Taniguchi, R. (2017). Analyzable Legal Yes/No Question Answering System using Linguistic Structures. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Kim, M.-Y., & Goebel, R. (2017). Two-step Cascaded Textual Entailment for Legal Bar Exam Question Answering. In *the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Kim, M.-Y., Goebel, R., Kano, Y., & Satoh, K. (2016). COLIEE-2016: Evaluation of the Competition on Legal Information Extraction/Entailment. In *Tenth International Workshop on Juris-informatics (JURISIN 2016)*.

Kim, M.-Y., Goebel, R., & Ken, S. (2015). COLIEE-2015: Evaluation of Legal Question Answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*. Keio University, Yokohama, Japan.

Morimoto, A., Kubo, D., Sato, M., Shindo, H., & Matsumoto, Y. (2017). Legal Question Answering System using Neural Attention. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Nanda, R., John, A. K., Caro, L. Di, Boella, G., & Robaldo, L. (2017). Legal Information Retrieval Using Topic Clustering and Neural Networks. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Nguyen, T.-S., Phan, V.-A., & Nguyen, L.-M. (2017). Recognizing entailments in legal texts using sentence encoding-based and decomposable attention models. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.

Yoshioka, M., & Onodera, D. (2017). A Civil Code Article Information Retrieval System based on Phrase Alignment with Article Structure Analysis and Ensemble Approach. In *4th Competition on Legal Information Extraction and Entailment (COLIEE 2017), 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*.