COLIEE-2015 : Evaluation of Legal Question Answering

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

University of Alberta, Canada[1]
National Institute of Informatics, Japan[2]
{miyoung2, rgoebel}@ualberta.ca, ksatoh@nii.ac.jp

Abstract. We present the details on evaluation of legal question answering of the Competition on Legal Information Extraction/Entailment (COLIEE)-2015. The COLIEE-2015 task consists of three sub-tasks: legal information retrieval (sub-task 1), recognizing entailment between articles and queries (sub-task 2), and combination of the previous two sub-tasks (sub-task 3). Participation was open to any group based on any approach, and the task attracted 6 teams. We received 6 submissions to the sub-task 1 (for a total of 12 runs), 4 submissions to the subtask 2 (for a total of 11 runs), and 4 submissions to the subtask 3 (for a total of 11 runs).

Keywords: legal question answering, recognizing textual entailment, information retrieval

1. Introduction

The Juris-Informatics workshop series has been created to promote community discussion on both fundamental and practical issues from various backgrounds such as law, social science, information and intelligent technology, logic and philosophy, including the conventional "AI and law" area.

The information extraction/reasoning from legal data is one of the important targets of JURISIN, including legal information representation, relation extraction, textual entailment, summarization, and their applications. Participants in the JURISIN workshops have examined a wide variety of information extraction techniques and environments with a variety of purposes, including retrieval of relevant articles, entity/relation extraction from legal cases, reference extraction from legal cases, finding relevant precedents, summarization of legal cases, and legal question answering.

In 2014, we held the first competition on legal information extraction/entailment (COLIEE-2014) on a legal data collection, and it helped establish a major experimental effort in the legal information extraction/retrieval field. We held the second competition (COLIEE-2015) 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 task, and our decomposition into three subtasks of retrieval, information extraction, and entailment.

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).

1) Phase 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 phase 1.

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

2.1 Phase 1 Details

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 consider 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 pairs of a query and relevant articles. 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 Phase 2 Details

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 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." The process of executing the queries over the relevant articles and generating the experimental runs should be entirely automatic. Test data includes only queries and relevant articles, but no "Y/N" label.

2.3 Phase 3 Details

Phase 3 combines Phase 1 and Phase 2. If a "Yes/No" legal bar exam question is given, a legal information retrieval system retrieves relevant Civil Law articles. Then, the task investigates the performance of systems that answer "Y" or "No" to previously unseen queries by comparing the meanings of the queries and retrieved Civil Law articles. Test data includes only queries, but no "Y/N" label, and no relevant articles.

3. The 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) Phase 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 Phase 1.

2) Phase 2 task requires some method of information extraction from both the question and the relevant 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.

3) Phase 3 task is combination of Phase 1 and Phase 2. It requires both of the legal information retrieval system and textual entailment system. If legal yes/no questions are given, each participant's legal information retrieval system retrieves relevant Civil Law articles. After that, each participant confirms a "Yes/No" entailment relationship between input yes/no question and the retrieved articles.

Participants can choose which Phase they will attempt, amongst the three sub-tasks as follows:

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

Table 1. Example of a query and a relevant article

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 2. Examples of sentence pairs holding different entailment relations

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

2. Sub-task 2: Recognizing Entailment between law articles and queries. Input is a pair of a question and relevant article(s), and output should be “Yes” or “No. (Phase 2)

3. Sub-task 3: Combination of sub-task 1 and sub-task 2. Input is a bar exam “Yes/No”question and output should be “Yes” or “No.” (Phase 3)

Table 1 shows an example of query and relevant articles.

In the entailment subtask, given two sentences A and B, a system must determine whether the meaning of B was entailed by A. In particular, a system is required to assign to each pair either the “Yes” label (when A entails B, viz., B cannot be false

when A is true), or the “No” label (when A does not entail B). Table 2 shows examples of sentence pairs holding different entailment relations.

The subtask 3 is combination of subtask 1 and subtask 2.

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, as described at the following URL:

http://sites.google.com/site/ntcir11riteval/

The RITEVAL format was developed for the general sharing of information retrieval on a variety of domains. The format of the JURISIN competition corpora derived from an NTCIR representation of 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 Phase 1). The pair label “Y” in this

example means the answer of query is “Yes", which is entailed from the relevant articles (relevant to Phase 2 and 3).

COLIEE training data was built from the bar exam (short answer test) civil code part published in 2006-2011, and consist of 6 XML files. Each file corresponds to one year’s publication. The total number of queries in the training data is 267, and each query has average 1.25 relevant articles. The test data size is 79 queries for Phase 1, extracted from the bar exam of 2012, and 66 queries for Phase 2 from the bar exam of 2013. For Phase 3, we use the same test data of Phase 1.

4. Evaluation Metrics and Baselines

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 Phases 1 and 2 are annotated, so simple information retrieval measures (precision, recall, F-measure [7], accuracy) can be used to rank each submission. As noted above, 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 Phase 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 Phase 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).

For Phase 3, the evaluation measure is the same with that of Phase 2.

For Phase 1, we consider the traditional TF-IDF [6] as the baseline model. For Phase 2, there is a balanced positive-negative sample distribution in the dataset (51.52% yes, and 48.48% no) for the formal run, so we consider the baseline for true/false evaluation is the accuracy when returning always “yes,” which is 51.52%.
Tables 3 and 4 presents the performances of the baseline models.

Table 3. Baseline of Phase one

Phase One Baseline	F-measure
TFxIDF with lemmarization (choosing the most relevant article)	0.4862

Table 4. Baseline of Phase two

Phase Two Baseline	Accuracy
When 'No' labels are all chosen	0.5152

5. Submitted Runs and Results

Overall, 6 teams participated in the task. Some participants submitted multiple runs for a task. We received 6 submissions to the subtask 1 (for a total of 12 runs), 4 submissions to the subtask 2 (for a total of 11 runs), and 4 submissions to the subtask 3 (for a total of 11 runs).

Tables 5 and 6 show the submitted systems' techniques. Because we have not received one participant's paper, the tables include 5 teams' runs and their techniques. We present the results achieved by runs against the Information Retrieval and Entailment subtasks in Tables 7, 8, and 9, respectively. Some systems performed well above the best baselines from Tables 3 and 4.

D. S. Carvalho et al. [2] extracted n-gram features from the query and articles using lexical and morphological characteristics, and then used their own relevance score for Phase 1. For Phase 2, the query and relevant article are represented in vector space model, and then they used AdaBoost (basic classifier: DecisionStump) to classify the queries between 'Yes' label and 'No' label.

Sushimita et al. [3] participated in only Phase 1, and used voting of three information retrieval techniques: Hiemstra, BM25 and PL2F.

Y. Kano [4] introduced their own keyword weighting method and snippet scoring after diving data into snippets . V.D.Tran et al. [5] participated in only Phase 1 and used Hidden Markov model (HMM) as a generative query model for legal information retrieval.

Table 5. Approaches of submitted systems for IR (Phase 1)

ID	Run	Approaches
UA [1]	UA1	Ranking SVM
	UA2	TFxIDF with lemmarization
Kano-lab[4]	Kanolab1	their own keyword weighting method snippet scoring (ptag=0.3)
	Kanolab2	their own keyword weighting method snippet scoring (ptag=0.4)
	Kanolab3	their own keyword weighting method snippet scoring (ptag=0.5)
	Kanolab4	their own keyword weighting method snippet scoring (topic=0.1)
	Kanolab5	their own keyword weighting method snippet scoring (topic=0.3)
JAIST [2]	JAIST01	Ranking by n-gram model
ALV [5]	ALV2015	Hidden Markov Model
	ALV2015-2	Hidden Markov Model with different parameter setting
Ismir [3]	ismir	Voting of three models: Hiemstra, BM25 and PL2F

Table 6. Approaches of submitted systems for the Entailment (Phase 2)

ID	Run	Approaches
UA [1]	UA	Convolutional Neural Network
Kano-Lab [4]	Kanolab1	Assigning Threshold to the snippet score (ptag=0.3)
	Kanolab2	Assigning Threshold to the snippet score (ptag=0.4)
	Kanolab3	Assigning Threshold to the snippet score (ptag=0.5)
	Kanolab4	Assigning Threshold to the snippet score (Section=0.1)
	Kanolab5	Assigning Threshold to the snippet score (Section=0.2)
	Kanolab6	Assigning Threshold to the snippet score (Section=0.4)
	Kanolab7	Assigning Threshold to the snippet score (Topic=0.1)
	Kanolab8	Assigning Threshold to the snippet score (Topic=0.3)
JAIST [2]	JAIST01	Adaboost (classifier: DecisionStump) using distance features and statistical features

The UA2 method in Table 7 uses TFxIDF with lemmatization which is the baseline

Table 7. IR results(Phase 1) on the formal run data

Run	Prec.	Recall	F-m.	Run	Prec.	Recall	F-m.
UA1 [1]	0.6329	0.4902	0.5525	Kanolab5[4]	0.3038	0.2353	0.2652
UA2 [1]	0.5570	0.4314	0.4862	JAIST01[2]	0.5663	0.4608	0.5081
Kanolab1[4]	0.3291	0.2549	0.2873	ALV2015[5]	0.3418	0.5294	0.4154
Kanolab2[4]	0.3291	0.2549	0.2873	ALV2015-2[5]	0.2250	0.3529	0.2748
Kanolab3[4]	0.3291	0.2549	0.2873	Ismir[3]	0.2778	0.2941	0.2857
Kanolab4[4]	0.3038	0.2353	0.2652				

Table 8. Entailment results (Phase 2) on the formal run data

Run	Accu.	Run	Accu.
UA [1]	0.6667	Kanolab5 [4]	0.5152
Kanolab1 [4]	0.6061	Kanolab6 [4]	0.5152
Kanolab2 [4]	0.6212	Kanolab7 [4]	0.5152
Kanolab3 [4]	0.6212	Kanolab8 [4]	0.5152
Kanolab4 [4]	0.5152	JAIST01 [2]	0.3788

Table 9. Combined results (Phase 3) on the formal run data

Run	Accu.	Run	Accu.
UA [1]	0.6582	Kanolab5 [4]	0.4937
Kanolab1 [4]	0.5696	Kanolab6 [4]	0.5317
Kanolab2 [4]	0.5949	Kanolab7 [4]	0.4557
Kanolab3 [4]	0.6203	Kanolab8 [4]	0.5696
Kanolab4 [4]	0.4684	JAIST01 [2]	0.5823

model. We can see that there are two systems which show better performance than the baseline model: UA1 and JAIST01. UA1 shows the best performance using ranking SVM, and it achieved significant increase of the F-measure compared to the baseline model. JAIST01 used their own ranking score from n-gram model. UA1, UA2 and Kanolab produced only one most relevant article for each query (in total 79 articles). JAIST01 produced 82 articles, and ALV2015 and ALV2015-2 produced 157 articles and 160 articles respectively. The reason that ALV2015 shows best recall is that they produced 157 results (average 1.99 articles per query), but their precision was not good. Ismir also produced 108 results, which is average 1.37 articles for each query.

Table 8 reports the results of Phase 2, where UA showed best accuracy, which is significantly better than the baseline performance. They used a convolutional neural network. Kanolab1, Kanolab2 and Kanolab3 also showed significantly increased

performance than the baseline, and they used their own threshold value from the snippet scoring to determine if the label is “yes” or “no.”

In the Table 9, results of Phase 3, Kanolab6, Kanolab8, and JAIST01 showed better performance than their performances in Phase 2. That means their answering of “yes” “no” questions is better when they used their own relevant articles than those marked as correct in the training set. This indicates the need for further detailed investigation.

6. Conclusion

We presented the results of the COLIEE-2015 competition. Three subtasks were offered: (1) Phase 1: finding relevant articles (information retrieval) (2) Phase 2: answering yes/no questions by comparing the meanings between a query and relevant articles (textual entailment), (3) combination of Phase 1 and Phase 2. There were 6 submissions to Phase 1 (for a total of 12 runs), 4 submissions to Phase 2 (for a total of 11 runs), and 4 submissions to Phase 3 (for a total of 11 runs), for a total of 6 participating teams. Various methods for Phase 1 are provided: ranking SVM, TFxIDF, n-gram features from the query and articles using lexical and morphological characteristics and their own relevance score. Some information retrieval techniques: Hiemstra, BM25 and PL2F, keyword weighting method and snippet scoring, and Hidden Markov Model have also been proposed. For Phase 2, the participants used AdaBoost using their own features, a convolutional Neural Network, and their own threshold from the snippet score. Most of the systems do not depend on the deeply linguistic features which require domain knowledge, but instead propose feature weighting and scoring methods. Some systems significantly outperformed baseline systems, and we need to deeply investigate how the performances of each Phase 1 and Phase 2 affect the overall performance of Phase 3.

Acknowledgements

This research was supported by the Alberta Innovates Centre for Machine Learning (AICML), the Natural Sciences and Engineering Research Council (NSERC), and also partially supported by JSPS KAKENHI Grant Numbers 26280091.

References

1. M-Y. Kim, Y. Xu, and R. Goebel, “A Convolutional Neural Network in Legal Question Answering”, Ninth International Workshop on Juris-informatics (JURISIN), 2015
2. D. S. Carvalho, M-T. Nguyen, C-X. Tran, and M-L. Nguyen, “Lexical-Morphological Modeling for Legal Text Analysis”, Ninth International Workshop on Juris-informatics (JURISIN), 2015
3. Sushmita, A. Kanapala, and S. Pal, “Legal Information Retrieval Task: Participation from ISM, Dhanbad”, Ninth International Workshop on Juris-informatics (JURISIN), 2015
4. Y. Kano, “Keyword and Snippet Based Yes/No Question Answering System for COLIEE 2015”, Ninth International Workshop on Juris-informatics (JURISIN), 2015

5. V. D. Tran, A. V. Phan, L. H. Trieu, and M. L. Nguyen, “An Approach for Retrieving Legal Texts”, Ninth International Workshop on Juris-informatics (JURISIN), 2015
6. K. J. Sparck. “A statistical interpretation of term specificity and its application in retrieval.” Journal of documentation 28.1, 11-21, 1972
7. G. Hripcsak, and A. S. Rothschild. “Agreement, the f-measure, and reliability in information retrieval”. Journal of the American Medical Informatics Association, 12(3), 296-298, 2005