Question Answering of Bar Exams by Paraphrasing and Legal Text Analysis

Mi-Young Kim[3(✉)], Ying Xu[1], Yao Lu[2], and Randy Goebel[1]

[1] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB, Canada
{yx2, rgoebel}@ualberta.ca
[2] iLab Tongji, School of Software Engineering, Tongji University, Shanghai, China
95luyao@tongji.edu.cn
[3] Department of Science, University of Alberta, Augustana Campus, Camrose, AB, Canada
miyoung2@ualberta.ca

Abstract. Our legal question answering system combines legal information retrieval and textual entailment, and exploits paraphrasing and sentence-level analysis of queries and legal statutes. We have evaluated our system using the training data from the competition on legal information extraction/entailment (COLIEE)-2016. The competition focuses on the legal information processing required to answer yes/no questions from Japanese legal bar exams, and it consists of three phases: legal ad-hoc information retrieval (Phase 1), textual entailment (Phase 2), and a combination of information retrieval and textual entailment (Phase 3). Phase 1 requires the identification of Japan civil law articles relevant to a legal bar exam query. For this phase, we have used an information retrieval approach using TF-IDF and a Ranking SVM. Phase 2 requires decision on yes/no answer for previously unseen queries, which we approach by comparing the approximate meanings of queries with relevant articles. Our meaning extraction process uses a selection of features based on a kind of paraphrase, coupled with a condition/conclusion/exception analysis of articles and queries. We also identify synonym relations using word embedding, and detect negation patterns from the articles. Our heuristic selection of attributes is used to build an SVM model, which provides the basis for ranking a decision on the yes/no questions. Experimental evaluation show that our method outperforms previous methods. Our result ranked highest in the Phase 3 in the COLIEE-2016 competition.

Keywords: Legal question answering · Recognizing textual entailment · Information retrieval · Paraphrasing

1 Task Description and Summary of Our Approach

Our approach to legal question answering combines information retrieval and textual entailment. We achieve this combination with a number of intermediate steps. For instance, consider the question "Is it true that a special provision that releases warranty can be made, but in that situation, when there are rights that the seller establishes on

S. Kurahashi et al. (Eds.): JSAI-isAI 2016, LNAI 10247, pp. 299–313, 2017.
DOI: 10.1007/978-3-319-61572-1_20

his/her own for a third party, the seller is not released of warranty." A system must first identify and retrieve relevant documents (typically legal statutes), and subsequently, identify a most relevant sentence. Finally it must extract and compare semantic connections between the question and the relevant sentences, and confirm a threshold of evidence about whether an entailment relation holds.

The Competition on Legal Information Extraction/Entailment (COLIEE) 2016[1] focuses on two aspects of legal information processing related to answering yes/no questions from legal bar exams: legal document retrieval (Phase 1), and whether there is a textual entailment relation between a query and relevant legal documents (Phase 2). In addition, Phase 3 is about combing them for the whole task.

We treat Phase 1 as an ad-hoc information retrieval (IR) task. The goal is to retrieve relevant Japan civil law statutes or articles that are related to a question in legal bar exams, from which we can confirm a yes or no answer based on deciding if there is an entailment relation between the question (or the negation of the question) and the relevant statutes.

We approach the information retrieval part of this problem (Phase 1) with two models based on statistical information. One is the TF-IDF model [1], i.e., term frequency-inverse document frequency. The idea is that relevance between a query and a document depends on their intersecting word set. The importance of words is measured with a function of term frequency and document frequency as parameters. Our terms are lemmatized words, which mean verbs like "attending," "attends," and "attended" are lemmatized as the same form "attend."

Another popular model for text retrieval is a Ranking SVM model [2]. We use that model to re-rank documents that are retrieved by the TF-IDF model. The model's features are lexical words, dependency path bigrams and TF-IDF scores. The intuition is that the supervised model can learn weights or priority of words based on training data in addition to, or as an alternative to TF-IDF.

The goal of Phase 2 is to construct yes/no question answering systems for legal queries, by heuristically confirming entailment of a query (or its negation) from relevant articles. The answer to a question is typically determined by measuring some kind of semantic similarity between question and answer. Because the legal bar exam query and relevant articles are complex and varied, we need to carefully determine what kind of information is needed for confirming textual entailment. Here we exploit a kind of paraphrasing based on term expansion and word embedding for semantic analysis, coupled with condition/conclusion/exception analysis on the query and relevant articles. After constructing a set of pre-trained semantic word embeddings using *word2vec*[2], we train the system to learn models for semantic matching between question and corresponding articles. These feature extraction methods are coupled with negation analysis, then used to construct an SVM model to provide the required yes/no answers.

[1] https://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2016/.

[2] https://code.google.com/p/word2vec.

2 Phase 1: Legal Information Retrieval

2.1 IR Models

Our information retrieval model is a combination of the term frequency–inverse document frequency (tf-idf) model and a support vector machine (SVM) re-ranking model. We will describe the two components in the following.

2.1.1 The TF-IDF Model

One of our baseline models is a tf-idf model implemented in Lucene, an open source IR system[3].

The simplified version of Lucene's similarity score of an article to a query is:

$$tf-idf(Q,A) = \sum_{t \in Q \cup A} \{\sqrt{tf(t,A)} \times [1 + \log(idf(t))]^2\} \quad (1)$$

The score *tf-idf*(*Q*,*A*) is a measure which estimates the relevance between a query Q and an article A. First, for every term t in query A, we compute *tf*(*t*,*A*), and *idf*(*t*). The score *tf*(*t*,*A*) is the term frequency of t in the article A, and *idf*(*t*) is the inverse document frequency of the term t, which is the number of articles that contain t. The final score is the sum of the scores of terms in both the article and the query. The bigger *tf-idf*(*Q*,*A*), the more relevance between the query Q and the article A.

The choice of terms in documents is as important as choosing the score functions. Instead of using the original words in a text, we lemmatize the text with the Stanford NLP tool [15]. After lemmatization, words such as *steal, stole, and steals* become *steal*. In this way, if there is *steal* in the question, but *stole* in the article, we can still retrieve the article as a match.

2.1.2 The Ranking SVM Model

Previous tf-idf models rank the articles based on frequency information. However, other features, such as the matched phrases between the article and the queries, are useful too. We use an SVM Ranking model to learn the importance of such features and then re-estimate the score of each retrieved article from the tf-idf output.

The ranking SVM model was proposed by [2]. That model ranks documents based on user's click through data; in our case, the correct articles in the training data. Given the articles retrieved from the *tf-idf* model, the ranking SVM will learn to rank correct articles higher than incorrect ones. More precisely, given the feature vector of a training instance, i.e. a retrieved article set given a query, denoted by $\Phi(Q,A_i)$, the model tries to find a ranking that satisfies constraints:

$$\Phi(Q,A_i) > \Phi(Q,A_j) \quad (2)$$

where A_i is a relevant article for the query Q, while A_j is less relevant.

[3] Lucene can be downloaded from http://lucene.apache.org/core/.

To use this ranking SVM, we incorporate the following types of features:

- Lexical words: the lemmatized normal form of surface structure of words in both the retrieved article and the query. In the conversion to the SVM's instance representation, this feature is converted into binary features whose values are one or zero, i.e. if a word exists in the intersection word set or not.
- Dependency pairs: word pairs that are linked by a dependency link, arising from a dependency parsing. The intuition is that, compared with the bag of words information, syntactic information should improve the capture of salient semantic content. Dependency parse features have been used in many NLP tasks, and improved IR performance [3]. This feature type is also converted into binary values.
- TF-IDF score (Sect. 2.1.1).

2.2 Experiments

The COLIEE legal IR task has several sets of queries with the Japan civil law articles as documents (1044 articles in total). Here follows one example of the query and a corresponding 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.*

Before the final test set was released, we received 8 sets of queries for a dry run. The 8 sets of data include 412 queries. We used the corresponding 8-fold leave-one-out cross validation evaluation. The metric for measuring our IR models is Mean Average Precision (MAP):

$$\mathrm{MAP}(\mathrm{Q}) = \frac{1}{|\mathrm{Q}|}\sum_{\mathrm{q}\in\mathrm{Q}}\frac{1}{\mathrm{m}}\sum_{\mathrm{k}\in(1,\mathrm{m})}\mathrm{precision}\,(\mathrm{R_k}) \tag{3}$$

where Q is the set of queries, and m is the number of retrieved articles. Rk is the set of ranked retrieval results from the first until the k^{th} article. In the following experiments, we set m as 3 for all queries, corresponding to the column MAP@3 in Table 1. The SVM's parameters are set according to the 8-fold cross validation IR performance. Given the top 20 articles returned by the tf-idf model, the SVM model extracts features for every article and trains according to the order that relevant articles are ranked higher than irrelevant ones.

Table 1 presents the results of using the different models. The result shows that the ranking SVM with all three features achieves the highest performance. We also show

Table 1. IR results on dry run data with different models.

Id	Models	MAP@3 (%)	Standard deviation (%)	Smallest (%)	Largest (%)	Average F-score @1 (%)
1	tf-idf with lemma	39.8	7.0	23.8	45.5	53.4
2	SVM-ranking with lemma	39.8	6.5	26.1	46.8	60.0
3	SVM-ranking with lemma and dependency pair	41.2	5.4	30.1	48.4	56.7
4	Model 3 plus tf-idf score	43.1	7.1	27.2	49.0	60.0

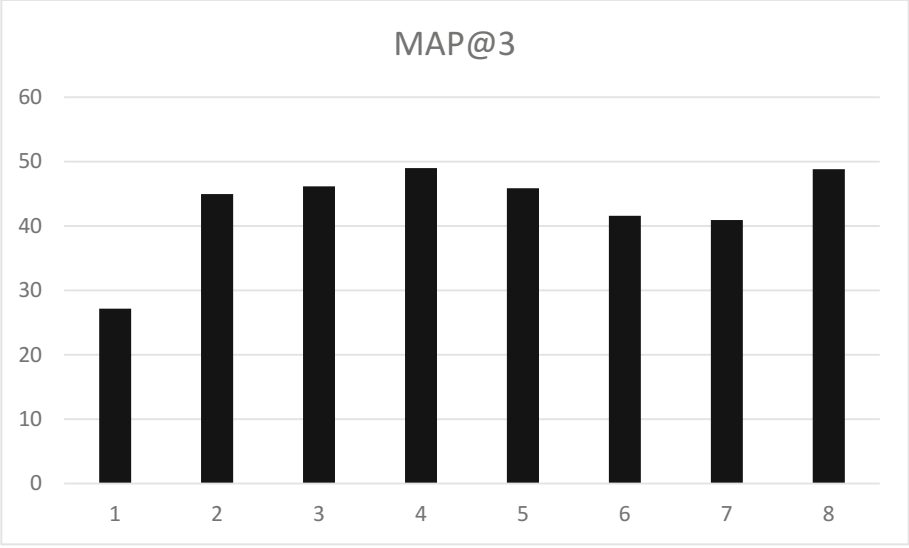

Fig. 1. MAP@3 for the 8 cross-validation set of Model 4.

the standard deviation of the cross-validation, the smallest and largest MAP@3 for 8 folds to show the effect of small training data. It seems the tf-idf model causes larger deviation than the SVM models. In the last column of Table 1, we show the F-score results of different models with the top first answers for every query. The F-score is used as the metric of the competition. We can observe that the SVM models are better than the tf-idf model. However, no difference is observed between the second model and the fourth model.

Figure 1 shows the MAP@3 values for every training fold for Model 4 in Table 1. It shows the model achieves a MAP@3 value larger than 40% for most of the folds.

Table 2 shows our IR result of the final SVM model on the test data and other systems' results. iLis7 [19] system with majority vote of decision tree, linear SVM, and CNN achieved the best result, but in Sect. 3, we will show that our method

outperformed iLis7 [19] and showed the best performance in answering yes/no questions when it is combined with our textual entailment methods.

Table 2. Our IR results on test data vs. other systems' results

Systems	Precision	Recall	F-score
JNLN1 [17]	0.6105	0.4427	0.5133
HUKB-1 [21]	0.6154	0.4886	0.5447
HUKB-2 [21]	0.6250	0.4962	0.5532
HUKB-3 [21]	0.6316	0.4580	0.5310
HUKB-4 [21]	0.6316	0.4580	0.5310
JNLN2 [22]	0.6211	0.4504	0.5221
iLis7 [19]	0.7272	0.5496	0.6261
JNLN3 [20]	0.6526	0.4733	0.5487
N01-1 [23]	0.3053	0.2214	0.2566
N01-2 [23]	0.4211	0.3053	0.3540
N01-3 [23]	0.4000	0.2901	0.3363
Our system (SVM-ranking with lemma and dependency pair and tf-idf score)	0.5895	0.4275	0.4956

Table 3. Query-article types

Query-article type	Proportion	Query-article type	Proportion
One article refers to another article	0.182	Question is a specific example	0.092
Multiple relevant articles	0.388	Multiple conditions	0.731
Exceptional case	0.148		

3 Phase 2: Answering Yes/No Questions

Our system uses a combination of word embedding for semantic analysis and paraphrasing for term expansion to predict textual entailment. Here we describe the entailment types and the extraction of features from sentences.

3.1 Entailment Types

We identify a variety of types of entailment as shown in Table 3. By classifying a yes/no problem as one of these types, we can determine what kind of further information is required to provide a decision on entailment.

Table 3 shows our list of query-article types. Note that one article can refer to another, such as "*If there is any latent defect in the subject matter of a sale, the provisions of Article 566 shall apply mutatis mutandis.*" This makes textual entailment more complex because we also need to analyze the meaning of the referred article.

In another case, one query can have multiple relevant articles, so we have to combine the multiple articles' meanings, or choose one as most relevant for determining entailment.

Note also that many statutes have exceptional cases, so we need to recognize if the query is included in the exceptional case or not. In addition, a query may be one example of the article case. There are also cases where some articles have multiple conditions for one conclusion, so we must then confirm if each condition is satisfied in the query. Overall, many query-article types also require the identification of negation and synonym/antonym relations to confirm the correct entailment.

The overall description of our procedure of textual entailment is as follows:

1. Find the most relevant article for a given query
2. Divide a query and the corresponding article into "Condition(s)," "Conclusion," and "Exception-condition(s)."
3. Term expansion using Paraphrasing
4. Negation and synonym detection
5. Extract features and perform learning using the features

In the following subsections, we explain each step in detail.

3.2 Finding the Most Relevant Article/Sentences

In case that there are multiple relevant sentences, we choose the article with the most overlapping words with the query. In the selected article, if there exist multiple regulations, we also choose the one regulation that has most overlapping words with a query.

3.3 Negation and Synonym Detection

We exploit a process for managing negation and antonyms as described in Kim et al. [10]. In addition, we approximate word semantic similarity by converting words to vector representations using the *word2vec* tool. The output of the semantic similarity is vector similarity. We used 1,044 legal law articles to train the *word embedding* by setting the *word2vec* vector dimension to 50 which has been most commonly chosen as the vector dimension in previous work.

3.4 Condition/Conclusion/Exception Detection

From our analysis of the structure of statutes we extract components based on the following rules:

$$conclusion := segment_{last}(sentence, keyword),$$

$$condition := \sum_{i \neq last} segment_i(sentence, keyword),$$

$$condition := condition \text{ [or] } condition$$

$$condition := sub\_condition \text{ [and] } sub\_condition$$

$$exception\_conclusion := segment_{last}(sentence, exception\_keyword),$$

$$exception\_condition := \sum_{i \neq last} segment_i(sentence, exception\_keyword),$$

$$exception\_condition := exception\_condition \text{ [or] } exception\_condition$$

$$exception\_condition := sub\_exception\_condition \text{ [and] } sub\_exception\_condition$$

So from keywords of a condition, we segment sentences. The keywords of the condition are as follows: “in case(s),” “if,” “unless,” “with respect to,” “when,” and “(comma).” After this segmentation, the last segment is considered to be a conclusion, and the rest of the sentence is considered as a condition. (We used the symbol $\sum$ to denote the concatenation of the segments.) We also distinguish segments which denote exceptional cases. Currently, we take the *exception_keyword* indication as “… this shall not apply, if (unless).”

The original bar law examinations in the COLIEE data are provided in Japanese and English, and our initial implementation used a Korean translation, provided by the Excite translation tool[4]. We chose Korean because we have a team member whose native language is Korean, and the characteristics of Korean and Japanese language are similar. In addition, the translation quality between two languages ensures relatively stable performance. Because our study team includes a Korean researcher, we can easily analyze the errors and intermediate rules in Korean. Therefore, the above rules may not be appropriate for all English sentences, because the segment order can differ.

The following is an example of condition and conclusion detection:

> <Civil law example> A person who employs others for a certain business, shall be liable for damages inflicted on a third party by his/her employees with respect to the execution of that business; Provided, however, that this shall not apply, if the employer exercised reasonable care in appointing the employee or in supervising the business, or if the damages could not have been avoided even if he/she had exercised reasonable care.

[4] http://excite.translation.jp/world/.

(1) Conclusion => shall be liable for damages inflicted on a third party by his/her employees with respect to the execution of that business.
(2) Condition => A person who employs others for a certain business
(3) Exception
Conclusion => this shall not apply (opposite of main conclusion)
Condition
Condition =>
Condition => if the employer exercised reasonable care in appointing the employee
Condition (OR) => in supervising the business
Condition(OR) => if the damages could not have been avoided even if he/she had exercised reasonable care.

3.5 Term Expansion Using Paraphrasing

There are many words with similar meanings but different lexical forms (e.g., 'obligor' vs. 'debtor', 'rescind' vs. 'cancel', 'lien' vs. 'privilege', etc.). To resolve these diverse terms, we use language translation-based paraphrasing. The idea of translation-based paraphrase is that translating from one language to another and then back, will often produce semantically similar but lexically distinct outputs. If we assume that the language translations preserve semantics, more or less, then lexically distinct terms can be considered as paraphrases. In our application of this idea, we translate the original English query/document into German, and then back-translate the German sentences into English. We then can detect pairs of words/phrases which can be considered as semantically related: the original English sentence and double-translated English sentence. We used Google translate[5], and chose German as the pivot language, which is a closely related to English, which we hope reduces the number of translation errors.

We performed double translation with 100 article laws in the Japanese Civil Code. We used the monolingual alignment tool of Sultan et al. [12] to create automatic word alignments in English. Table 4 shows examples of detected paraphrases using language translation. We can see that it also detects plural forms and past tense forms, in addition to words with similar meanings. We extract the top 100 paraphrases, and manually extracted corresponding Korean words in the Korean-translated Query-Article text.

Table 4. Examples of detected paraphrases

Original word	Paraphrased word	Original word	Paraphrased word
Year	Years	Establishes	Sets
Makes	Made	Purpose	Aim
Warranties	Guarantees	Matter	Area
Released	Relieved	Pledge	Commitment
Assigned	Transferred	Demand	Claim
Respect	Relation	Referred	Designated

[5] https://translate.google.com/.

3.6 Supervised Learning with SVM

Since we cannot anticipate the impact of each linguistic attribute, we use a machine learning algorithm that learns what information is relevant in the text to achieve our goal. We have compared our method with SVM, as a kind of supervised learning model. Using the SVM tool included in the Weka [4] software library[6], we performed cross-validation for the 412 questions. We used a linear kernel SVM because it is popular for real-time applications as they enjoy both faster training and classification speeds. Even though our system does not require much time for training, we chose a linear kernel to see the training performance for this simplest kernel. We used the following features:

(a) Word Lemma
(b) Lexical semantic features
(c) Negation feature
(d) Sentence analysis feature (condition, conclusion, and exception).

For concept features, we have exploited word embedding using *word2vec*. When we use word embedding, we assume the concepts of two words are the same if their cosine similarity in vector space is larger than 0.8.

The detailed features that we use are as follows:

Feature 1: $If\ \exists_{i,j}\{(\text{concept}(w_i)\ ,\ \text{Query}_{\text{condition}})\ \bigcap\ (\text{concept}(w_j)\ ,\ \text{Article}_{\text{condition}})\}$

Feature 2: $If\ \exists_{i,j}\{(\text{concept}(w_i)\ ,\ \text{Query}_{\text{conclusion}})\ \bigcap\ (\text{concept}(w_j)\ ,\ \text{Article}_{\text{conclusion}})\}$

Feature 3: $If\ \exists Article_{sub\_condition}\bigcap \exists_{i,j,k}\{(\text{concept}(w_i)\ ,\ \text{Query}_{\text{condition}})\ \bigcap\ (\text{concept}(w_j)\ ,\ \text{Article}_{\text{sub\_condition}k})=\varnothing\}$

Feature 4: $If\ \exists_{i,j}\{(\text{concept}(w_i)\ ,\ \text{Query}_{\text{condition}})\ \bigcap\ (\text{concept}(w_j)\ ,\ \text{Article}_{\text{exception\_condition}})\}$

Feature 5: $If\ \exists Article_{sub\_exception\_condition}\bigcap \exists_{i,j,k}\{(\text{concept}(w_i)\ ,\ \text{Query}_{\text{condition}})\ \bigcap\ (\text{concept}(w_j)\ ,\ \text{Article}_{\text{sub\_exception\_condition}k})=\varnothing\}$

Feature 6 : $If\ neg\_level(\text{Query}_{\text{condition}})\ =\ neg\_level(\text{Article}_{\text{condition}})$

Feature 7 : $If\ neg\_level(\text{Query}_{\text{conclusion}})\ =\ neg\_level(\text{Article}_{\text{conclusion}})$

Feature 8 : $If\ neg\_level(\text{Query}_{\text{condition}})\ =\ neg\_level(\text{Article}_{\text{exception\_condition}})$

Features 1 and 2 check if there are overlapping concepts between a query condition (conclusion) and its relevant article condition (conclusion). Feature 3 checks if there is an overlapping word between a query condition and its relevant article sub-condition. Because the article sub-condition is connected with other sub-condition(s), using "*and*" as a connector, the query should include the meanings of all the article sub-conditions. Feature 4 checks if there are overlapping concepts between a query condition and its article exception-condition. We want to check if the query is included in the exceptional case using the feature. Feature 5 confirms that there is no overlapping word between a query condition and its relevant article sub-exception-condition. Features 6,

[6] The SVM function in Weka is provided by libsvm https://www.csie.ntu.edu.tw/∼cjlin/libsvm/, and the linear kernal is from liblinear https://www.csie.ntu.edu.tw/∼cjlin/liblinear/.

7, and 8 check the negation levels between the query condition, article condition, query conclusion, article conclusion, and article exception-condition. The negation level (*neg_level(segment)*) is computed as following: if [negation + antonym] occurs an odd number of times in the segment, its negation level is 1. Otherwise if the [negation + antonym] occurs an even number of times, including zero, its negation level is 0.

4 Phase 2: Experimental Results

4.1 Comparison of Our System's Performance with Others

In the general formulation of the textual entailment problem, given an input text sentence and a hypothesis sentence, the task is to make predictions about whether or not the hypothesis is entailed by the input sentence. We report the accuracy of our method in answering yes/no questions of legal bar exams by predicting whether the questions are entailed by the relevant civil law articles.

There is a balanced positive-negative sample distribution in the dataset (51.70% yes, and 48.30% no) for a dry run of COLIEE 2016 dataset, so we consider the baseline for true/false evaluation is the accuracy when always returning "yes," which is 51.70%. Our total data for the dry run has 412 questions.

Table 5 shows the experimental results. An SVM-based model showed accuracy of 62.14% when we did not use word embedding but used the lexical form of each word; the method of Kim et al. [10] showed 60.92% and that of Kim et al. [11] showed 61.65%. Our SVM augmented system outperformed Kim et al. [10, 11]. The differences were significant using the Wilcoxon Signed Rank Test at the level of significance of 0.05. We guess the reasons that our current system shows better results than the previous systems [10, 11] are as follows: (1) we analyzed queries in more detail and detected multiple conditions such as "and/or" connections, and then performed entailment based on the "and/or" logics. (2) We did paraphrasing as term expansion.

Table 5 also shows the experimental results arising when we adjust some of the features in our method. For example, the accuracy was reduced by 1.70% when we removed paraphrasing, and the accuracy was reduced by 1.47% when we used word

Table 5. Experimental results on dry run data for Phase 2

Method	Accu. (%)
(a) Baseline	51.70
(b) Our method using cross-validation with Supervised learning (SVM) not using word embedding but using lexical word itself	62.14
(c) Our method using cross-validation with Supervised learning (SVM) using word embedding	60.67
(d) Cross-validation using Kim et al. [10]	60.92
(e) Cross-validation using Kim et al. [11]	61.65
Without term expansion using paraphrasing from (b)	60.44
Without neg_level() from (b)	49.27

embedding. This suggests that word embedding does not help capture the semantics better than the lexical word by itself. We can guess that it may be because of the small training data for *word2vec* training. This suggests that we need to construct a higher volume of legal text data for *word2vec* training, and then check the performance of word embedding. When we did not use the negation feature, the accuracy became lower by 12.87%, which demonstrates the importance of the negation feature.

Table 6 shows the experimental results on the COLIEE-2016 test data. The test data size is 70 queries for Phase 2 (extracted from the bar exam of 2015), and 95 queries for Phase 3 (extracted from the bar exam of 2014) which are the same with the test data for Phase 1. Our accuracy on test data is 55.71% for Phase 2, and 55.79% for Phase 3. As shown in Table 7, our system showed best performance when two phases are combined (Phase 3), even though our Phase 1 and Phase 2 systems were not the best in the COLIEE 2015 competition [16]. Our system also performed paraphrasing, and detected condition-conclusion-exceptions for the query/article; our system extracted the article segment for which the query is semantically related. In contrast to other systems (except for Carvalho et al. [17]) that recognized textual entailment from the whole article to the query, our system compared the approximate semantics from a specific article segment to the approximated semantics of the query.

Table 6. Experimental results on formal run data

Method	Accu. (%)
Phase 2 baseline when 'yes' labels are all chosen	52.86
Phase 2 system (entailment)	55.71
Phase 3 system (1) (TF-IDF and entailment)	46.32
Phase 3 system (2) (ranking SVM lemma and dependency bigram as features (a) and entailment)	54.74
Phase 3 system (3) (adding IR score as features into (a) and entailment)	55.79

4.2 Error Analysis

From unsuccessful instances, we manually classified the error types as shown in Table 8. The biggest error arises, of course, from the semantic similarity error, and we believe our word embedding is not sufficient for estimating semantic similarity. In the future, we will try to include the bar exam text in the training data for the word embedding. The second biggest error is because of complex constraints in conditions. As with the other error types, there are cases where a question is an example case of the corresponding article, and the corresponding article embeds another article. We also found cases that indicate the need to do more extensive temporal analysis.

It will be interesting if we compare our performance using Korean-translated sentences with that using original Japanese sentences. We would expect the system using original sentences to show improved performance, because there would be no translation errors. As future work, we will construct a Japanese system using paraphrase/synonym/antonym dictionaries for Japanese, and then analyze how the translation affects performance.

Table 7. IR+Entailment results (Phase 3) on the formal run data in the COLIEE-2016

Run	Accu.	Run	Accu.
JNLN1 [17]	0.4000	iLis7 [19]	0.5368
KIS-1 [18]	0.5158	JNLN3 [20]	0.4737
KIS-2 [18]	0.5158	Our system (1)	0.4632
KIS-3 [18]	0.5263	Our system (2)	0.5474
KIS-4 [18]	0.5263	Our system (3)	**0.5579**

Table 8. Error types

Error type	Accuracy (%)	Error type	Accu. (%)
Specific example case	9.62	Semantic similarity error	28.85
Incorrect detection of the most similar article sentence	10.90	Constraints in condition	25.00
Incorrect detection of condition, conclusion, and mismatch	11.54	Etc.	14.10

5 Related Work

A previous textual entailment method from Bdour and Gharaibeh [5] provided the basis for a yes/no Arabic question answering system. They used a kind of logical representation, and compared the logical representation between queries and documents. This method may be appropriate for the task where queries and documents have similar logical representations so it is easier to confirm entailment from one logical representation to another. However, our task's entailment type is more complex, so we take an approach that approximates the logical content of queries and documents, rather than attempt any complete transformation to a logical form.

Nielsen et al. [6] extracted features from dependency paths, and combined them with word-alignment features in a mixture of an expert-based classifier. Zanzotto et al. [7] proposed a syntactic cross-pair similarity measure for RTE. Harmeling [8] took a similar classification-based approach with transformation sequence features. Marsi et al. [9] described a system using dependency-based paraphrasing techniques.

Many methods have been proposed for paraphrasing. One of the methods is the idea of semantic parsing via paraphrasing [13]. They transform a sentence into a logical form, and then convert logical forms to canonical form using the Freebase database. Subsequently, they obtain an association between the original sentence and canonical forms. However, hundreds of logical/canonical forms have been generated per sentence in their method, and the method does not show how to choose the best amongst them.

The method of Zhang et al. [14] also uses a pivot language for paraphrasing. Like us, they translate one language to another, then re-translate from the translated language into the original language. They then obtain a paraphrasing set between the original utterance and double-translated utterance. They showed improved performance in paraphrase detection using the pivot language translation, so we also employ the

language translation-based paraphrasing. But instead of their use of the GIZA++ alignment, we used the monolingual alignment tool of Sultan et al. [12], because GIZA++, which is for alignment between two different languages, did not show good performance for our dataset.

6 Conclusion

We have described our most recent implementation for the Competition on Legal Information Extraction/Entailment (COLIEE)-2016 Task.

For Phase 1, legal information retrieval, we implemented a Ranking-SVM model for the legal information retrieval task. By incorporating features such as lexical words, dependency links, and tf-idf score, our model shows better mean average precision than tf-idf.

For Phase 2, we have proposed a method to answer yes/no questions from legal bar exams related to civil law. We used an SVM model using paraphrasing and pre-trained word embedding and query/article condition/conclusion/exception analysis. We show improved performance over a previous system, and paraphrasing and negation detection contributed to the performance. In the COLIEE 2016 competition, our system combining the Phase 1 and Phase 2 ranked highest in the accuracy of answering yes/no questions. As future work, we will train word2vec by larger texts not by articles to get the benefit of word embedding, and also try different kernels for SVM training to check if the kernel selection can increase the entailment performance.

Acknowledgements. This research was supported by the Alberta Machine Intelligence Institute (www.amii.ca). We are indebted to Ken Satoh of the National Institute for Informatics, who had the vision to create the COLIEE competition.

References

1. Jones, K.S.: A statistical interpretation of term specicity and its application in retrieval. In: Willett, P. (ed.) Document Retrieval Systems, pp. 132–142. Taylor Graham Publishing, London (1988)
2. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 133–142. ACM, New York (2002)
3. Maxwell, K.T., Oberlander, J., Croft, W.B.: Feature-based selection of dependency paths in ad hoc information retrieval. In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics, (vol. 1: Long Papers), pp. 507–516. Association for Computational Linguistics, Sofia, August 2013
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
5. Bdour, W.N., Gharaibeh, N.K.: Development of yes/no Arabic question answering system. Int. J. Artif. Intell. Appl. **4**(1), 51–63 (2013)
6. Nielsen, R.D., Ward, W., Martin, J.H.: Toward dependency path based entailment. In: Proceedings of 2nd PASCAL Challenges Workshop on RTE (2006)

7. Zanzotto, F.M., Moschitti, A., Pennacchiotti, M., Pazienza, M.T.: Learning textual entailment from examples. In: Proceedings of 2nd PASCAL Challenges Workshop on RTE (2006)
8. Harmeling, S.: An extensible probabilistic transformation-based approach to the third recognizing textual entailment challenge. In: Proceedings of ACL PASCAL Workshop on Textual Entailment and Paraphrasing (2007)
9. Marsi, E., Krahmer, E., Bosma, W.: Dependency-based paraphrasing for recognizing textual entailment. In: Proceedings of ACL PASCAL Workshop on Textual Entailment and Paraphrasing (2007)
10. Kim, M.-Y., Xu, Y., Goebel, R.: Alberta-KXG: legal question answering using ranking SVM and syntactic/semantic similarity. In: 8th International Workshop on Juris-Informatics (JURISIN), 2014
11. Kim, M.-Y., Xu, Y., Goebel, R.: A convolutional neural network in legal question answering. In: JURISIN Workshop (2015)
12. Sultan, M.A., Bethard, S., Sumner, T.: Back to basics for monolingual alignment: exploiting word similarity and contextual evidence. Trans. Assoc. Comput. Linguist. **2**, 219–230 (2014)
13. Berant, J., Percy, L.: Semantic parsing via paraphrasing. In: Proceedings of Conference of the Association for Computational Linguistics (ACL), pp. 1415–1425 (2014)
14. Zhang, W., Ming, Z., Zhang, Y., Liu, T., Chua, T.S.: Exploring key concept paraphrasing based on pivot language translation for question retrieval. In: AAAI, pp. 410–416 (2015)
15. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60 (2014)
16. Kim, M.-Y., Goebel, R., Kano, Y., Satoh, K.: COLIEE-2016: evaluation of the competition on legal information extraction and entailment. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016)
17. Carvalho, D.S., Tran, V.D., Tran, K.V., Lai, V.D., Nguyen, M.-L.: Lexical to discourse-level corpus modeling for legal question answering. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: JNLN1)
18. Taniguchi, R., Kano, Y.: Legal yes/no question answering system using case-role analysis. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: KIS)
19. Kim, K., Heo, S., Jung, S., Hong, K., Rhim, Y.-Y.: An ensemble based legal information retrieval and entailment system. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: iLis7)
20. Do, P.-K., Nguyen, H.-T., Tran, C.-X., Nguyen, M.-T., Minh, N.L.: Legal question answering using ranking SVM and deep convolutional neural network. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: JNLN3)
21. Onodera, D., Yoshioka, M.: Civil code article information retrieval system based on legal terminology and civil code article structure. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: HUKB)
22. Nguyen, T.-S., Phan, V.-A., Nguyen, T.-H., Trieu, H.-L., Chau, N.-P., Pham, T.-T., Nguyen, L.-M.: Legal information extraction/entailment using SVM-ranking and tree-based convolutional neural network. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: JNLN2)
23. John, A.K., Di Caro, L., Boella, G., Bartolini, C.: Team-normas' participation at the COLIEE 2016 bar legal exam competition. In: Tenth International Workshop on Juris-Informatics (JURISIN) (2016). (Submission ID: N01)