# 法律文書処理の国際コンテスト型ワークショップ COLIEE2018 の開催

狩野芳伸[1] 吉岡真治[2] Mi-Young Kim[3] Yao Lu[3] Juliano Rabelo[3] 清田直希[1] Randy Goebel[3] 佐藤健[4]
[1]静岡大学 [2]北海道大学 [3]University of Alberta [4]国立情報学研究所

**概要**：法律文書の処理は、その直接的応用が期待される実用的なタスクであると同時に、構文、意味役割、論理など高度かつ多様な解析技術を必要とする挑戦的なタスクである。我々は、法律文書処理の評価タスクとして、司法試験の自動解答を題材に数年にわたり毎年コンテスト型国際ワークショップ COLIEE（Competition for Legal Information Extraction and Entailment）を開催してきた。COLIEE では、多択式の民法短答式試験を二値の正誤問題として扱い、民法の条文を知識源として、解答に必要な条文を推測する情報検索タスクと、正誤を答える含意関係タスクを設定している。また原文の日本語に加え、英語に翻訳したものも提供し、各国から多数のチームが参加してきた。2018 年に開催した COLIEE 2018 では、成文法である我が国の司法試験に加え、判例法であるカナダの判決集を用いたタスクを新たに追加した。本稿では前者を中心に、COLIEE 2018 の概要、成果、参加システムと問題の分析を報告する。

## 1. はじめに

法律文書の処理は、自然言語処理の応用として大きな期待の寄せられている分野の一つである。法律文書の処理には、単に分野適応が必要というだけでなく、構文、意味役割、論理など高度かつ多様な解析技術を必要とする複合的な要素がある。技術の達成度を測り研究を促進するため、定量的かつ客観的な評価が必要である。

我々は我が国の司法試験をそのような評価として採用し、コンテスト型ワークショップ COLIEE（Competition for Legal Information Extraction and Entailment）を数年にわたり開催してきた。司法試験は我が国において弁護士・検察官・裁判官になるために合格が必須の毎年行われる国家試験であり、多択式の問題である短答式試験と、文章で答える論文式試験からなる。科目として主に憲法・民法・刑法の三種類がある。COLIEE ではこのうち民法短答式試験を対象とした。さらに、多択式を二択にブレークダウンするとともに、解答に関連する条文を検索する情報検索タスク（Information Retrieval）と、二択を Yes/No で答える含意関係タスク(Entailment)を用意した。また知識源として民法の全条文を配布し、問題文共々原文の日本語に加え英語に翻訳したものを配布している。COLIEE 2018 ではこれらのタスクに加え、新たにカナダ連邦裁判所の判例データを用いたタスクを追加した。本稿では、COLIEE 2018 の概要、結果と分析を報告する。

## 2. COLIEE 2018 の概要

COLIEE は、COLIEE 2014、COLIEE 2015 [1], COLIEE 2016 [2]、COLIEE 2018[3]と人工知能学会国際シンポジア（JSAI-isAI）のひとつである JURISIN（法情報学ワークショップ）において開催してきた。COLIEE 2017[4]は法と人工知能に関するトップカンファレンスである ICAIL（International Conference of Artificial Intelligence and Law）において開催し、COLIEE 2019 も ICAIL のワークショップとして開催予定である

### 2.1 COLIEE Case Law (Task1, 2)

COLIEE 2018 から新たに追加された COLIEE Case Law タスクでは、vLex Canada 社から提供されたカナダ連邦裁判所の判例データを用いた。カナダは判例法(Case Law)を用いている。判例データは、各判決についてどの判例(noticed case)を用いたかがアノテーションされている。情報検索を行う Task 1 は、200 件の判例の中から、与えられた判決に対して必要とされた複数の判例を答える設定とした。含意関係検索を行う Task 2 は、判決と必要な判例を既知として、判例内のどの段落が判決を含意するかを答える設定とした。

### 2.2 COLIEE Statute Law (Task3, 4)

我が国では成文法（Statute Law）を用いている。司法試験から検索された問題文に対し、民法条文から（場合によって複数の）関連する条文を返す情報検索タスク（Task 3)、関連条文を与えずに Yes/No の二値解答を行う質問応答タスク（Task 4）からなる。

我が国の司法試験民法短答式試験から作られたタスクオーガナイザー配布のデータは、訓練（および開発）データおよびテストデータからなる。いずれも原文の日本語版と、英訳の双方を提供している。COLIEE では毎年、最新の年度の司法試験問題からテストデータを作成し、前年のテストデータを含む過去の問題に正答を付して訓練（および、うち最新年度を開発）データとしている。

```
<pair label="Y" id="H18-2-2">
<t1>
（緊急事務管理）
第六百九十八条　管理者は、本人の身体、名誉又は財産に対する急迫の危害を免れさせるために事務管理をしたときは、悪意又は重大な過失があるのでなければ、これによって生じた損害を賠償する責任を負わない。
</t1>
<t2>
車にひかれそうになった人を突き飛ばして助けたが，その人の高価な着物が汚損した場合，着物について損害賠償をする必要はない。
</t2>
</pair>
```

図 1　COLIEE 含意関係タスク問題の例

もともとの問題は多択式であるため、多択解答を指示する部分を削除したうえで、これを二択に展開している。

<pair>タグは問題一つ分に相当し、訓練データの場合は label 属性に正答が付与される。Id 属性は年度を含む一意な問題 ID である。<t1>が民法条文、<t2>が問題文に相当する。図 1 に例を示す。

COLIEE 2018 訓練データの問題総数は 651 であった。テストデータは2017年に相当し､問題数は69であった。

タスクオーガナイザー配布の知識源は全 1044 条からなる民法の条文である。知識源は原文の日本語版と、英訳の双方を提供している。

### 2.3 参加チームと手法概要

各国から結果提出があった。Case Law（カナダ判例）タスクには 13 チームの参加があり、Task 1 は 6 提出（12 run）、Task 2 は 4 提出（8 run）であった。Statute Law（司法試験）は Task 3 に 8 チーム（17 run）、Task 4 に 3 チーム（7 run）の提出があった。

Task 1 では多くの参加者が機械学習を用いているが、その中でも語彙的特徴量とサマリ属性の潜在特徴量を組み合わせた手法が最も良い結果を得ている。

Task 2 で最も良い結果を得たシステムは、類似度に基づく特徴量を Random Forest で処理したものであった。

Task 3 で最も良い結果であったシステムは、クエリを構成するのに適切なキーワードを選ぶのに tag cloud アルゴリズムを用い、最終的な結果を出力するのに Terrier IR プラットフォームを用いたものであった。

| Run | Prec. | Recall | F1 |
|---|---|---|---|
| Baseline | 0.2649 | 0.4102 | 0.3219 |
| HUKB1 | 0.4974 | 0.3084 | 0.3808 |
| HUKB2 | 0.4047 | 0.3037 | 0.3470 |
| JNLP-r=2.5 | 0.5464 | 0.6550 | 0.5958 |
| JNLP-k=10 | 0.6763 | 0.6343 | 0.6546 |
| Smartlaw | 0.2871 | 0.4308 | 0.3446 |
| UA | 0.3725 | 0.3227 | 0.3458 |
| UA-postproc | 0.3484 | 0.4038 | 0.3741 |
| UA-smote | 0.3539 | 0.3927 | 0.3723 |
| UBIRLED-1 | 0.1329 | 0.6232 | 0.2191 |
| UBIRLED-2 | 0.1955 | 0.7202 | 0.3075 |
| UBIRLED-3 | 0.5614 | 0.1017 | 0.1723 |
| UL | 0.5638 | 0.3021 | 0.3934 |

表 1　Task 1(カナダ判例，情報検索)の評価結果

| Run | Prec. | Recall | F1 |
|---|---|---|---|
| Baseline | 0.0405 | 0.5094 | 0.0751 |
| Smartlaw | 0.0465 | 0.1509 | 0.0711 |
| UA | 0.2381 | 0.2830 | 0.2586 |
| UA-100 | 0.1905 | 0.2264 | 0.2069 |
| UA-500 | 0.2381 | 0.2830 | 0.2586 |
| UBIRLED-1 | 0.0484 | 0.8302 | 0.0914 |
| UBIRLED-1 | 0.0495 | 0.9245 | 0.0940 |
| UBIRLED-1 | 0.0467 | 0.7925 | 0.0881 |
| UNCC0 | 0.0330 | 0.0566 | 0.0417 |

表 2　Task 2(カナダ判例，含意関係)の評価結果

| run id | L | 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 | 0.628 | 0.7381 | 0.7528 | 0.809 | 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 | 0.413 | **0.7126** | 0.7398 | 0.764 | 0.8202 | 0.9213 |
| Smartlaw | E | 138 | 57 | 0.6042 | 0.413 | 0.7005 | 0.7036 | 0.7079 | 0.764 | 0.8315 |
| JNLP2 | E | 138 | 56 | 0.5997 | 0.4058 | 0.6981 | 0.7296 | 0.7528 | 0.809 | 0.9101 |
| SPABS_bm25 | E | 138 | 55 | 0.5821 | 0.3986 | 0.6739 | 0.707 | 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 | 0.47 | 0.4494 | 0.4607 | 0.5056 |
| UB1 | E | 69 | 31 | 0.4171 | 0.4493 | 0.413 | 0.5355 | 0.573 | 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 | 0.215 | 0.1377 | 0.2536 | 0.2638 | 0.3371 | 0.4494 | 0.573 |
| 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 |

表 3　Task 3(司法試験，情報検索)の評価結果

L はデータの言語、ret は出力条文数、rel は正しい出力条文数を表す

| Team | Language | Correct Answers (69 questions in total) | Accuracy |
|---|---|---|---|
| BaseLine | N/A | 35 (answers No to all) | 0.5072 |
| UA | ? | 44 | 0.6377 |
| KIS_Frame | Japanese | 39 | 0.5652 |
| KIS_mo3 | Japanese | 38 | 0.5507 |
| KIS_dict | Japanese | 37 | 0.5362 |
| KIS_SVM | Japanese | 36 | 0.5217 |
| KIS_Frame2 | Japanese | 35 | 0.5072 |
| UE | English | 33 | 0.4783 |

表 4　Task 4(司法試験, Y/N 質問応答)の評価結果

Task 4 で最も良い結果を得たシステムは、条件部・結論部・例外部を検出するルールと、辞書ベースの否定判断を行い、内部的に機械翻訳と韓国語の解析器をもちいたものであった。

さらに詳細は JUSIRIN 2018 の Proceedings にある各チームの論文を参照されたい。

### 2.4 評価結果

評価メトリクスとして、Task 1/2 については precision（適合率）、recall（再現率）、F1-measure（F 値）を、Task 4 については accuracy（正解率）を算出している。Task 3 については、Precision、Recall に加え、全ての正解文書について、見付けた時点での精度の平均をとったものの課題毎の平均である AP（Average Precision）、上位 n 件目までの結果を検索結果としたときの再現率(見付けた正解文書/ 全ての正解文書)である $R_n$（Recall@n）を計算した。表 1〜4 に各タスクの評価結果を示す。

## 3. 分析と議論

### 3.1 カナダ判例タスク（Case Law, Task 1/2）

Task 2 は Task 1 よりはるかに難しく、人間でも適切な含意関係から正しい段落を選ぶのは困難である。Task 1/2 は新規タスクのため、Task 2 の評価手法、アンバランスデータの対処、機械学習には不十分と思われるデータサイズの不足など今後さまざまに改善の余地がある。

### 3.2 司法試験情報検索タスク（Statute Law, Task 3）

情報検索タスクでは、関連すると判断した少数の条文を返す結果と順位付きで 100 位までの関連条文を返す結果の二種類の結果提出を行い、前者の結果に対する精度と再現率ならびに、再現率に重きを置いた調和平均である F2 尺度を利用した評価を行った。また、後者の結果に対して、MAP（Mean Average Precision:平均精度(正解文書を発見した時の精度の平均)の課題ごとの平均）、$R_n$（Recall@n：n 位までの文書を検索結果としたときの再現率）によりシステムの評価を行った（表 3）。

ただし、課題ごとの難しさの議論をするためには、上位で見つけることも困難なのかどうかという観点を考慮するため、順位付きのリストを用いて分析を行った（図 2,3）。各棒グラフの長さは、全ての参加者の提出結果に対する各課題ごとの結果の平均である。基本的に、左側が AP（Average Precision:正解文書を見つけた時の精度の平均）が大きい（簡単な）課題である。図 3 については、最初の 2 問が 3 つ関連条文を持ち、残りは、関連条文が 2 つものである。

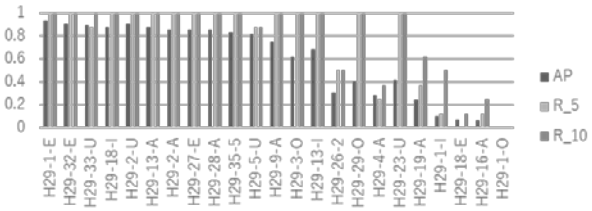


図 2. 正解文書が 1 つの質問に対する AP,$R_5$,$R_{10}$ の参加者平均

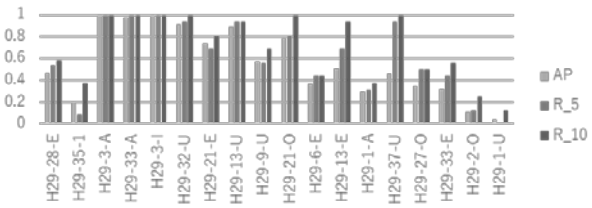


図 3 正解文書が複数の質問に対する精度、再現率、AP,$R_5$,$R_{10}$ の参加者平均

問題が条文に似ているような簡単な問題については、今回の法律の問題にチューニングしていないシステムでも、簡単に見付けられる一方で、ほとんどのシステムが正解条文を上位にランキングできないような、とても難しい問題があることが確認できた。このような問題では、用いられている単語が実際の事例で用いられる具体的な文章を条文で用いられている概念に対応づける必要がある。しかし、このような難しい問題は数が少なく、提出されたシステムは、簡単な問題を確実に解く方が性能が向上することもあり、これらの難しい問題については、まだ考察が不十分のように思われる。

### 3.3 司法試験質問応答タスク（Statute Law, Task 4）

質問応答タスクでの最もよい評価値は 63 ポイントで。質問応答タスクは二値分類であり、ランダムで 50 ポイントが見込める一方、問題の難しさを考えると一見かなりの性能を達成しているようにも見える。しかし詳細に分析すると、必ずしも十分な性能を達成しているかは不明であることがわかる。

まず、テストデータで数十という数は安定した評価には不足である。各チームから年度による評価値の揺れが大きく、数ポイント以上はあることが報告されている。実際、昨年度の最高評価値は同一チームの同一システムで 70 ポイントを超えていた。

訓練データで数百という数も、end-to-end の教師付き機械学習のみでシステムを実装するには全く不十分である。そうした手法をとって仮に成功したのであれば、高々数百程度の過去問とよく似た問題ばかりが出題されたか、表層的な次元数の低い特徴量で処理できたということに

なるが、実際の問題を見ればそのように単純な手法だけで解答可能とは考えられない。

改めて図 2 の例を用いて分析を試みる。問題文には、A「車にひかれそうになった人」を B「突き飛ばして」C「助けた」とあるが、これが条文の A「急迫の危害」を B「免れさせるため」に C「事務管理をした」ときは、D「悪意又は重大な過失があるのでなければ」にあたるはずである。A、B いずれも、条文レベルの抽象的表現と具体的な表現との包含関係を判定する必要がある。C に至っては、一般的な感覚では何が事務管理なのか不明である。D はおそらく「助けた」ということから判断ができようが、いずれも対応しうる表現のバリエーションは膨大なことが想像されるうえ、文脈によって異なる判断が必要なこともあろう。

一方で、一部の問題は非常に解答が容易であることがわかっている。そうした問題は条文のごく一部を改変した文章が問題文となっており、表層レベルの処理でもある程度の確率で解答可能であると考えられる。

技術的に困難な問題が多数であることは、研究のベンチマークとしては適している。現実の法律文書処理に適用するとしても、背後にある構文構造、人物の役割と関係性、論理、抽象性など多様で複雑な構造を的確に処理する必要がある。テストデータを対象に、解答に必要な要素を手作業で分類した結果を表 5 に示す。

| チーム毎の正答数と正解率 / 問題分類 | 該当問題数 | UA | accuracy | UE | accuracy | KIS_mo3 | accuracy | KIS_Frame | accuracy |
|---|---|---|---|---|---|---|---|---|---|
| 箇条書き | 3 | 1 | 0.33 | 2 | 0.67 | 1 | 0.33 | 2 | 0.67 |
| 番号 | 3 | 2 | 0.67 | 2 | 0.67 | 1 | 0.33 | 2 | 0.67 |
| 含意 | 5 | 2 | 0.4 | 2 | 0.4 | 1 | 0.2 | 2 | 0.4 |
| 係り受け | 5 | 3 | 0.6 | 1 | 0.2 | 2 | 0.4 | 4 | 0.8 |
| 条文検索 | 5 | 3 | 0.6 | 2 | 0.4 | 3 | 0.6 | 4 | 0.8 |
| 言い換え | 5 | 2 | 0.4 | 4 | 0.8 | 3 | 0.6 | 3 | 0.6 |
| 否定 | 7 | 5 | 0.71 | 3 | 0.43 | 5 | 0.71 | 7 | 1 |
| 法律用語 | 7 | 4 | 0.57 | 2 | 0.29 | 2 | 0.29 | 3 | 0.43 |
| 一般用語 | 9 | 5 | 0.56 | 5 | 0.56 | 4 | 0.44 | 4 | 0.44 |
| 述語項 | 9 | 8 | 0.89 | 3 | 0.33 | 5 | 0.56 | 5 | 0.56 |
| 動詞言い換え | 13 | 7 | 0.54 | 6 | 0.46 | 7 | 0.54 | 4 | 0.31 |
| 格役割 | 15 | 8 | 0.53 | 6 | 0.4 | 9 | 0.6 | 6 | 0.4 |
| 曖昧性解消 | 17 | 9 | 0.53 | 7 | 0.41 | 8 | 0.47 | 9 | 0.53 |
| 照応 | 20 | 13 | 0.65 | 5 | 0.25 | 12 | 0.6 | 13 | 0.65 |
| 形態素 | 25 | 18 | 0.72 | 16 | 0.64 | 20 | 0.8 | 16 | 0.64 |
| 人物関係 | 26 | 14 | 0.54 | 11 | 0.42 | 13 | 0.5 | 10 | 0.38 |
| 人物役割 | 27 | 16 | 0.59 | 12 | 0.44 | 14 | 0.52 | 13 | 0.48 |
| 条件 | 31 | 19 | 0.61 | 9 | 0.29 | 13 | 0.42 | 16 | 0.52 |

表 5 テストデータの問題タイプ分類

## 4. おわりに

現実の法律分野の課題に対応するには、表層的な処理だけでは不足である。数年にわたり開催してきた、司法試験の自動解答を試みる COLIEE タスクは、今年度よりさらに判例法であるカナダ連邦裁判所のデータを用いたタスクを追加し、情報検索・含意関係・質問応答の各タスクにおいて、現在不足している技術要素を浮き彫りにするベンチマークとして技術発展と議論を促進する役割を果たしている。我々はさらに現実的な応用タスクとして、科学研究費助成金による「裁判過程における人工知能による高次推論支援」プロジェクトを開始した。法的推論、機械学習、議論学、法学の各分野の専門家に加え法曹界の実務家を交えて構成しており、裁判過程の自動化を研究するものである。自然言語処理の研究としては、民法あるいは刑法など分野を限定することで、必要な語彙数がある程度少数ですむと期待される。COLIEE とともに、より本質的、かつ複雑な構造を解析する糸口になるのではないか。今後も COLIEE タスクを継続しつつデータの増加や評価手法の改善、さらに現実的な法律文書処理タスクへの応用を行い、法律処理にとどまらない基盤的な言語処理技術の発展に貢献していきたい

**謝辞** 本研究の一部は、JST CREST および文部科学省科学研究費助成金の補助を受けたものである。

## 参考文献

[1] M.-Y. Kim, R. Goebel, and S. Ken, “COLIEE-2015: Evaluation of Legal Question Answering,” in *Ninth International Workshop on Juris-informatics (JURISIN 2015)*, 2015.

[2] M.-Y. Kim, R. Goebel, Y. Kano, and K. Satoh, “COLIEE-2016: Evaluation of the Competition on Legal Information Extraction/Entailment,” in *Tenth International Workshop on Juris-informatics (JURISIN 2016)*, 2016.

[3] M. Yoshioka, Y. Kano, N. Kiyota, and K. Satoh, “‘Overview of Japanese Statute Law Retrieval and Entailment Task at COLIEE-2018,’” in *Twelfth International Workshop on Juris-informatics (JURISIN 2018)*, 2018.

[4] Y. Kano, M.-Y. Kim, R. Goebel, and K. Satoh, “Overview of coliee 2017,” *Epic Ser. Comput.*, vol. 47, pp. 1–8, 2017.