Vol.2017-NL-234 No.2
2017/12/20

司法試験自動解答の国際コンテスト型ワークショップ COLIEE と法律文書処理

狩野芳伸[†1]　佐藤健[†2]

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抽出タスクと、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。本稿ではこれまでの COLIEE の概要、成果、参加システムと問題の分析を報告する。また、新規科研費プロジェクト「裁判過程における人工知能による高次推論支援」についても紹介し、法律文書処理の今後の展望を議論する。

キーワード：COLIEE、司法試験、法律文書処理

COLIEE International Workshop for Solving Legal Bar Exam and Legal Document Processing

YOSHINOBU KANO[†1]　KEN SATOH[†2]

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抽出する情報抽出タスク（Information Extraction）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。

本稿では、これまでの COLIEE の概要を報告するとともに、関連して今後の法律文書処理の研究計画についても紹介する。

2. COLIEE の概要

COLIEE は、COLIEE 2014、COLIEE 2015 [1], COLIEE 2016 [2]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[3]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）において開催した。

2.1 タスク設定

COLIEE は与えられた問題文に対し、民法条文から関連する条文を返す情報抽出タスク（Phase 1）、問題文と与えられた関連条文について二値の判定をする含意関係タスク（Phase 2）、関連条文を与えずにこの二つを同時に行う質問応答タスク（Phase 3）の三つのサブタスクからなる。COLIEE 2017 においては、時期の問題から Phase 2 を実施せず、Phase 1 を Task 1、Phase 3 を Task 2 と呼称して実施した。以下で各タスクを定式化する。

情報抽出タスクは、与えられた問題文 Q を回答するのに関連する民法条文のサブセット S1, S2,..., Sn を民法全体から抽出するタスクとなる。すなわち、解答すべき条文は複数のこともある。

含意関係タスクは、与えられた問題文 Q を、同時に与え

* †1 静岡大学 情報学部
Faculty of Informatics, Shizuoka University
†2 国立情報学研究所
National Institute of Informatics

られた民法条文のサブセット S1, S2,..., Sn が含意するか、すなわち Entails(S1, S2, ..., Sn , Q)または Entails(S1, S2, ..., Sn , not Q).を回答するタスクである。

質問応答タスクは、与えられた問題文 Q について、民法全体を用いて Yes あるいは No の二値から解答するタスクである。

2.2 知識源

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。図 1 にその冒頭部を示す。

民法（第一編第二編第三編）
第一編　総則
第一章　通則
（基本原則）
第一条　私権は、公共の福祉に適合しなければならない。
２　権利の行使及び義務の履行は、信義に従い誠実に行わなければならない。
３　権利の濫用は、これを許さない。
（解釈の基準）
第二条　この法律は、個人の尊厳と両性の本質的平等を旨として、解釈しなければならない。
第二章　人
第一節　権利能力
第三条　私権の享有は、出生に始まる。
２　外国人は、法令又は条約の規定により禁止される場合を除き、私権を享有する。
第二節　行為能力
…

図 1　民法の冒頭部

条文は階層的に記述されており、編・章・節・款・目のレベルで区分されている。このすべてが常に定義されているわけではない。条はそれらで区分される基本単位であり、さらに項、項の下位にさらに号をもつことがある。第一項の番号は省略されている。条に対して、その前に条の内容を簡潔に説明する（）でくくられた見出しがつくことがある。これらが規則的に改行および全角空白区切りで記述されているため、機械的に構造を判断できる。

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　タスクデータセット

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。

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。COLIEE 2017 の時点で、訓練データは 2006 年から 2015 年に相当する 10 個の XML ファイルからなり、各年度が 1 ファイルに収められている。訓練データ中の問題総数は 570 であった。テストデータは 2016 年に相当し、問題数は 78 であった。

XML ファイルの形式は、NTCIR RITEVAL[4]タスクで用いられた形式に準拠している a。図 2 に例を示す。

```
<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 2　COLIEE 含意関係タスク問題の例

<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。

2.4　評価手法

評価メトリクスとしては情報抽出・含意関係・質問応答いずれのタスクについても、precision（適合率）、recall（再現率）、F-measure（F 値）、accuracy（正解率）を算出している。情報抽出については、複数回答がある場合はそれぞれを別個にカウントして算出した。

2.5 COLIEE 2017 の結果

本節では、最新の結果である COLIEE 2017 の参加チームとその評価について述べる。COLIEE 2017 では各国から 10 チームの結果提出があり、情報抽出タスクに 9 チーム（17 セット）、質問応答タスクに 8 チーム（20 セット）の提出があった。表 1 に情報抽出タスクの参加チームとアプローチ、表 2 に質問応答タスクの参加チームとアプローチ、表 3 に情報抽出タスクの評価結果、表 4 に質問応答タスクの評価結果を示す。

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

表 1　COLIEE 2017 情報抽出タスク参加チーム

チーム ID	アプローチ
HUKB [5]	article structure analysis, phrase matching, rank SVM with 15 similarity scores and alignment scores, ensemble
iLis7-1 [6]	TF-IDF, LSM, LDA, Word2Vec, LSA
JAISTNLP [7]	TF-IDF
JNLP [8]	ranking related n-gram collections, term order probabilities, relevance disambiguation.
NOR [9]	LDA
UA [10]	TF-IDF, language model

表 2　COLIEE 2017 質問応答タスク参加チーム

チーム ID	アプローチ
iLis9 [11]	TF-IDF, negation detection
JAISTNLP [7]	ranking, encoding-based and attention neural network
KIS [12]	case-role based linguistic analysis, predicate-argument structures
NAIST [13]	Word2vec, attention neural network
NOR [9]	CNN, LSTM
UA [10]	Korean dependency parser, Excite Japanese/Korean machine translation, semantic dictionary, k-means clustering

表 3　COLIEE 2017 情報抽出タスクの評価結果

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

表 4　COLIEE 2017 質問応答タスクの評価結果

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

2.6 分析と議論

情報抽出タスクでも質問応答タスクでも、最もよい評価値で 60～70 ポイント程度となっている。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。

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るか、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。

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。

3. 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、情報抽出・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。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。COLIEE オーガナイザーである University of Alberta の Randy Goeble, Mi-Young Kim の両氏に深謝申し上げる。

参考文献

[1] Chang, C. L. and Lee, R. C. T.. Symbolic Logic and Mechanical Theorem Proving. Academic Press, 1973, 331p.

[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] Y. Kano, M.-Y. Kim, R. Goebel, and K. Satoh, “Overview of COLIEE 2017,” in *The 16th International Conference on Artificial Intelligence (ICAIL 2017)*, 2017.

[4] S. Matsuyoshi, Y. Miyao, T. Shibata, C.-J. Lin, C.-W. Shih, Y. Watanabe, and T. Mitamura, “Overview of the NTCIR-11 Recognizing Inference in TExt and Validation (RITE-VAL) Task,” in *the 11th NTCIR (NII Testbeds and Community for information access Research) workshop*, 2014, pp. 223–232.

[5] M. Yoshioka and D. Onodera, “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)*, 2017.

[6] S. Heo, K. Hong, and Y.-Y. Rhim, “Legal Content Fusion for Legal Information Retrieval,” in *the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017)*, 2017.

[7] T.-S. Nguyen, V.-A. Phan, and L.-M. Nguyen, “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)*, 2017.

[8] D. S. Carvalho, V. Tran, K. Van Tran, and L. M. Nguyen, “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)*, 2017.

[9] R. Nanda, A. K. John, L. Di Caro, G. Boella, and L. Robaldo, “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)*, 2017.

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

[11] B. Jung, C. Soh, K. Hong, S. Lim, and Y.-Y. Rhim, “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)*, 2017.

[12] Y. Kano, R. Hoshino, and R. Taniguchi, “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)*, 2017.

[13] A. Morimoto, D. Kubo, M. Sato, H. Shindo, and Y. Matsumoto, “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)*, 2017.