# 低資源な法ドメイン含意タスクにおけるデータ拡張

伊藤 光一　山田 寛章　徳永 健伸
東京工業大学 情報理工学院
{ito.k.bo@m, yamada@c, take@c}.titech.ac.jp

## 概要

法ドメインではアノテーションが高コストのため学習データが不足する問題がある．本稿では，COLIEE TASK 4 を用いて，ラベル付き学習データのルールベースによる拡張と，言語モデル事前学習の際の擬似的な学習データ拡張の効果検証を行う．実験の結果，提案手法である反対解釈によるデータ拡張手法が最良の性能を示した．

## 1　はじめに

事前学習済み言語モデル BERT[1] により，自然言語処理分野は大いに進歩した．BERT をファインチューニングすることで，様々なベンチマークタスクにおいて，従来の手法を超える性能が報告されている．法分野における含意関係認識タスクでも，BERT を使用したモデルが最高スコアを達成している [2]．しかし，BERT のような事前学習済み言語モデルを特定タスクにファインチューニングするには，一定の規模の学習データが必要となる．これは，ラベル付き言語資源の少ないドメインでは課題となっており，法分野も例外ではない．本稿では，日本語法分野含意関係認識タスクを対象として，法律文を用いたルールベースのデータ拡張手法の提案と検証を行う．

また，法分野などのドメインに特化した事前学習済み言語モデルの入手が困難な場合，Wikipedia 等で学習された既存モデルをドメイン固有のデータで追加事前学習を行う必要がある．法分野の場合，ラベルなしデータであっても入手が困難であることから，ファインチューニングによるドメイン適合にもデータ拡張が必要となる．このため，本稿では，追加事前学習の際に対照学習を用いた擬似的なデータ拡張手法の提案と検証も行う．

## 2　対象タスク：COLIEE TASK4

本稿では COLIEE (Competition for Legal Information Extraction and Entailment) 2021[2] の TASK4 を対象タスクとする．COLIEE 2021 は，法と人工知能に関するトップカンファレンス (ICAIL) において開催された法律情報の抽出と含意に関するコンペティションである．このうち TASK4 では，日本の司法試験の民法選択式問題を含意タスクと見なしている．具体的には，司法試験の問題文が前件として，関連する民法条文が仮説として対となり入力として与えられており，民法条文が問題文を含意するか否かを解答する 2 値分類問題となっている．COLIEE では，過去の司法試験から作成された問題が学習データとして配布され，開催当時最新の司法試験から作成された問題がテストデータとして使われている．年度ごとの問題の数，含意ラベルの数を Appendix A.1 に示す．

## 3　ルールベースのデータ拡張

### 3.1　COLIEE 2021 SOTA

Aoki ら [3] は，BERT モデルを用い，民法条文を利用したルールベースによる含意学習データの拡張とアンサンブルを組み合わせることで，COLIEE 2021 年度 TASK4 で最高スコアを記録した．

Aoki らは，図 1 に示すように，論理構造を特定して民法の各条 $R$ に対して $R$ が含意する文 $S_i$ を作成し，$S_i$ を使用して COLIEE の拡張データ $Q_i$ を作成した．$S_i$ の作成は，文末及び段落情報を使用して民法を分割し，$R$ 内の参照をルールベースで展開することで行う．COLIEE の拡張データ $Q_i$ の前件 $P_{Q_i}$ として $S_i$ を使用する．仮説 $H_{Q_i}$ として $S_i$ を与えて含意ラベルを付与し，正例とする．また，仮説 $H_{Q_i}$ として $S_i$ の文末をルールベースで否定させた文を与えて非含意ラベルを付与し，負例とする．

この手法により，正例 2,074 件を含む 4,062 件の

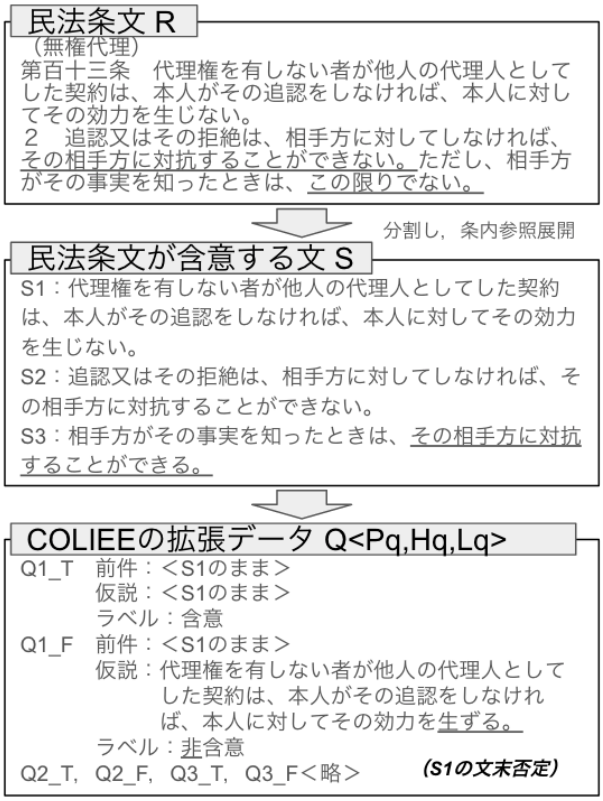


**図 1**　Aoki らのデータ拡張手法

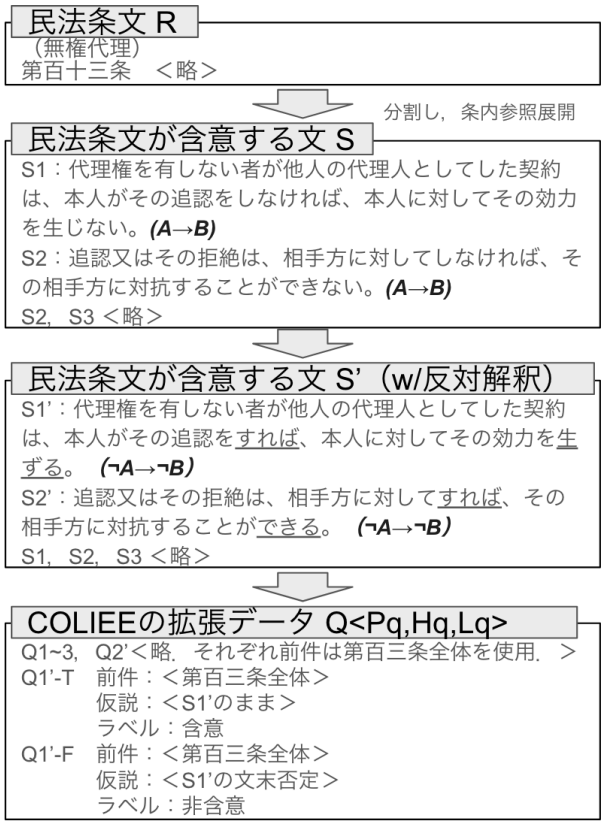


**図 2**　反対解釈を使用したデータ拡張手法

拡張データを作成する．本研究では Aoki らのデータ拡張手法をベースラインとする．

### 3.2　反対解釈を用いた追加拡張

命題論理の体系では命題 $A$，$B$ に対して $A \to B$ ならば裏 $\neg A \to \neg B$ は正しいとは限らない．一方，民法の分野では，「反対解釈 [4]」と呼ばれる概念によって，ある条件の下での規定がある時，条件を満たさないものについては反対の規定がされていると解釈されることがある．

本研究では Aoki らのデータ拡張に加えて反対解釈を用いたデータ拡張を提案する．

反対解釈を用いたデータ拡張の概要を図 2 に示す．$S_i$ に対してさらに辞書を用いて特定の表現を置換することで反対解釈に相当する新たな $S'_i$ を作成する．既存の $S$ に加え作成された $S'_i$ を $S'$ の集合に加えることで，再拡張された $S'$ を生成する．

今回用意したルールでは民法から拡張できた件数が 254 件 (内正例 127 件) と多くなかったため，民法以外も含めた条文集から拡張した 496 件 (内正例 248 件) を使った追加実験も併せて行なった．条文集で使用した法律を Appendix B に示す．

### 3.3　拡張データの与え方の変更

Aoki らの拡張データは，入力文対の前件と仮説の文末を比べることで正解できる．本手法では，COLIEE と問題形式を揃えて拡張データの質を高めることを目的に，図 2 で示すように拡張データの前件 $P_{Q_i}$ として条文一条全体 $R$ を与える．

## 4　文埋め込みの事前学習

事前学習済みモデルに対して SimCSE (Simple Contrastive Learning of Sentence Embeddings)[5] を用いた文埋め込み学習をしてから，COLIEE タスクでファインチューニングする手法を提案する．よりよい文埋め込みを獲得することをねらい，SimCSE を用いる．SimCSE は教師あり，教師なし，どちらにおいても，文埋め込みの良さを測るのに使われる類似度ベンチマークである STS (Semantic Textual Similarity)[6] において教師なしで優れたスコアを記録した．

一般ドメインでは，大規模なラベル付き NLI データセットが日本語でも入手可能 [7, 8, 9] である一方，日本語法ドメインには同様のデータは存在しない．

そこで，一般ドメインにおける文埋め込み性能向上のためには，教師あり学習による追加事前学習をする手法を採用し，法ドメインのデータを用いた文埋め込み性能の改善には教師なし SimCSE を用いた追加事前学習を採用した．以下の実験における実験設定は，入力トークン長を 512 へと変更した点を除いてオリジナルの SimCSE と同一である．ただし，条文集以外はオリジナルと同規模の 100 万文を使用して 1 エポックの学習をおこなったが，条文集を使用した学習は，およそ 1 万文で 100 エポックに変更した．

オリジナルの SimCSE と対応して，最良のモデルを選ぶ指標として，JGLUE （Japanese General Language Understanding Evaluation）[7] の JSTS のスピアマンの順位相関係数を使用した．125 ステップごとに JSTS の文対をそれぞれモデルに入力し，スピアマンの順位相関係数が最も大きいステップのモデルを選ぶ．SimCSE を適用するモデルとして，日本語 WikiBERT[1)]，及び JLBERT-FP[10] を使用した．JLBERT-FP は，日本語法分野に特化した BERT モデルである．

### 4.1 教師なし文埋め込み学習

SimCSE の教師なし学習は，BERT のドロップアウトを利用した単純なデータ拡張を使用した対照学習である．同一文に対して擬似的に拡張をした文埋め込み同士が近く，異なる文に対する文埋め込み同士が遠くなるように学習する．

教師なし学習においては，法ドメインデータの判決書データ，条文集に加えて，一般ドメインデータの Wikipedia を用いる．判決書データには平成十二年から令和二年までの民事事件下級審判決 52,967 件を用いている．判決書データ及び Wikipedia については，ランダムに 100 万文を抽出した．条文集はおよそ 1 万文を抽出した．

一部手法は一般ドメインデータにおける文埋め込み学習後に法ドメインデータで文埋め込み学習をしており，実験結果の表 1 において右矢印（→）で示す．

### 4.2 教師あり文埋め込み学習

SimCSE の教師あり学習は，NLI のデータセットを利用した対照学習である．NLI のデータセットは前件，仮説ペアに対する含意，非含意がラベリングされたデータセットであり，SimCSE では，前件の文埋め込みに対し，含意ラベルが付与された仮説の文埋め込みが近くなるように学習する．

ここでは NLI データセットに，JSNLI[9] を使用した．SimCSE の構造上，前件に対して仮説が含意，非含意いずれも存在する三つ組を用意する必要があったため，JSNLI から 137,563 セットの三つ組を作成し，学習に使用した．

## 5 実験設定

### 5.1 ファインチューニング

条文と問題文を [SEP] トークンで区切り，BERT に与える．クロスエントロピーロスを使い，エポック数は 5，学習率は 1e-5，バッチサイズは 12，入力トークン最大長は 512 である．Aoki らは最大長 256 で実験を行なっていたが，トークン長の分布 (Appendix A.3) から 512 が適切だと判断した．

事前学習済みモデルには，日本語 WikiBERT，JLBERT-FP，及び SimCSE による事前学習済みモデルを使用した．

### 5.2 評価

性能指標には正解率 (Accuracy) を用いる．COLIEE 2021 データセットは合計でも 806 件に留まり（A.1），年度毎のデータの特徴も異なること（A.2）から，モデルの性能評価では以下の方法を取る．

平成十八年度から令和元年度のうち，1 年度をテストデータとし，残りのうち 1 割を検証データ，9 割を学習データとする実験を，各年度について行い計 14 年度分実施することを 1 回の実験と定義する．1 実験全体のスコアは，マイクロ平均を用いる．各手法について 5 回ずつ実験を行い，その平均を各手法のスコアとする．各手法の正解率の平均 について，並べ替え検定 (有意水準 5%, 両側検定) を行った．

## 6 実験結果

実験結果を表 1 に示す．提案手法 (#3) によって，ベースライン (#1) から正解率が向上し，全手法を通して最高性能となった．一方で，今回提案したデータ拡張手法 (#3, #4) とベースライン (#1) との間に統計的な有意差は確認できなかった．ベースラインは，利用する事前学習済み言語モデル及び SimCSE で利用するデータの種類に関わらず，データ拡張な

1) https://github.com/cl-tohoku/bert-japanese より cl-tohoku/bert-base-japanese-whole-word-masking を使用．

**表 1** 手法ごとの評価

| # | モデル | SimCSE | データ拡張 | 正解率 |
|---|---|---|---|---|
| 1 | J | なし | Aoki | 60.37** |
| 2 | J | なし | なし | 52.08* |
| 3 | J | なし | Ours(民法) | 61.24 |
| 4 | J | なし | Ours(+条文集) | 59.58 |
| 5 | J | Wiki→ 判決書 | Aoki | 59.53 |
| 6 | J | Wiki→ 判決書 | なし | 49.78* |
| 7 | J | Wiki→ 法律 | Aoki | 58.06 |
| 8 | J | Wiki→ 法律 | なし | 50.37* |
| 9 | J | Wiki | Aoki | 60.05 |
| 10 | J | Wiki | なし | 51.02* |
| 11 | J | 判決書 | Aoki | 58.98 |
| 12 | J | 判決書 | なし | 51.66* |
| 13 | J | 教師あり NLI | Aoki | 60.20 |
| 14 | J | 教師あり NLI | なし | 53.57* |
| 15 | L | なし | Aoki | 58.76 |
| 16 | L | なし | なし | 56.25* |
| 17 | L | 判決書 | Aoki | 57.94 |
| 18 | L | 判決書 | なし | 52.66* |
| 19 | L | 法律 | Aoki | 60.05 |
| 20 | L | 法律 | なし | 53.77* |

(学習済みモデルの J は日本語 WikiBERT，L は JLBERT-FP．*は**に対して有意差を確認．)

しの結果に対して常に高い正解率を示し，統計的な有意差も確認できた．

SimCSE を用いた手法 (#5 – #14, #17 – #20) は，いずれも Aoki らの手法 (#1) 及び提案手法 (#3) を下回る正解率となった．

Aoki らのデータ拡張を適用した場合の性能向上幅を，日本語 WikiBERT (#1, #2) と JLBERT-FP (#15, #16) の間で比較すると，日本語 WikiBERT を用いた手法でより大きい．SimCSE 及びデータ拡張をどちらも用いない手法 (#2, #16) の中では，JLBERT-FP が高い性能を発揮した．

## 7 考察

反対解釈を用いた手法 (#3) はベースライン (#1) に対して改善が見られた．しかし，改善幅が小さく，統計的な有意差までは確認できなかった．

問題の内容を確認するために令和元年度をテストデータとした実験を抽出して分析した．令和元年度データにおいて，反対解釈が必要である問題は少なくとも 3 件 (問 3，問 21，問 53) 存在した．令和元年度データに対する予測結果を手法 #1, #3, #4 間で比較したところ，手法 #1 では 0/3 問正解だったのに対し，手法 #3 及び #4 では 1/3 問正解であった[2)]．

2) 各手法について 5 回実験を実施しているため，結果を多数決によってマージしている．

このことから反対解釈を用いた拡張は部分的には有効に作用している可能性がある．改善幅が小さい原因は，反対解釈によるデータ拡張量が少なかったことが挙げられる．反対解釈による拡張データはルールベースによって作成した．反対解釈が可能な条文を厳密に抽出するようにルールを設計したため，本来は拡張に利用可能な条文であっても排除している可能性がある．

拡張データ件数を確保するため条文集を追加した手法 #4 は，手法 #3 を下回る性能を示した．これは民法向けに構築した拡張ルールをそのまま他の法典に適用したことが原因と推測できる．

SimCSE を利用した文埋め込み学習手法は Aoki らの手法に対して改善が見られなかった．SimCSE におけるモデル選択の指標が適切でないことが原因として考えられる．JSTS のような大規模な意味的類似度計算タスクデータセットは日本語法ドメインでは存在しないため，実験では一般ドメインデータである JSTS を用いてモデル選択を行ったが，本来は法ドメインデータを用いることが望ましい．法ドメインにおける低資源問題を回避するため教師なし SimCSE の活用を検証したが，モデル選択に必要な検証データとして一定量のラベル付き法ドメインデータが確保できない点が，依然としてボトルネックとなってしまった．

## 8 結論

本稿では，法ドメインの含意タスクの低資源問題を回避するため，1) 反対解釈によるデータ拡張手法, 2) SimCSE を用いた文埋め込み学習の利用を検証した．

従来のデータ拡張手法に更に反対解釈によるデータ拡張を行うことで，性能が向上することを示した．今後は，反対解釈ルールの拡張や言語生成モデルを活用した生成的なデータ拡張の手法の検討を行う．

一方で，SimCSE の有用性は示せず，教師なし学習であっても検証用データすら不十分なドメインの場合は最良モデルの選択が難しいことを確認した．

## 謝辞

本研究で使用した判決書データは株式会社 LIC から提供を受けたものである．本研究の一部は，JST, ACT-X，JPMJAX20AM の支援を受けたものである．

## 参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. **The Review of Socionetwork Strategies**, Vol. 16, No. 1, pp. 111–133, Apr 2022.

[3] Yasuhiro Aoki, Masaharu Yoshioka, and Youta Suzuki. Data-augmentation method for bert-based legal textual entailment systems in coliee statute law task. **The Review of Socionetwork Strategies**, Vol. 16, No. 1, pp. 175–196, Apr 2022.

[4] 早田幸政. 入門 法と憲法. ミネルヴァ書房, 2014.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In ***SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)**, pp. 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[7] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. 言語処理学会 第 28 回年次大会発表論文集, 2022.

[8] 谷中瞳, 峯島宏次. JSICK: 日本語構成的推論・類似度データセットの構築. 人工知能学会第 35 回全国大会, 6 2021.

[9] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会第 244 回自然言語処理研究会, 7 2020.

[10] 宮崎桂輔, 菅原祐太, 山田寛章, 徳永健伸. 日本語法律分野文書に特化した BERT の構築. 言語処理学会 第 28 回年次大会 発表論文集, 2022.

# A COLIEE

## A.1 COLIEE の基本情報

COLIEE の年度ごとの問題数と含意ラベル数を表 2 に示す．

**表 2** COLIEE の年度ごとの数

| 年度 | データ数 | 含意ラベル |
|---|---|---|
| H18 | 36 | 16 |
| H19 | 37 | 22 |
| H20 | 41 | 27 |
| H21 | 54 | 30 |
| H22 | 47 | 21 |
| H23 | 41 | 21 |
| H24 | 79 | 36 |
| H25 | 60 | 29 |
| H26 | 74 | 43 |
| H27 | 49 | 25 |
| H28 | 49 | 20 |
| H29 | 58 | 27 |
| H30 | 70 | 36 |
| R01 | 111 | 59 |
| all | 806 | 412 |

## A.2 年度間での傾向の異なり

手法 #1 の実験について年度ごとの結果を表 3 に示す．

**表 3** 年度ごとの実験の正解率

| 年度 | 正解率 | 標準偏差 | データ数 |
|---|---|---|---|
| H18 | 63.33 | 2.32 | 36 |
| H19 | 55.67 | 5.60 | 37 |
| H20 | 61.46 | 2.67 | 41 |
| H21 | 55.92 | 7.56 | 54 |
| H22 | 54.04 | 9.10 | 47 |
| H23 | 60.00 | 7.43 | 41 |
| H24 | 62.02 | 8.72 | 79 |
| H25 | 64.66 | 7.20 | 60 |
| H26 | 49.18 | 3.39 | 74 |
| H27 | 57.95 | 3.70 | 49 |
| H28 | 64.89 | 1.70 | 49 |
| H29 | 57.93 | 6.28 | 58 |
| H30 | 58.85 | 4.67 | 70 |
| R01 | 62.70 | 0.80 | 111 |

## A.3 トークン長

H18 年から R01 年のデータ 806 個に対するトークン長の分布を図 3 に示す．256 より大きいデータは 118 個，512 より大きいデータは 12 個ある．

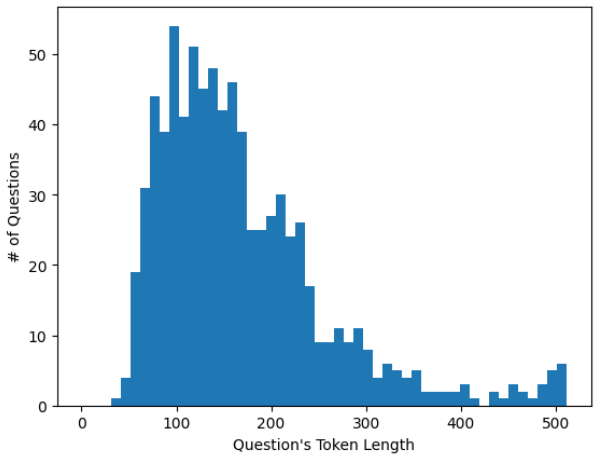


**図 3** トークン長の分布

# B 条文集

反対解釈及び SimCSE で使用した憲法，法律を以下に示す．以下から成る条文集から 10,488 文を抽出した．e-Gov 法令検索[3] のデータを利用した．

- 行政機関の保有する情報の公開に関する法律
- 行政事件訴訟法
- 行政手続法
- 行政不服審査法
- 刑事訴訟法
- 刑法
- 個人情報の保護に関する法律
- 国家賠償法
- 戸籍法
- 借地借家法
- 商業登記法
- 商法
- 宅地建物取引業法
- 賃貸住宅の管理業務等の適正化に関する法律
- 日本国憲法
- 不動産登記法
- マンションの管理の適正化の推進に関する法律
- 民事執行法
- 民事訴訟法
- 民事保全法
- 民法
- 利息制限法
- 労働関係調整法
- 労働基準法
- 労働組合法

3) https://elaws.e-gov.go.jp/