要件に対する効果の生成を経由した法律分野の自然言語推論

チェ ジョンミン [1,2]　本多右京 [1,2,3]　渡辺太郎 [1]　乾健太郎 [2,4]
[1] 奈良先端科学技術大学院大学 [2] 理化学研究所
[3] 株式会社サイバーエージェント [4] 東北大学
choi.jungmin.ce6@is.naist.jp, honda_ukyo@cyberagent.co.jp
taro@is.naist.jp, kentaro.inui@tohoku.ac.jp,

概要

法律分野における自然言語推論（NLI）は，前提（法律）と仮説（法的問題に関する記述）の間の含意関係を予測するタスクである．一般的な NLI に対する現在の state-of-the-art である，事前学習済みの言語モデルを用いた手法は，法律分野においてはそれほど有効でない．これは、前提と仮説の抽象度の違い，ならびに法律文の特殊性によるものと考えられる．本研究では，法律文が，原則として要件と効果によって構成されるという特性に着目し，NLI のタスクをより容易なサブタスクに分解することによってこの問題に対処し，既存手法を上回る性能を得た．

1　はじめに

自然言語推論 (NLI) とは，前提文と仮説文のペアが与えられたとき，それらの関係を，前提が仮説を包含する場合は含意，仮説が前提に反する場合は矛盾，どちらの関係も成立しない場合は中立 (neutral) に分類するものであり，自然言語処理の中心的な課題の一つであるといえる。法律分野における，このタスクの最も自然な形態は，「法律問題に関するある言明が法律に照らして正しいかどうかを判定する」というものであるが，これは法曹資格を審査する司法試験の一部を構成するものであり，法律家の仕事を自動化するシステムの開発のためには不可欠といえる．

法律分野に特有の NLI の難しさは，アノテーションのコストが高いためにデータが不足しがちで，モデルに法的概念の意味や具体例を学習させにくいというところにある．

最近の研究では，Transformer ベースの事前学習済み言語モデルによって，前提と仮説のベクトル表現を求め，これを線形変換によって分類するものが主流になっている [1, 2].

われわれは，法律分野の文章の特性を利用した，根本的に異なるアプローチを提案する．まず，法律文は要件と効果で構成される，という点に注目する．法律に関する仮説も，要件に対応する箇所（以下，単に要件部という）と効果に対応する箇所（以下，単に効果部という）で構成される．人間が法律文を現実の状況に適用する際には，1) 事実に適合する要件を法律文から選び，2) その効果を特定し，事実にあてはめて具体的な効果部を導出する，という過程を辿る．そこでわれわれは，モデルの学習においても，このように多段階の手順を踏むことで学習が容易になるという仮説を立てた．すなわち，タスクを，1) 仮説の要件部に対応する法律要件を特定，2) 法律効果を仮説の設定にあてはめて，仮説が前提に含意されるような効果部を生成，3) 生成された効果部と仮説の効果部を比較，というサブタスクに分割して解く．

また，要件と効果の区別を考慮したデータ拡張も行う．

提案手法が有効であると考えられる根拠は，以下のようにまとめられる．1) 生成された効果部と仮説の効果部に焦点を絞ることで比較が容易になる，2) 条文の論理構造に基づいた擬似データ作成によって，複数のルールから，状況に適合するルールを選ぶという問題にモデルを習熟させることができる，3) 事前学習済みの言語モデルが保有している世界知識（腕時計は動産にあたる，など抽象的な概念と具体物の関係）を使うことができる．

さらに，生成された結果部は，モデルの予測の説明ととらえることもできる．なぜなら，モデルは仮説の要件部を受けて，法律に照らして適切と考えられる効果部を出力するよう学習しているからである．

本研究では，日本の司法試験問題を元に作成され

[前提] **占有者がその占有を奪われたときは、占有回収の訴えにより、**その物の返還及び損害の賠償を請求することができる。占有回収の訴えは、占有を侵奪した者の特定承継人に対して提起することができない。ただし、その承継人が侵奪の事実を知っていたときは、この限りでない。
[仮 説] **Bが海外出張のため１週間大学を留守にしていた間に，Cが甲を盗み出して現に所持している場合，Bは，Cに対し，**占有回収の訴えにより甲の返還を求めることができる。
[Label] 1 (含意)

図 1　法律 NLI の例．COLIEE[3] タスク 4 より抜粋．要件は太字+青字，効果は下線+赤字で示している．

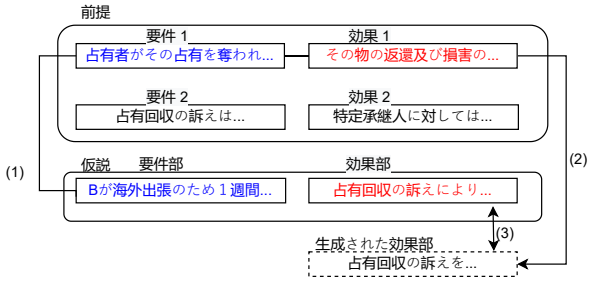

図 2　提案手法の概要．(1) 前提 (一つまたは複数の条文) に含まれる要件から，仮説の要件部と合致するものを同定する．　(2) その要件に対応する効果を，仮説の設定に合わせて書き換え，出力する．(2) 生成された効果と，仮説の効果部を比較し，含意例か矛盾例に分類する．

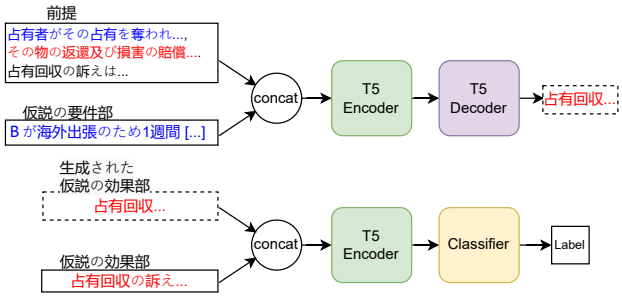

図 3　手法の手順. (1) 仮説を，要件部と効果部に分ける．(2) encoder-decoder モデルを，前提と仮説の要件部から導かれる順当な効果部を出力するよう訓練する．(3) encoder モデルと分類器を，生成された効果と仮説の効果部を入力したとき，含意例であれば 1 を，矛盾例であれば 0 を出力するよう訓練する．

た日本語の法律 NLI データセットで実験を行った．その結果，最先端の手法やベースラインと比較して，大幅な性能の向上を確認した．

2　先行研究

法律分野の NLI においては，前提は法律、たとえば条文の部分集合であり，仮説はその法律に関連する状況を記述し，その状況の法的帰結について述べるものである．仮説が前提によって含意されているかどうか，すなわち，仮説が法律に照らして正しいかどうかを予測することが具体的なタスクとなる．

一般的な NLI に関して大きな成功を収めた，BERT[4] など事前学習モデルによるベクトル表現を分類するという手法は，そのままでは，法律分野のデータに関してそれほどの性能を示さないことが報告されている [3]．その原因は，法律文書に使われる抽象的な語彙と，一般的な語彙の乖離により，モデルが事前学習した知識を適用することが難しいこと，法律分野のデータが不足していることにあると考えられる．近年の法律分野の NLI 研究においては，ベクトル表現の分類という一般的な枠組みを踏襲しつつ，上述の問題を解決するための補助的な工夫を行った手法が主流になっている．現時点での state-of-the-art である [1] は，ドメインデータを補うため，条文をそれ自身とペアとして擬似正例，条文とその否定形をペアとして擬似負例を作成する．[2] は，シソーラスを用い，各単語に対応する意味的カテゴリ情報を付与し，抽象的な概念と具体物の関連をモデルに明示的に与えることで，前提と仮説の抽象度の差を緩和している．これらの手法は，いずれも一定の有効性を示しているものの，正解率は一般的な NLI データには遠く及ばず，改善の余地が大きいものとみられる．

3　提案手法

われわれは，効果部の生成と分類という二段階から成る手法を提案する．提案手法の概要を図 2 に示す．まず，仮説の要件部を受けて，前提から導かれる順当な効果部を生成する．すなわち，もし含意例であれば仮説の効果部となるべき系列を出力する．そして，生成された効果部が，実際の仮説の効果部と一致するか否かを分類する．具体的には，以下のようにモデルを学習させることを目指す．1）前提に含まれる複数の要件から，仮説の要件部に合致するものを見つける，2）前提の中から対応する効果を見つけ、仮説の状況に当てはめて具体的な効果を生成する、3）実際の仮説の効果と比較する．手法の詳細な手順については、 図 3 を参照されたい．

3.1　効果部の生成

生成モデル $P(c|p,h)$ は，前提 p と仮説の要件部 h を与えられ，そこから導かれる効果部 $\hat{c}$ を生成する．まず、各仮説を要件部と効果部に分割する．厳

密な分割のルールを同定することはできないが，検証データでは、ほとんどの場合、最後のカンマの後のどこかで分割されていることがわかる．そこで、最後のカンマ以降の全てのトークンで分割し，それぞれを例として扱う，というヒューリスティックを用いる．仮説中にカンマが存在しない少数の場合においては，最後から８番目のトークンから分割を開始する．

3.2 分類

生成の学習が終わった後，全ての例に関してモデルに効果部を生成させる．そして，生成された効果部と仮説の効果部を連結して一つの系列とし，ベクトル表現に変換したものを，線形変換による分類器で含意と矛盾に分類する．より形式的にいえば，$\hat{c}$ と c という 2 つの入力が与えられたとき，分類モデル $P(y|\hat{c}, c)$ によってラベル y を予測する．これが可能であると考えられる根拠は次のようにまとめられる．モデルは，生成の学習の際，含意例のみを用いているため，どのような例についても，「含意例として妥当な効果部」を生成するはずである．したがって，実際に含意例であれば生成された効果部と仮説の効果部が類似し，矛盾例であれば相違していると期待される．短い系列である効果部同士を比較するだけで分類できるため，問題をより簡単な形に落とし込めているといえる．

3.3 データ拡張

提案手法は，事前学習済み言語モデルの知識を利用することで，データ不足の影響を緩和することを想定しているが，法律分野で使用される語彙は通常の語彙とは大きく異なるため，ドメインデータをなるべく多く学習しなければならないことに変わりはない．そこで，条文から擬似的な事例を作成する．[1] は，各条文を個別の規範に分割し、各規範を前提，それ自身を仮説にして含意例を作成し，他方で各規範を前提，その否定を仮説として矛盾を作成する．この擬似データの目的は，モデルに文とその否定の違いを理解させることにあるが，われわれはさらに，モデルに複数の規範から仮説に対応する一つの規範を選んで効果部を生成させるように擬似データを設計する．ここでは，2 つの要件効果ペアが接続詞によって連結された条文に注目する．1 の例で説明する．前提中の条文の一つは，「占有回収の訴えは、[...] 提起することができない。」と，「その承継人が [...] この限りでない。」という二つの規範から成る．この条文を前提とし，規範の内の一つ，たとえば，「占有回収の訴えは、[...] 提起することができない。」を仮説として含意例ができる．他方で，この条文と，規範の効果「占有を侵奪した者の特定承継人に対して提起することができない。」と「占有を侵奪した者の特定承継人に対して提起することができる。(「この限りでない」を書き下した[1])」を入れ替え，「占有回収の訴えは、占有を侵奪した者の特定承継人に対して提起することができる。」とした偽の規範は矛盾例を成す．もう一方の規範を使っても同様のことがいえる．[2]

4 実験

COLIEE ワークショップ [3] において提供された法律 NLI データセットを使用する．[3][4]．図 1 のように，前提は一つまたは複数の民法の条文であり，仮説は前提に照らして真偽が定まる一つの言明である．ラベルは含意または矛盾の二値であり，中立は存在しないことに注意されたい．元の訓練データとテストデータは，それぞれ 806 例と 81 例から成る．本実験では，元の訓練データを 81 例の検証データと 725 例の訓練データにランダムに分割した．回答欄の指示などのノイズを除去すると，前提および仮説を連結した平均の文長は約 120 トークンである．680 例の擬似データは，テキスト生成と分類学習の両方に使用する．

COLIEE にしたがい，正解率を主たる評価指標として採用する．また，分析のために精度，再現率，F1 スコアも併記する．

生成モデルとして，一般的な日本語コーパスで事前学習した T5 モデル[5]を使用する．

以下の手法を試す．まず，これまでの state-of-the-art である [1] を再現し，これを “Aoki” と表記する．

1) この書き下しは，[1] の提供する辞書を使用して行った．
2) われわれの手法では，このように，「ただし」を含む条文しかデータ拡張に使用しないため，作成される擬似例の数は 680 となり，[1] の 3,351 例を大きく下回る．しかし，前提中に複数の規範が含まれ，仮説に直接関係するのはそのうちの一つ，という性質は法律 NLI 特有であり，このような例を集中的に学習することが特に重要であると考えられることから，本研究ではこれを採用した．
3) データセットは 2021 年度コンペティションのタスク 4 に使用されたものである
4) COLIEE データセットでは，人物は A，B のように記号で表されている，日本語の T5 トークナイザはこの記号を語彙として持っていないため，人物名をよくある日本の姓に置き換えた．
5) https://github.com/sonoisa/t5-japanese

表 1 テストデータにおける手法別のスコア．スコアは 10 回の試行の平均．カッコ内は標準偏差．Aoki は [1] の報告，Aoki† は本論文著者らによる再現

model	Acc.	Prec.	Recl.	F1
Aoki	63.9±2.2	–	–	–
Aoki†	63.7±4.5	62.8±7.3	65.8±19.1	61.9±7.8
Aoki w our aug	54.1±6.2	48.3±18.4	64.5±33.2	51.9±20.6
T5 baseline 1				
w Aoki's aug	60.0 ±4.0	56.4±4.6	71.1±9.0	62.4±3.1
T5 baseline 1				
w our aug	58.3±4.1	61.8±7.6	36.1±21.9	41.4±16.5
T5 baseline 2				
w Aoki's aug	61.1±4.9	57.9±5.4	66.3±4.9	61.6±3.5
T5 baseline 2				
w our aug	68.5±1.6	66.9±3.5	66.1±4.6	66.3±1.4
TG w\o aug	61.1±7.4	69.5±14.5	58.0±6.7	61.1±4.2
TG w Aoki's aug	65.6±4.9	61.2±6.5	**77.1** ± 5.3	**67.9** ± 2.5
TG (Ours)	**71.0** ± 1.3	**71.3** ± 2.2	64.2±5.1	67.4±2.5
TG w BERT	69.0±3.1	67.6±3.5	65.3±5.4	66.3±3.8

また，[1] と同じアーキテクチャで，擬似データのみ提案手法に置き換えた実験を行い，これを “Aoki w our aug” とする．“T5 baseline 1” は，[5] の手法にしたがい，前提と仮説を単に連結して T5 に入力し，含意であれば「はい」，矛盾であれば「いいえ」を出力するように訓練するという手法である．“T5 baseline 2” は，生成の学習をスキップし，前提と仮説を直接 T5 の encoder に入力することを除けば，提案手法と同じである．“T5 baseline 1,2” に関しても，われわれのデータ拡張手法と [1] のデータ拡張手法を試し，それぞれ，“w Aoki's aug”, “w our aug” を付け加えて示す．最後に，提案手法とほぼ同じだが，分類段階のエンコーダーとして，T5 ではなく BERT を用いる手法である “TG w BERT” の実験を行う．

提案手法は “TG (Ours)” と表記し，データ拡張を全く行わないものを “TG w\o aug”，[1] の擬似データを使ったものを “TG w Aoki's aug” と表記する．

ハイパーパラメータについての詳細は B に示す．

4.1 結果

表 1 は 10 回の試行の平均値と標準偏差を示す．提案手法は，他の手法と比較して，最も高い精度と正解率を達成した．“TG” と “T5 baseline 2” では提案手法のデータ拡張の方が効果が高く，“Aoki” では [1] のデータ拡張の方が高かった．“T5 baseline 1” ではデータ拡張による有意な差は見られなかった．提案手法のデータ拡張は，要件と効果という法律文書の論理構造をモデルが学習するよう設計されているため，要件に対する効果の生成を行う提案手法のアーキテクチャと組み合わさったとき最も効果を発揮するのは期待通りの結果である．他の手法では，前提に含まれる規範のうちどれが仮説に関連しているかを明示的には学習しないため，提案手法のデータ拡張が奏功しなかったと考えられる．

いずれのデータ拡張を採用するにせよ，提案手法は “T5 baseline 1” と “2” を上回る性能を見せ，提案手法の成功が T5 に負うものではないことを示した．“T5 w BERT” は “Aoki” より高い性能を示している．前者が後者と違う点は，前提・仮説の代わりに生成された効果部と仮説の効果部を BERT に入力することのみであり，このことから，生成によって効果部の比較に問題を収斂させることが有効であることが確認できる．

4.2 分析

表 2 に示した例のように，提案手法の生成は期待に沿った挙動を見せている．38 件の含意例のうち，細部まで仮説と意味的に等しい効果部を生成することに成功した例は 27 件であった．失敗した 11 件においては，1) 前提と仮説の間の概念をの関連性を認識するために外部知識を必要とする，例えば，所有物を贈り物として受け取った人は「特別承継人」である等，2) 仮説中に 3 人以上の人物が登場する，という特徴が見られた．43 件の矛盾例のうち，モデルが仮説の効果部と明らかに相違する効果部を生成することに成功した例は 35 件あった．失敗した 8 件においては，1) 法律用語としても日常用語としても使われるがその意味が全く異なる語彙が含まれる，2) 要件部が，逆接または順接の接続詞で終わっており，接続詞に正しく続くような効果部を生成するよう誘導されてしまう，という特徴が見られた．

5 おわりに

本研究では，仮説を要件部と効果部に分割し，前提に基づき，要件部に対する効果部を生成することを経由して，含意関係を予測する手法を開発した．これは予測の性能を向上させるとともに，予測のための説明を提供するものであり，NLI にとどまらず，質問応答タスクへの発展の可能性を示している．

謝辞

貴重なデータを提供していただいた COLIEE 運営者の皆様に感謝いたします．実験に関しては Tohoku NLP グループの皆様に有益な助言をいただきました．感謝いたします．

参考文献

[1] 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, pp. 175–196, 2022.

[2] Mi-Young Kim and Juliano Rabelo. Bm25 and transformer-based legal information extraction and entailment. In **ICAIL/COLIEE**, 2021.

[3] Juliano Rabelo, Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, and Ken Satoh. Summary of the competition on legal information extraction/entailment (coliee) 2021. In **ICAIL/COLIEE**, 2021.

[4] 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.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**, 2015.

A　生成の結果

表 2　生成された効果部の例

生成された効果部	仮説の効果部	ラベル
の抵当権者その他の第三者の承諾を得ることを要しない。	Fの承諾が必要である。	矛盾
その動産の所有権の取得を主張することはできない。	Aが即時取得により宝石の所有権を取得することはない。	含意
A の代金支払債務は当然に消滅する。	損害賠償請求権から優先弁済を受けることができる。	矛盾
占有者に対してその物の回復を請求することができる。	Dに宝石の回復を請求することができる。	含意

B　ハイパーパラメータ

TG，TG w\o aug，TG w Aoki's aug に関しては，最適化手法として Adam[6]，損失関数として cross-entropy loss を用いる． 学習率は，生成では 1e-4，分類では 1e-5 と設定する．バッチサイズは生成では 16，分類では 32 である．

T5 baseline 1 に関しては， 最適化手法として Adam，損失関数として cross-entropy loss を用いる．学習率は，1e-4． バッチサイズは 64.

T5 baseline 2，T5+BERT に関しては，最適化手法としては Adam，損失関数として cross-entropy loss を用いる．学習率は，生成では 1e-4，分類では 1e-5 と設定する．バッチサイズは，生成では 16，分類では 32 である．

[1] の再現では，[1] で報告されている通りのハイパーパラメータを用いた．最適化手法は Adam，損失関数は cross-entropy loss，学習率は 1e-5，バッチサイズは 12 である．

C　アンサンブルによる予測

[1] はアンサンブルでの予測の結果を報告しているため，本研究でも同様のアンサンブルを行う．10 回の試行を行い，その全ての組み合わせに関して，ロジットの算術平均を取ることでアンサンブルを作る．検証データで予測の正解率が上位 3 位のアンサンブルを ensemble 1， 2， 3 と名づけ，テストデータでの正解率を示す．アンサンブルに関しても TG が最も高い性能を示している．

表 3

model	ensemble 1	ensemble 2	ensemble 3
Aoki	67.90	67.9	70.4
Aoki†	67.9	69.1	61.7
T5 baseline 1 w our aug	63.0	65.4	65.4
T5 baseline 2 w our aug	67.9	67.9	69.1
TG w\o aug	61.7	66.7	63.0
TG w Aoki's aug	72.8	74.1	72.8
TG (Ours)	74.1	72.8	75.3
TG w BERT	71.6	71.6	70.4