

# 見出し情報を考慮した階層型 RNN による 日本語判決書のための修辞役割分類

山田寛章<sup>1</sup> Simone Teufel<sup>2,3</sup> 徳永健伸<sup>2</sup>

<sup>1</sup> 東京工業大学  
情報理工学院

<sup>2</sup> 東京工業大学 情報理工学院  
Tokyo Tech World Research Hub Initiative

<sup>3</sup> University of Cambridge  
Computer Laboratory

yamada.h.ax@m.titech.ac.jp simone.teufel@cl.cam.ac.uk take@c.titech.ac.jp

## 1 はじめに

我々は、日本国の判決書に対する情報アクセスの容易化・効率化を目指し、検索の手掛かりとなる判決の要約を機械的に生成することを目指している。判決書にはその判決に至るまでの詳細な議論が記録されており、その議論・論理構造には一定の共通性がある。これは、執筆を担当する者が高度に教育された裁判官であることや、裁判の流れ・裁判制度の構造に起因する記述内容の規則性によるものである。このような判決書に内包される共通の議論構造を適切に利用できれば、高い品質の自動要約を実現することが可能となる。

判決書の階層的な議論構造を一括して抽出することは難度が高く、その性能評価も難しい。そこで、我々は判決書の要約のために必要十分な粒度の情報を抽出する4つの小タスクを定義した [1]。

1. 修辞役割分類 判決書からの基本的な議論構造を表現するために、先行研究 [2, 3] に従って表 1 に示す修辞役割を設計した。この小タスクは文書中の各文に対して最も適切な修辞役割を一つ割り当てる。
2. 争点特定 判決書における議論の中心的話題を示す「争点」が記述されている文を特定する。
3. 争点関連付け 特定した「争点」と関連する文書中の各文を Issue Topic リンクによって関連付ける。
4. **FRAMING** リンク修辞役割分類によって特定された修辞役割の内、FRAMING-main と FRAMING-sub (または BACKGROUND) 間の関係を FRAMING リンクによって関連付け、議論上の支持関係を表現する。

本稿では、4 タスクの中で主要な役割を果たす修辞役割分類の自動化及びその性能改善について議論す

る。これまでの日本の判決書における修辞役割分類の研究 [4] では、Conditional Random Field (CRF) を用いた分類器を構築し F=0.63(マクロ平均値) の性能を達成していたものの、BACKGROUND(F=0.32) 及び CONCLUSION(F=0.39) の重要な役割について相対的に分類性能が低くなっており、改善の余地があった。本稿では、文間文脈を考慮可能な階層型 RNN をベースとするモデルを用いることで、日本国判決書における修辞役割分類の性能が従来の CRF による分類器に比べて向上することを示す。また、判決書中に出現する見出し情報を扱う専用のネットワークを階層型 RNN に追加することで修辞役割分類の性能が向上することを示す。

## 2 関連研究

修辞役割分類分析 [5] は科学研究論文の要約のための手法として提案されたものであるが、法的文書の分野においても、同様のアプローチを用いて抽出型要約または重要箇所抽出を行う研究が存在する。Hachey ら [2] は英国の判例法システム下における判決書を対象として、文書中の各文に対する修辞役割分類を機械学習を用いて行い、各文が判決書中で果たす修辞役割 (例: 結論, 法律の引用等) を弁別した上で要約生成を行った。また、修辞役割と同様のアプローチを利用して Farzindar ら [6] らがカナダの司法制度下における要約生成、Saravanan ら [3] がインド司法制度下における要約の研究を報告している。日本国判決書の自動要約については、阪野ら [7] が重要文を他の文と弁別する抽出型要約に取り組んでいる。

科学研究論文 [8] や利用規約文書 [9] における修辞役割分類に類似のタスクでは、階層型 RNN を用いて文間の文脈情報を利用可能なモデルを構築し、分類性能の向上が報告されている。同様に法的文書において

表 1: 日本国判決書のための修辞役割 [1]

ラベル	説明
FACT	証拠や証言, 状況やイベントを明示して法廷の場に明示的に事実として提示している文.
FRAMING-main	CONCLUSION を導くための立論を構成する部分であって, CONCLUSION を直接支える事実認定や主張を行っている文.
FRAMING-sub	CONCLUSION を導くための立論を構成する部分であって, FRAMING-main を支持する間接事実の認定を行う等, その根拠や理由を示している文.
CONCLUSION	主要事実の認定など, 争点の対象となっているトピックについての結論, または裁判全体の結論を述べている文.
IDENTIFYING	議論対象となる話題・問題を特定している文.
BACKGROUND	過去の裁判例や法条を直接引用または参照して, 適用している文.
OTHER	上記のいずれにも適合しなかった文.

表 2: 各修辞役割の分布

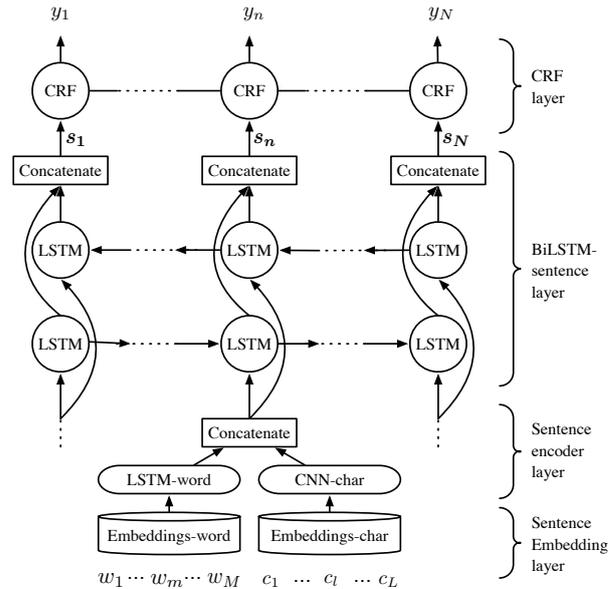
修辞役割	文数 (%)	修辞役割	文数 (%)
FACT	23.3	CONCLUSION	4.0
FRAMING-main	19.1	IDENTIFYING	2.1
FRAMING-sub	10.7	BACKGROUND	0.9

も, Tran ら [10] は Hachey らのデータに対して CNN と RNN を組み合わせた文間文脈を考慮可能な階層型の深層学習モデルを, Bhattacharya ら [11] はインド最高裁の判決書に対する修辞役割分類データに対して階層型 RNN を提案している. 本稿で報告する階層型 RNN モデルは Bhattacharya らによるモデルと同様のものに, 見出し情報を取り扱う専用エンコーダを追加したものとなる.

### 3 判決書コーパス

本稿では, 修辞役割分類, 争点特定, 争点関連付け, **FRAMING** リンクの各タスクについての正解データがアノテーションされた日本国判決書に対する議論構造注釈付きコーパス [4] を修辞役割分類の実験用データとして用いる. 本コーパスは, 裁判所 Web ページからダウンロード可能な, 下級審裁判所より公開された民事事件裁判例を対象として収録したものである. 実際にアノテーションされた文書は人手による判決要旨が付与済みであるものに限定されている.<sup>1</sup> コーパスは 120 文書, 48,370 文 (約 32 億文字) から構成される. コーパス中の各修辞役割の出現数を表 2 に示す.

<sup>1</sup>コーパス中には, 2 人の異なる作業者によって注釈付けされた 8 文書が含まれる.



$w$ : 入力文からの単語  
 $c$ : 入力文からの文字  
 $s$ : ある文に対応する文レベル LSTM からの出力ベクトル  
 $y$ : 予測された修辞役割の出力

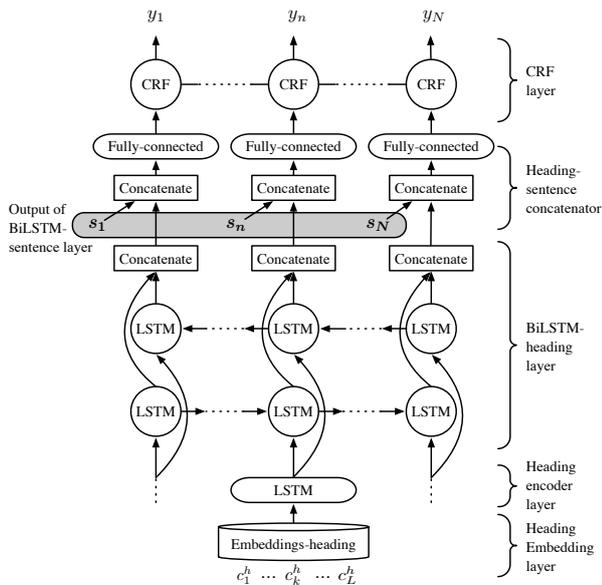
図 1: hBiLSTM-CRF モデル

## 4 修辞役割分類モデル

本稿では 4 種のタスクのうち, 修辞役割分類を対象とする. 判決書における修辞役割分類は文書中の各文を表 1 に示される 7 つのクラスの何れかに分類する多クラス分類である. 本稿では先行研究 [4] に従い, 単語バイグラム, 文位置, 文長, キーワード, モダリティ表現, 日本語機能語表現, 法律の題名を素性とする CRF をベースラインモデル (**CRF-base**) として実装した.

### 4.1 階層型 RNN による分類モデル

本稿では, BiLSTM-CRF を元にした階層型 RNN モデル (**hBiLSTM-CRF**) を日本語で記述された判決書に適用する. BiLSTM-CRF は BiLSTM と CRF を組み合わせたモデルであり [12], 固有表現抽出や品詞解析といった単語系列ラベリングのために提案された. 本モデルはこれを文の系列ラベリングに拡張したものである. 本モデルは, BiLSTM 層によって前後の文脈の情報を取り込み, CRF 層によって入力系列に対してもっともらしい出力系列を求めることで, 入力される文の系列に対してその文脈に応じた修辞役割系列を予測する. 本モデルは, 単語系列から文ベクトルを生成する層 (Sentence encoder), BiLSTM により文ベクトルを文脈化する層 (BiLSTM-sentence), 予測結果を出力する CRF 層 (CRF) の 3 つから構成される (図 1). なお, Sentence encoder は, 単語分散表現を受け取り文ベク



$c^h$ : 入力見出しからの文字  
 $s$ : 図1中におけるベクトル  $s$  に同じ

図2: 見出し情報を考慮した hBiLSTM-CRF モデル

トルを出力する LSTM-word と文字分散表現<sup>2</sup>を受け取り文ベクトルを出力する CNN[14]である CNN-char から構成され、LSTM-word と CNN-char からの出力を結合したものが出力となる。

本モデルは文の系列を入力とする。この入力系列は、サイズ  $w$  のウィンドウを文書頭から文書末まで一文毎に移動することで得られる。ある文書中の  $i$  番目の文を  $S_i$  とした時、その文書における  $n$  番目の入力系列は  $\{S_{n-(w-1)/2}, \dots, S_n, \dots, S_{n+(w-1)/2}\}$  となる<sup>3</sup>。本モデルからは各文に対応する出力が  $w$  個得られるが、このうち入力系列の  $S_n$  に対応する出力を最終出力と見なす。

## 4.2 見出し情報を考慮した分類モデル

文書中に出現する見出しは、人間の読者がその文書を読む際に文のトピックや修辞役割が変化することを把握する手掛かりとなる。同様に、修辞役割分類を行うモデルが見出し情報を利用できれば、分類性能の向上に寄与するはずである。

前節の hBiLSTM-CRF に、見出し情報を取り扱う専用のネットワークをモデルに追加した (hBiLSTM-CRF+H)。この見出し用のネットワークは heading encoder, BiLSTM-heading, heading-sentence concatenator の3つから構成される (図2)。このネットワークへの入力は入力文が見出しか否かで処理が異なる。ある文が見出しであるならばその文がそのまま入力となり、見出しでないならば特殊トークン “.” が入力となる。

<sup>2</sup>判決書を用いて GloVe[13] による事前学習を行った分散表現を用いる。

<sup>3</sup> $n$  は奇数とする。

heading encoder は文字ベースの BiLSTM エンコーダであり、文字分散表現を受け取って見出しベクトルを生成する。BiLSTM-heading では、BiLSTM-sentence と同様に文脈化された見出しベクトルを生成する。BiLSTM-sentence の出力と BiLSTM-heading の出力は結合後に全結合層を経由して CRF 層へと入力される。本モデルの入出力系列の取り扱いは hBiLSTM-CRF に準じる。

また、比較対象として CRF-base に入力が見出しか否かを示す二値の素性を追加したモデル CRF-base+H を実装した。

## 4.3 見出し自動アノテーション

見出し情報を利用するためには、入力データに見出しであるか否かの属性が付与されている必要性があるが、本実験で利用するコーパスには付与されていない。このため、ルールベースの見出し検出器を構築した。検出器は文末の読点の有無と文長を元に見出しか否かを判定する分類器である。hBiLSTM-CRF+H 及び CRF-base+H はこの検出器の出力を見出し判定に利用する。

実験に使用するコーパスの内、修辞役割分類の訓練・テストには使用しない文書から 5 文書 (全 2,061 文中 622 文が見出し) を用いて、本検出器の性能評価を行った。正解データは本稿の第一著者が付与した。結果は  $F = 0.89$  ( $R = 0.99$ ,  $P = 0.81$ ) であった。

## 5 実験

### 5.1 実験設定

実験データとして3節で示した 120 文書からなるコーパスを使用した。このうち 10 文書を開発データとし、ハイパーパラメータの選択及び見出し検出器の開発に用いた。本実験では、開発データを除く 110 文書を用いて、前節に挙げた CRF-base, CRF-base+H, hBiLSTM-CRF, hBiLSTM-CRF+H の 4 モデルの訓練及び評価を 5 分割交差検定にて実施した。hBiLSTM-CRF, hBiLSTM-CRF+H については、ウィンドウ幅  $w$  の値を 11,21,31 の複数の設定で実験を行った。F 値 (マクロ平均) について、Randomisation test ( $\alpha=0.05$ , two-tailed,  $R=100,000$  samples) を用いて検定を行った。

### 5.2 実験結果

表3に各モデルの性能を示す。全モデル中で hBiLSTM-CRF+H( $w=11$ ) が最も高い F 値 ( $F=0.654$ ) を示し、CRF-base( $F=0.630$ ) 及び CRF-base+H( $F=0.632$ ) のベースラインモデルに対して統計的に有意な差を示し

表 3: 各モデルの性能 (マクロ平均値)

モデル	精度	再現率	F 値
CRF-base	0.681	0.603	0.630
CRF-base+H	0.685	0.605	0.632
hBiLSTM-CRF ( $w = 11$ )	0.663	0.635	0.647
hBiLSTM-CRF ( $w = 21$ )	<b>0.686</b>	0.629	0.651
hBiLSTM-CRF ( $w = 31$ )	0.673	0.615	0.638
hBiLSTM-CRF+H ( $w = 11$ )	0.679	<b>0.636</b>	<b>0.654</b>
hBiLSTM-CRF+H ( $w = 21$ )	0.657	0.628	0.640
hBiLSTM-CRF+H ( $w = 31$ )	0.653	0.620	0.633

表 4: 修辞役割別の各モデルの F 値

Category	hBiLSTM-CRF	hBiLSTM-CRF+H
BACKGROUND	0.319	<b>0.341</b>
CONCLUSION	0.415	<b>0.449</b>
FACT	<b>0.890</b>	0.879
FRAMING-main	0.642	<b>0.651</b>
FRAMING-sub	<b>0.527</b>	0.474
IDENTIFYING	0.798	<b>0.806</b>
OTHER	0.972	<b>0.975</b>

hBiLSTM-CRF は  $w = 21$ , hBiLSTM-CRF+H は  $w = 11$ .

た. 同様に, hBiLSTM-CRF( $w=21$ )( $F=0.651$ ) も CRF-base( $F=0.630$ ), CRF-base+H( $F=0.632$ ) に対して有意に高い性能を示した. 概して階層型 RNN ベースのモデル群は CRF ベースのモデル群よりも高い分類性能を達成した. このことから, 日本語で書かれた判決書においても, 文レベルでの BiLSTM-CRF 適用による文ベクトルの文脈化と予測系列の最適化が有効といえる.

F 値のマクロ平均値に関しては, hBiLSTM-CRF と hBiLSTM-CRF+H の間に統計的な有意差は確認できなかった. しかし, 修辞役割別の F 値 (表 4) に着目すると, BACKGROUND ( $F=0.341$ ), FRAMING-main ( $F=0.651$ ), CONCLUSION ( $F=0.449$ ) について hBiLSTM-CRF+H が有意に高い性能を示していることが確認できた. これらは結論や重要な論拠を提示する修辞役割であり見出し周辺に出現する傾向があるため, hBiLSTM-CRF+H モデルへの見出し情報の追加が有効に働いたと考えられる.

## 6 おわりに

本稿では日本国判決書に対する修辞役割分類について, 階層型 RNN の適用と判決書中の見出し情報の利用による性能改善を提案した. 階層型 RNN を用いることで分類性能が向上すること, 及び重要な修辞役割について見出し情報の利用することで分類性能が向上することを示した.

## 参考文献

[1] Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. Annotation of argument structure in Japanese legal documents. In *Proceedings of the 4th Workshop on Argument*

*Mining*, pp. 22–31. Association for Computational Linguistics, 2017.

- [2] Ben Hachey and Claire Grover. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, Vol. 14, No. 4, pp. 305–345, 2006.
- [3] M. Saravanan and B. Ravindran. Identification of Rhetorical Roles for Segmentation and Summarization of a Legal Judgment. *Artificial Intelligence and Law*, Vol. 18, No. 1, pp. 45–76, 2010.
- [4] Hiroaki Yamada, Simone Teufel, and Takenobu Tokunaga. Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artificial Intelligence and Law*, Vol. 27, No. 2, pp. 141–170, Jun 2019.
- [5] Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, Vol. 28, No. 4, pp. 409–445, 2002.
- [6] Atefeh Farzindar and Guy Lapalme. LetSum, an automatic Legal Text Summarizing system. *Jurix*, pp. 11–18, 2004.
- [7] 阪野慎司, 松原茂樹, 吉川正俊. 機械学習に基づく判決文の重要箇所特定. pp. 1075–1078. 言語処理学会, 2006.
- [8] Di Jin and Peter Szolovits. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3100–3109, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
- [9] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. Obligation and prohibition extraction using hierarchical RNNs. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 254–259, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [10] V. D. Tran, M. L. Nguyen, K. Shirai, and K. Satoh. An approach of rhetorical status recognition for judgments in court documents using deep learning models. In *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 1–6, Oct 2019.
- [11] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. Identification of rhetorical roles of sentences in Indian legal judgments. In *Proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX2019)*, pp. 3–12, December 2019.
- [12] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, Vol. abs/1508.01991, , 2015.
- [13] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [14] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.