機械学習に基づく判決文の重要箇所特定

阪野 慎司† 松原 茂樹‡ 吉川 正俊†

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

banno@dl.itc.nagoya-u.ac.jp

1 はじめに

近年，判決文の電子化に伴い，判決文の要約に関する研究が諸外国で進められている [1, 2, 7]．日本においては，最高裁判所判例集 [3] や LEX/DB データベース [4] などで大量の電子化された判決文がインターネット上で閲覧可能であるものの，その要約研究はほとんど行われていない．判決文の 1 つの要約といえる判例は，裁判官，検察官，弁護士など司法の専門家に対して極めて重要な情報である．例えば，裁判においては，判例は各自の主張を行うときの根拠としてしばしば引用される．しかし，過去数十年にわたる判決文が存在しており，また日々増加する文書群であるため，目的の判例を探すことは容易ではない．このため，判決文には判例の効率的な検索，理解を支援するために要旨が付与されている．要旨の作成は，現在，専門家により人手で行われておりその負担は小さくない．

本稿では，要旨の内容に該当する箇所を判決文から自動的に抽出する手法を提案する．要旨の内容が記述されている判決文の「理由」から，判決文の各文に対する特徴を抽出する．特徴は，形態素情報や文の長さなどの基本的な言語情報と，判決文の種類などの判決文に特有な情報を組み合わせる．抽出した特徴を用いて機械学習を行い，要旨の内容が記述されている文 (重要文) を獲得する．獲得された重要文，及び，機械学習結果を用いて最終的な重要文を決定する．得られた重要文から接続詞など，要旨に不必要な表現を削除し，重要箇所を特定する．本手法の有効性を評価するために判例コーパスを対象に重要箇所特定実験を行った．

2 判例と判決文

2.1 判例

判例は文献 [5] によると「裁判の理由の中で裁判所の示した法律上の判断」と定義されている．判例は，その裁判の事実関係の中で，争われている論点に対する法律上の判断でなければならないため，常に判決文から読み取られる必要がある．判例は，憲法や法律とは違い法令として認められていないが，これらの法令の条項には述べられていない具体的事例に対する法律上の判断を行っていることから「事実上の法律」とみなされている．実際の裁判においては，他の法令と同様に過去の判例を引用することにより，裁判官，検察官，弁護士などの専門家は自己の意見を主張している場合もある．それゆえ，過去に行われた裁判で示された判例は，全て今後行われる裁判で参照される可能性がある．

件名 不当利得返還請求事件（最高裁判所 平成15(オ)386、平成15(受)390 第二小法廷・判決 破棄差戻し）

原審 H14.11.28 東京高等裁判所 （平成14(ネ)1142）

主 文

原判決を破棄する。
本件を東京高等裁判所に差し戻す。

理 由

上告代理人及川智志外102名の上告受理申立て理由について
1 原審が確定した事実関係は，次のとおりである。
(1) 上告人は，貸金業の規制等に関する法律(以下「法」という。)3条所定の登録を受けて貸金業を営む被上告人との間で，平成7年5月19日，上告人が被上告人から手形割引，金銭消費貸借等の方法により継続的に信用供与を受けるための基本的事項について合意した(以下，この合意を「本件基本契約」という。)。上告人は，被上告人に対し，本件基本契約の合意内容を記載した「手形割引・金銭消費貸借契約等継続取引に関する承諾書並びに限度付根保証承諾書」を差し入れ，その後，被上告人からの借入金の増額に伴い，5回にわたり，上記書面とほぼ同一内容の書面を作成し，提出した。被上告人は，これらの書面の提出を受ける都度，上告人に対し，その写し(以下「本件各承諾書写し」とい

図 1: 最高裁判所判例集の判決文

2.2 判決文

判決文は，ある事件の裁判について書かれた文書であり，事実，理由，及び，主文など様々な情報が記載された文書である．以下，本研究で対象とするインターネット上で公開されている最高裁判所判例集の判決文について説明する．最高裁判所判例集の判決文の例を図 1 に示す．

最高裁判所の判決文は，図 1 のように「件名」,「原審」,「主文」,「理由」などの要素から構成されている．判例が記述される「理由」は,「主文」にたどり着くまでの過程が記述される.「理由」の内容，及び，記述される順序は，図 2 に示す通りである．基本的な判決文の流れとしては，図 2 の通りに記述される．図 2 において，追加，反対，捕捉意見は傍論と呼ばれ，判例が直接記述されないことが知られており [5]，判例を探すときの大きな手がかりの一つとなる．

図 1 に示すとおり,「理由」は文章で記述され，その文量が非常に多い判決文も存在する．この大量の文の

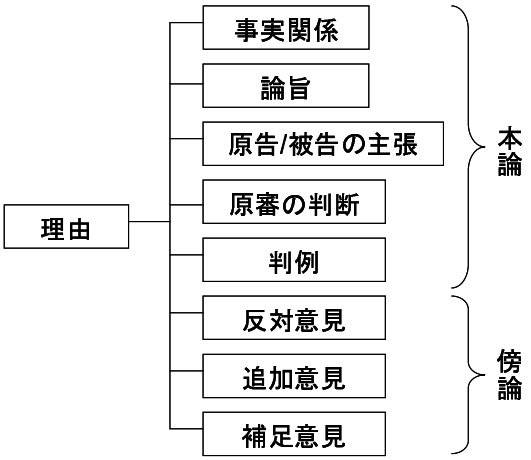

図 2: 理由に記述される内容

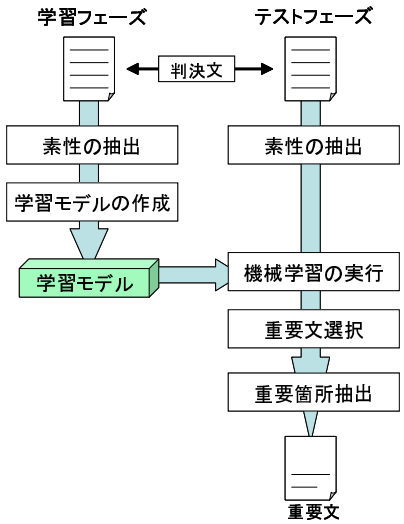

図 3: 処理の流れ

中から判例を検索したり，理解するには時間がかかるため，予め判決文には，付加情報が専門家により人手で付与されている．最高裁判所判例集の判決文には，判決日などの基本情報，判示事項，要旨，参照・法条という情報が付与される．これらの付加情報の中でも，要旨は，「理由」から判例に該当する内容をまとめたもので，重要な付加情報の一つである

3 機械学習を用いた重要箇所抽出

本手法の概要を図 3 に示す．本手法では，判決文の「理由」の各文に対して重要文を抽出するための素性を獲得する．獲得した素性を使用して機械学習を行い，各文に対するスコアを算出する．得られたスコアを基準として，各判決文に対して重要文を選択する．得られた重要文から不必要な表現を削除し，重要箇所を抽出する．

以下，本手法で使用した素性，重要文選択方法，重要箇所特定方法について順に述べる．

3.1 選択素性

素性は，一般的に使用される言語情報に判決文というドメインを考慮した判決文に特徴のある素性を組み合わる．判決文に特化した素性を用意することにより，優れた重要文の抽出が可能となる．

形態素

本手法では，原則として全ての形態素を素性の候補とする．判決文の各文に対して形態素解析を行い，各形態素の基本形をそれぞれ 1 つの素性とする．しかし，全ての形態素を素性とするのは適切ではない．形態素の中には，句点，読点などのようにほとんど全ての文で出現する形態素や，人名や地名などの固有名詞のように出現頻度が極めて少ない形態素が存在する．出現数が非常に多い形態素は，各文の特徴付けにほとんど意味をなさない．また，出現数が極めて少ない形態素は，その文を特徴付けるには有効であるが，その他の文で出現する可能性はほとんどないため，素性として選んでも効果はあまりない．本手法では，これらの不適切な形態素を素性から削除するために，学習データに対する総出現数を用いて出現数の少ない形態素を除去し，情報検索における idf 値に相当する isf(inverse sentence frequency) 値を利用する．形態素 t に対する isf 値は次の式で求められる．

$$isf(t) = \log \frac{N}{sf(t)} + 1$$

なお，N は学習データの総文数を，$sf(t)$ は形態素 t の出現する文数を表す．形態素 t の総出現数 $TF(t)$ は，閾値 K 以上のものを獲得対象の形態素とし，$isf(t)$ は，各文ごとにソートし，値の高いものから割合 C 以内の形態素を獲得対象とする．

数詞

算用数字や漢数字は，判決文において使用法が決まっている．算用数字は，最近の判決文における裁判に関係する数値上のデータ (個数，金額，日時等) を記述するときに使用する．漢数字は，それ以前の判決文におけるデータの記述や，熟語の一部として使用されている．これらの形態素は，一般的に事件の背景や，判断を下す上での材料として使用されるため，重要文にはあまり出現しない．重要文かどうかを判断する一つの材料としての利用が考えられるが，各数字を形態素として分割する必要はない．そのため，これらの数詞をそれぞれ統合して一つの素性として扱う．

文の長さ

要旨の内容を含む文は，法律上の判断が記述されているため，ある程度の長さを持つ．極端に短い文は，重要文になり得る可能性が低いと考えられるため，こうした文と区別をする．ある文 S に含まれる総形態素数を $Total(S)$ とすると，

$$Length(S) = \begin{cases} 1 & (Total(S) \geq 12) \\ 0 & (otherwise) \end{cases}$$

として素性を与える．なお，閾値となる形態素数 12 は，判決文データの分析に基づいて定められる．

文の位置

重要文が出現する位置はその文書の性質によって変化することが知られている．判決文においても基本的な文書構成が定められており，重要文が出現する位置には一定の傾向が見られる．判決文の要旨に該当する文は，本論においてのみ出現することが判明している．従って，本論におけるその文の相対的な出現位置を計算し，その割合で 4 分割し，それぞれ素性を付与する．

本論/傍論

判決文には，その裁判で中心として扱われる争点と扱われない争点が存在する．中心となる争点に関係して記述される本論と，そうではない傍論とで判決文の本文を大別することができる．判決文の要旨に該当する内容は，本論の位置にのみ出現することが知られており，本論かどうかの素性を与えることにより，これらの区別を行う．

判決文の種類

判決文はその結論に基づいて複数の種類に分類することができる．判決文データの分析から判決文の種類に応じて重要文が記述される位置の傾向が異なることが明らかになっている．判決文は大きく分類して 4 種類 (棄却，破棄自判，破棄差戻，却下) 存在する．実際の判決文は，これら 4 種類の組み合わせから構成されており，4 種類それぞれに素性を与えることにより，判決文の種類に関する情報を与えることができる．

3.2 重要文選択

重要文の選択は，次の 2 つの条件に従う．

1. SVM が正例と判定した文は重要文である．
2. 判決文中に正例がない場合は，SVM の結果の上位 M 個を重要文と認定する．

すなわち，本研究では，SVM の出力結果に基づいて重要文の選択を行う．SVM が正例と判断した文については，無条件に重要文であるとする．SVM が負例と判断した文に関しては，その文が含まれる判決文全体の状況に応じて重要文と判断する．本研究では，判決文には必ず要旨箇所が記述されている文が存在するものと仮定しているため，SVM の結果である判決文に重要文が含まれない場合には，SVM の出力結果の値の上位 M 文をその判決文の重要文とする．このようにして得られた重要文を各判決文におけるシステムの出力結果とする．

3.3 重要箇所特定

得られた重要文から要旨生成に不必要な内容を削除することにより重要箇所を特定する．前節までで得られた重要文には，接続詞をはじめとした実際の要旨には全く出現しない表現が含まれる．このような不必要な表現が削除されないまま残ると要旨の生成が困難になる．不必要な表現を削除するために，判例コーパスから観察される特徴から規則を作成する．

判例コーパスの重要文，及び，重要箇所から以下の特徴が観察された．

- 重要箇所と不必要な表現との境界には，読点が存在する．
- 不必要な表現は文頭からある読点まで連続して存在する．
- 丸括弧などの中で記述される補足内容は，要旨には記述されない．
- コーパスの各データによって，重要文に含まれたり含まれなかったりする表現が存在する．

上記の特徴に基づき，重要箇所特定の手順を以下に説明する．まず，要旨に含まれない丸括弧などの箇所を削除する．重要箇所と不必要な箇所の境界は，ほとんどが読点であるため重要文を読点で分割する．不必要な箇所は重要文の先頭から重要箇所の先頭までの間に存在し，重要箇所の途中に現れることはない．そのため，文頭から読点ごとに，削除するかどうかの判断を行う．しかし，重要箇所に含まれる表現の中には，要旨に必要ない表現も含まれている．本手法では，これらの表現も削除対象とみなし，削除するための規則を作成する．削除規則は，コーパスにおける節末表現の出現頻度，及び，品詞情報を用いる．

4 評価実験

4.1 実験概要

本節では，重要文抽出手法について評価する．構築した判例コーパスの判決文を用いて重要文抽出実験を行った．判例コーパスは，最高裁判所判例集の判決文を対象に構造情報と要約情報を付与したもので，1989 年から 2004 年までの 624 判決文が XML 文書形式で保存されている．判例コーパスの 624 判決文を訓練データ 574 判決文，テストデータ 50 判決文に分割した．テストデータには全部で 1,739 文が含まれ，そのうち重要文は 139 文である．実験で用いたパラメータはそれぞれ，単語の出現頻度の閾値 $K = 20$，isf 値の閾値 $C = 0.67$，重要文選択数 $M = 2$ とする．なお，形態素解析には ChaSen[9] を，SVM には SVM_Light[6] を使用した．SVM に使用するカーネル関数は次式を使用した．

$$(x \cdot y + 1)^2$$

実験の評価には，精度，再現率，F-値を使用する．各評価指標はそれぞれ以下の式で求められる．

$$Precision(P) = \frac{Count}{Sys\_Out}$$

$$Recall(R) = \frac{Count}{I\_Sen}$$

表 1: 実験結果

	精度 (%)	再現率 (%)	F-値 (%)
言語情報のみ	46.8	31.4	37.6
本手法	61.2	37.4	46.4

$$F\ measure = \frac{2 * P * R}{P + R}$$

ここで，*Count* はシステムが出力した重要文のうち正解となる文の総数を，*Sys_Out* はシステムが出力した重要文の総数を，*I_Sen* はテストデータに含まれる重要文の総数となる．

4.2 実験結果

機械学習の素性として，形態素，文に関する言語情報のみを用いた場合と，判決文に特有の情報を言語情報に関する素性に追加した本手法との間で比較を行った．結果を表 1 に示す．表 1 から，判決文特有の情報を追加することにより，精度，再現率ともに向上し，F-値で 8.8%向上した．言語表現からは判別できない文に対する分類性能が向上し，言語的特長の少ない重要文をより多く抽出できたといえる．

生成された重要文を見ると,「したがって、一個の債権の一部についてのみ判決を求める旨を明示して訴えが提起された場合において、当該債権の残部を自働債権として他の訴訟において相殺の抗弁を主張することは、債権の分割行使をすることが訴訟上の権利の濫用に当たるなど特段の事情の存しない限り、許されるものと解するのが相当である。(Ｈ１０．０６．３０ 第三小法廷・判決 平成６（オ）６９８ 不当利得)」における「したがって」や「相当である。」など重要文に頻出する表現 [8] が出現する文が多い．このような重要文に典型的な文は抽出可能であるが，複数の重要文が 1 つの要旨を構成する場合の途中の文など，あまり特徴が見られない文の抽出はそれほどできていない．その文から得られる情報だけでは，こうした文を取得するのは難しく，その前後の文から得られる情報を素性とする必要がある．

また,「すなわち、金銭債権の数量的一部請求訴訟で敗訴した原告が残部請求の訴えを提起することは、特段の事情がない限り、信義則に反して許されないと解するのが相当である。(同判決文) 」などの文のように，重要文に頻出する表現を用いた傍論の文に対して正しく分離できており，判決文情報の効果を確認できた．

5 おわりに

本稿では，機械学習に基づく判決文から要旨に該当する箇所を抽出する手法を提案した．判決文から形態素や文長などの言語情報，及び，判決文に特有な情報を素性として用いることにより判決文に特化した学習を可能とする．機械学習の結果を用いて要旨に該当する文を選択することにより，より多くの重要文を抽出できる．得られた重要文から不適切な表現を削除することにより，重要箇所を特定する．

今後の課題としては，重要文抽出性能の向上と重要箇所抽出の評価が挙げられる．現在の方法では，形態素に対する素性数が多く，適切でない形態素が多く含まれている．従って，これらの形態素を削除したり，同じ性質を持つ形態素を統合するなどして素性の洗練を行い性能の向上を図るつもりである．判決文に特有な素性としても，判決文の役割 (事実関係，判例など) に応じて与えられる素性や，前後の文との関係から得られる情報を追加する予定である．重要箇所抽出に関しては，重要箇所抽出単体，及び，全体の評価を行う予定である．また，SVM 以外の機械学習手法を用いた評価についても今後進める予定である．

謝辞 本研究を進めるにあたって，貴重なご意見を頂いた名古屋大学大学院法学研究科の松浦好治先生，Frank Bennett 先生，角田篤泰先生ならびに，大学院情報科学研究科の外山勝彦先生，小川泰弘先生に感謝いたします．本研究の一部は，科研費基盤研究 (B)(2) によります．

参考文献

[1] Ben Hachey and Claire Grover: Automatic Legal Text Summarisation: Experiments with Summary Structuring, Proceedings of the 10th International Conference on Artificial Intelligence and Law, pp.75-84 (2005).

[2] Ben Hachey and Claire Grover: Sequence Modelling for Sentence Classification in a Legal Summarisation System, Proceedings of the 2005 ACM Symposium on Applied Computing, pp.292-296 (2005).

[3] 最高裁判所判例集：http://courts.go.jp/

[4] LEX/DB データベース：
http://www.tkclex.ne.jp/index.html

[5] 中野 次雄．判例とその読み方，有斐閣，(2002).

[6] Thorsten Joachims. Making large-Scale SVM Learning Practical: Advances in Kernel Methods - Support Vector Learning, B.Scholkopf and C. Burges and A. Smola (ed.) MIT-Press (1999).

[7] Atefeh Farzindar and Guy Lapalme: LetSum, an automatic Legal Text Summarizing system, homas F. Gordon (ed.), Legal Knowledge and Information Systems, Jurix 2004: the 17th Annual Conference, pp.11-18 (2004).

[8] 阪野 慎司，松原 茂樹，吉川 正俊：手がかり表現に基づく判決文の自動要約，言語処理学会第 11 回年次大会発表論文集，pp.193-196 (2005).

[9] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara : Japanese Morphological Analysis System ChaSen version 2.2.1. (2000).