刑事判決書の特徴表現パターン抽出に関する複数手法の検討

Experimental Investigation of the Utility of Semi-syntactic Analysis for Phrase Pattern Extraction from Judicial Decisions

千本 達也[†1] 山本 大介[†2] 竹内 和広[†1] 三島 聡[†3]
Tatsuya Senbon Daisuke Yamamoto Kazuhiro Takeuchi Satoshi Mishima
†1 大阪電気通信大学 情報通信工学部 情報工学科, 寝屋川市初町 18-8
Osaka Electro-Communication University, 18-8 Hatsumachi, Neyagawa, Osaka
†2 大阪電気通信大学 工学研究科 情報工学専攻
Osaka Electro-Communication University, Graduate School of Engineering
†3 大阪市立大学大学院 法学研究科
Osaka City University, Graduate School of Law

あらまし:刑事判決書を収集し、その文書集合から特徴的な表現パターンを抽出することを検討する。具体的には、文の句構造を中心とした要素を解析し、文書群すべてから、その機能表現をノードとして、木構造に組織化する。その上で、木に対して複数の枝刈りといった編集をすることにより、当該パターンが特徴的か否かを検討する。また、木を表現し編集するデータ構造について、データ量、編集における計算効率性の観点から工夫を行った。

Summary: In this paper, we propose a data structure to store the various superficial syntactic sequences, which reflects the combination of content expressions and functional expressions in a certain set of specialized documents. We investigate some algorithms that extract characteristic patterns from the previous judicial decisions with the data structure. As a consequence of the experiment on the limited data, we confirm the efficiency of the extracted patterns from the viewpoint of information extraction and retrieval.

キーワード:情報検索, 表層的統語解析, 機能表現
Keywords: information retrieval, superficial syntactic analysis, functional phrases

1. はじめに

本稿では、裁判員裁判において適正・妥当な量刑審理・量刑判断を確保するための基礎資料を提供することを目的とし、既存の刑事裁判決書の蓄積から、複雑な条件下で検索をする、あるいは、情報抽出した結果を計量的に分析するための基礎技術として、自然言語で記述された刑事判決書を収集し、その文書集合から特徴的な表現パターンを抽出することを検討する。

具体的には、表層的な統語構造におけるパターンの組み合わせをできるだけ広く、現実的なデータ構造を用いて保存しておき、その保存データに基づいて特定の文書集合に特徴的な表現パターンを抽出する複数手法を提案する。本稿では、上記データ構造の実装を行い、実際の刑事裁判の判決書をデータとして、それぞれの手法により抽出された表現パターンが類似情報抽出や情報検索といった応用上、どのような利点があるかを検討したい。

2. 文構造の表層的解析

2.1 文型パターン

専門文書に対して高度な検索やテキストマイニングといった処理を行うためには、形態素解析の辞書や係り受け解析のモデルの調整といった言語処理基盤の再整備・調整が課題となってきた。

実際、日本語の統語解析の一つとしてよく用いられる係り受け解析モジュールの cabocha[1)]を法律文書に対して適用した場合、単語に関しては登録されていな

い単語が多く、また、独特の長い文を持つ文体から、係り受けの解析間違いも多い

他方、統語的な文解析の結果表示には、特定の統語構造のパターンを類型化した文型を利用した表現方法もある。この文型と呼ばれる概念は、特に日本語学習の分野では盛んに用いられている。例えば、日本語能力検定では、学習者があやまりがちな表現の出現環境を同様のパターンによって整理した書籍が販売されている。

文型は一般的には、次のような例文における、「N は、V」、「N1 に N2 を送る」といった形の、特徴的な文構造を表示する。先のカギ括弧内に示したパターンを本稿では文型パターンと呼ぶ。

「太郎は、花子に花を送る」

一般化すると、文型パターンは、N, V といった意味的あるいは統語的性質で抽象化可能な非終端記号と文構造の特徴付け要因となる文書中に実現された実文字列の組み合わせによって記述できる。後者の実文字列は多くの場合、機能表現と呼ばれる表現である。

文型パターンは、文構造の特徴を示すため、パターン内の係り受け構造と整合性を持つ。例えば、「N は、V」、「N1 に N2 を送る」を例にすれば、係り関係はそれぞれ、{(N は、→V)}, {(N1 に→送る), (N2 を→送る)}である。この様に、文型パターンが合致する文の端的に文構造を示している。

文型パターンは、パターン中の非終端記号の抽象度が様々である。例えば、「N1 に N2 を送る」の「送る」という動詞は、「贈る」「あげる」「プレゼントする」といった動詞に代えた場合も、同様の文構造的な特徴をもちうる。このような当該部分に当てはまる複数の対象を抽象化した非終端記号を恣意的に導入するため、非終端記号の性質が統語的なもの、意味的なもの、それらが複合的にからみあったもの、さらには談話的な性質をもつもの、といったように抽象化のレベルや基準が様々である。このことは抽象化に対応する語の集合を規定するという形は可能ではあるが、抽象化の理由は自然言語による記述に頼らざる得ない部分があり、人間が文特徴を把握することを志向した文解析の表示方法といえる。

2.2 機能表現の定義

野田 [2]は、形態素解析用の電子化辞書 Unidic[3] に収録されている、接続詞・連体詞・助動詞・助詞・名詞-助動詞語幹・形状詞-助動詞語幹、及び、松吉ら [4]が編纂した日本語機能表現辞書に収録されている表現を整理した文型パターン分析用の機能表現の集合を定義した。松吉らの日本語機能表現辞書は、「からすると」や「ざるを得ない」のように、国立国語研究所の研究では複数の短単位から構成され機能的な働きをする長単位を、助詞・助動詞型に属する複合辞として扱っている。これらを複合辞の活用や音韻的変化について調査・整理し、機能語に対して認定される表現を、機能型に属する機能表現と定義している。そして、この定義に従い、『日本語表現文型:用例中心・複合辞の意味と用法』[5]、『使い方の分かる類語例解辞典 新装版』[6]に収録されている表現から、341 種の見出し語をもつ 16,771 個の出現形について整理を行い辞書化した。この野田が整理した機能表現(F)を扱う。また、名詞・動詞・形容詞などの、機能表現と対となるものを内容表現(C)とする。そして、ある文における内容表現と機能表現の並びを CF 系列とする。

2.3 CRF による機能表現系列解析

機能表現系列を解析するタスクは、一般に系列ラベリング問題と呼ばれる。ここで、いくつかの要素が連なったものを系列と呼び、系列内のそれぞれの要素にラベルを付けることを系列ラベリングと呼ぶ。系列ラベリングは、ある文では動詞として機能する単語が、直前に形容詞が来た場合には、名詞として機能する場合があるなど、ある要素のラベルが系列内の他のラベルに依存するような場合を扱う。

ここで、機能表現系列を特定するという系列ラベリング問題を処理するために、条件付き確率場(Conditional Random Fields, CRF)[7]を利用する。CRF による識別モデルは形態素解析 [8]や未知語抽出 [9]などの系列ラベリング問題において、隠れマルコフモデル(Hidden Markov Model, HMM)による生成モデルや MEMM(Maximum Entropy Markov Model)よりも高い精度で解析できることが報告されている。他にも、固有表現抽出 [10][11]などでの利用例があり、多くの言語処理タスクにおいて利用される機会が増えている。

本稿では、CRF を用いて各節中の文字系列に対して、各文字に対応するラベルをラベリングする。ラベルの種類は、C と F の 2 種類である。CRF の実装は、CRF++[8]を用いて野田が整備したツールを使用する。

機能表現		begin	を	ていた
内容表現の文字数		7	3	0
内容表現の文字種	漢字	1	1	0
	ひらがな	1	1	0
	カタカナ	0	0	0
	数字	0	0	0
	アルファベット	0	0	0

図 1：CF 系列の抽出

3. 提案手法

3.1 文の解析処理

本節では、刑事判決書の文型的特徴を持った表現の型を機械的に抽出するための前処理として、CF 系列を抽出する。ただし、CF 系列中の内容表現は、実文字列を文字するのではなく、各機能表現にセットとなる内容表現の最大文字列数や文字種などの情報を付与することで、抽象化して扱う。これは、文の持つ機能的な役割の特定と、各機能表現間に入りうる内容表現を制限するためのものである。

入力された節から、文型パターンを抽出するための前処理として、CF 系列を抽出する手続きを以下に示す。

1. 刑事判決書中には記号を伴った箇条書きや特殊な記法によって記された事件番号、裁判の日付などが存在し、これらが以降の解析時にノイズとなるため、テキストから取り除くか、代替記号に置換するクリーニングと呼ばれる作業を行う。また、これ以降の解析において、文を句読点で区切られた節として解析するために、句読点で文を区切り、それぞれ別の系列として扱う。それから、句読点を削除した。さらに、括弧等の記号で囲まれている文中の機能表現は、それらを埋め込まれている文の機能表現の系列中に含んでしまうと以降の解析時に上手く処理することができないため、正規表現を利用して抽出し、別の節として処理を行った。
2. CRF を用いて、節中の機能表現系列を解析し、節を機能表現と内容表現に分解する。
3. 機能表現系列の先頭に begin というダミーの機能表現を追加する。これは以降の解析において、機能表現の先頭情報を保持するためのものである。これにより、機能表現の系列は末尾から見て、内容表現と機能表現の組の繰り返しである CF 系列となる。各機能表現には対となる内容表現の文字数と文字種の情報が付与される。文字種は、内容表現中に特定の文字種が出現するならば 1、そうでないならば 0 というようにフラグの形で情報を保持する。

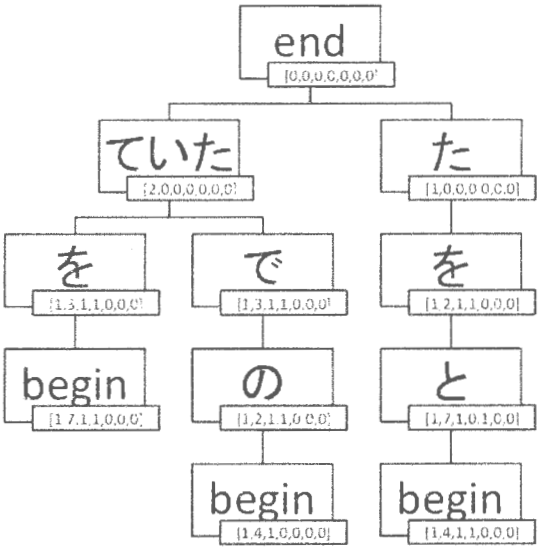
以上、一連の流れで、例として「繰り返し揉め事を起こしていた。」という入力節に対し、前処理を施した様子を図 1 に示す。処理により得られる CF 系列（内容表現に関する情報は省略）は「ていた-を-begin」となる。

3.2 文解析結果の保存法

刑事判決書の各文に対して、3.1 節で述べた前処理を行い、抽出した各 CF 系列から文型パターンを抽出するために、各 CF 系列を一つに統合した CF 系列木として構造化する。各 CF 系列を木として構造化したのはデータサイズの圧縮と、次節において、文型パターンを抽出する際の手続きのためである。木の各ノードは機能表現であり、異なる CF 系列をまとめる際、節末を含む共通の機能表現の部分系列を持つ場合は、それらを一つに統合した。

複数の CF 系列に対して、それらを一つ木構造にするための手順を以下に示す。

1. まずは機能表現の系列に着目した際に、同一の機能表現系列を持つ CF 系列を一つに統合する。そして、統合した CF 系列の数を、その CF 系列の出現数として付与する。ただし、CF 系列を統合する際、各機能表現に付与されている内容表現の情報については、文字列数は大きい方の値を保持し、各文字種については、論理和をとる。
2. 木の根として、end という根ノードを作成する。これは、各 CF 系列の末尾に相当する。
3. 各 CF 系列の末尾の機能表現から順に、end ノードに、再帰的に各機能表現をノードとして追加する。末尾から見た各 CF 系列の並び順は、木における階層と対応する。例えば、end が 0 番目の階層ならば、各 CF 系列の末尾の機能表現は、end の子として追加されるので、1 番目の階層となる。
4. 機能表現をノードとして追加する際、同時に機能表現とセットとなっている内容表現に関する情報と、各 CF 系列に 1.で付与した出現数を、そのノードの属性として持たせる。具体的には、出現数（weight）と木における階層（level）、そして、内容表現の文字数（length）、内容表現の文字種のフラグとして、漢字（kanji）・ひらがな（hira）・カタカナ



※ [weight, length, kanji, hira, kata, num, alpha]

図 2：CF 系列木の抽出

(kata)・数字(num)・アルファベット(alpha)を持たせる。また、ノードを追加する際、すでに同じ機能表現を持つ、既存ノードが存在する場合は、これらの情報に単に上書きせずに、それぞれ更新を行う。weight は追加ノードのものを既存ノードのものに加算、length は値が大きい方を保持、文字種については、それぞれ論理和をとる。

以上の一連の処理の流れで、以下の各節に対し、生成した CF 系列木を図 2 に示す。各節の機能表現は下線で表した。

(1) 繰り返し揉め事を起こしていた。
(2) 被害者方の離れで生活していた。
(3) 強い憤りと精神的ストレスを感じた。

3.3 特徴的文型パターンの抽出

CF 系列木を用いて、刑事判決書の各節に対して、文型パターンを抽出した。文型パターンは、一つの節に対して複数抽出される。

文型パターンの抽出方法について、図 3 に示す CF 系列木と入力節を用いて解説する。これは解説のために擬似的に設定したものである。丸で囲まれた数字は内容表現、アルファベットは機能表現をそれぞれ表す。図 5 のように、E ノードを根とする木を生成する。これは、文型パターンを抽出するための CF 系列木とは異なる木である。この木を文型パターン木とする。文型パターン木から、CF 系列木と整合がある当該文書群に特徴的な文型パターンを抽出する。E ノードを、図 4 に示す 3 つの木の成長規則を用いて成長させていく。ここで、S は任意時点での処理済みの部分木、X は任意の追加ノード、＊は全機能表現に対して整合を持つ追加ノードである。各成長規則を以下に示す

a) S に X を追加する。ただし、CF 系列木中の S の子に X が存在しない場合は、成長を止める。
b) S に＊を追加し、＊に X を追加する。ただし、CF 系列木中の S−＊の子に X が存在しない場合は、成長を止める。
c) S に X を追加せず、level と文中の対象文字を一つ進める。

ただし、機能表現系列の一致だけではなく、CF 系列木中の対応するノードが持つ内容表現に関する情報と整合が取れなければ成長を止めた。具体的には、文型パターン木を成長させる際、解析対象節中のある機能表現と対となる内容表現の文字数が、CF 系列木中の対応する機能表現が保持する文字数より大きいならば、成長を止めた。また、文字種についても同じく、解析対象節中のある機能表現と対となる内容表現の文字種を、CF 系列木中の対応する機能表現が保持していなければ、成長を止めた。これは、各機能表現間に入る内容表現に制限をかけるためである。

最終的に、図 5 における level:5 の B ノードのように、CF 系列木において対象の機能表現の子に begin、つまり、その機能表現の直前の内容表現から、CF 系列が始まるという情報が CF 系列木中に、存在したものを文型パターンとして抽出する。また、各ノードには、3.2 節と同じように、入力節の内容表現の情報をそれぞれ持たせた。

4. 実験

4.1 文型パターン被覆率

刑事判決書 1,171 件の約 640,000 節において、各節に対して CF 系列を抽出し、CF 系列木を生成した。CF 系列木を生成するまでの流れを図 6 に示す。そして、まとめた CF 系列木を用いて、前述の約 640,000 節から、判決書の発行年月が古いものから約 1,000 節を取り出し、各節に対して複数の文型パターンを抽出した。実験に用いた刑事判決書の情報を表 1 に示す。各節に対して抽出された文型パターンの一部を表 2 に示す。各節の下線部は機能表現を表す。そして、各節の文型パターン①は、対象節の全機能表現を並べた状態、つまり CF 系列に相当する。文型パターン②は機能表現数が①に比べて少なく、文型パターン③は、＊を含む文型パターンであり、＊は一定の CF 系列と合致す

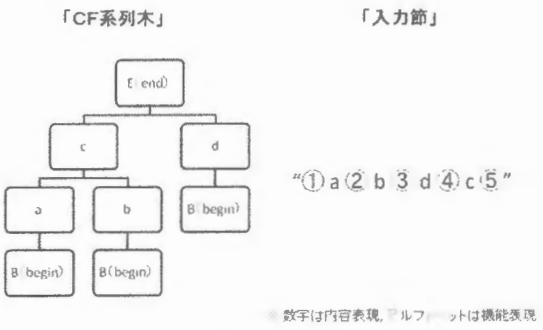


図 3：CF 系列木と入力節

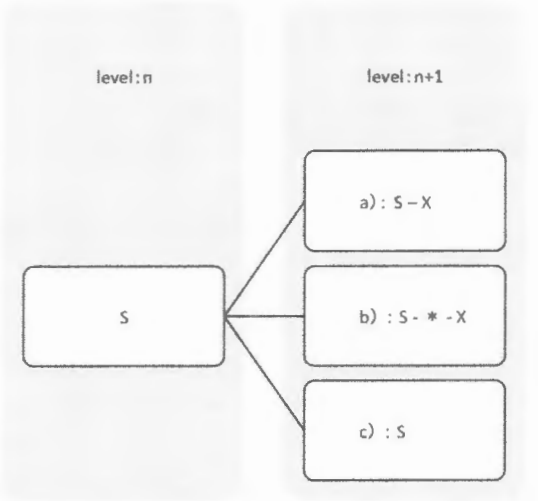


図 4：文型パターン木の成長規則

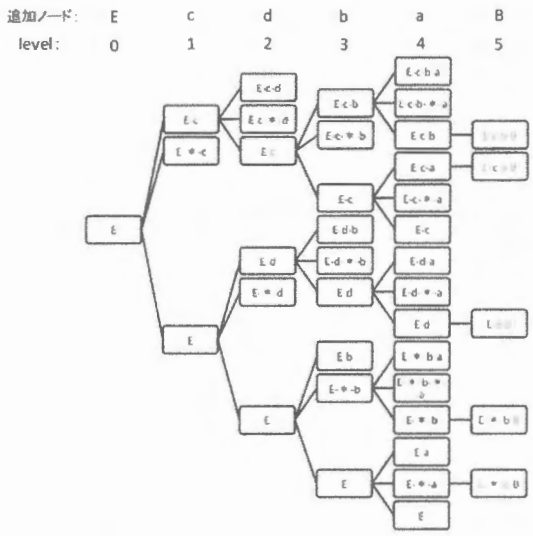


図 5：文型パターン木の成長過程

る。これは、部分木の抽象化に相当する。また、文型パターンは簡易化のために、文型パターン数と各機能表現と対となる内容表現に関する情報などを省略して、CF 系列のみを記載している。

抽出した全文型パターンにおいて、整合を持つ節が多いものについては頻出する表現であると考えられる。そこで、n 個の機能表現からなる文型パターンについて、2 つ以上の節と整合を持つ被覆率 P を算出した。P の算出方法を以下に示す。ここで、ただ 1 つの節と整合を持つ文型パターンの数を PN1、2 つ以上の節と整合を持つ文型パターンの数を PN2 とする。

$$P = \frac{P2}{P1 + P2}$$

機能表現数毎の文型パターンの被覆率を図 7 に示す。n = 1〜9 までの結果を出しているが、これは n = 10 以上の文型パターンが抽出されなかったためである。結果としては、n = 3, 4 の文型パターンだけで、全体の約 55%以上を占めており、被覆率 P についても、n = 2 についで高い。よって、刑事裁判決書の特徴な文型パターンの機能表現数は 3 または 4 で調整することが有効であることが推測される。

4.2 CF 系列木を枝刈りしたときの被覆率変化

CF 系列木に対して、複数の枝刈りを行い、そこから抽出された文型パターンの被覆率 P の変化を調査した。ただし、前節の結果を踏まえ、調査を行う文型パターンは n = 3, 4 とする。

CF 系列木に対して、以下の提案手法を用いて枝刈りを行う。各提案手法を閾値 m で適用した場合の CF 系列木から、n = 3, 4 の文型パターンを生成し、m = 1〜5 まで変化させたときの n = 3 と n = 4 の文型パターンの P の平均の変化を図 8 に示す。

A. あるノードに着目したときに、そのノードの各子の weight が m 以下の子をカットする。ただし、ノードが根である場合は処理を行わない。
B. あるノードに着目したときに、その親の weight を parentWeight とする。そして、各子の weight の下位 m 個の和が parentWeight よりも小さいとき、それらの子をカットする。ただしノードが根である、または子が一つしかない場合は処理を行わない。
C. あるノードに着目したときに、その親の weight を parentWeight とする。そして、各子の weight の下位 m 個の積が parentWeight よりも小さいとき、それらの子をカットする。ただしノードが根である、または子が一つしかない場合は処理を行わない。

実験の結果より、手法Cの閾値m = 3のときのPの値が他手法よりも高かった。被覆率Pは前述でも述べた通り、実験対象の全節のうち、2つ以上の節に整合する文型パターンの被覆率を表している。これは、多くの節に当てはまる頻出する文型パターン数を表しているので、手法Cは、文型パターン抽出に用いた節集合に対して、頻出する表現を多く保持することがわかった。

4.3 文型パターンによる内容表現の抽出特性

4.2節において、手法Cのm=3で枝刈りしたCF系列木から抽出した文型パターンを用いて、情報抽出と情報検索への応用を考えてみたい。情報抽出も情報検索も文型パターンによる内容表現の抽出が基本となる。

従来の内容表現抽出では、内容表現を抽出するため、ストップリストや形態素解析を用いるが、本稿の手法では、文の表層的な統語的枠組みである文型パターンと合致することにより、内容表現を抽出することができる。例として、表2の「被告人に逆らえなくなったというCの供述が不自然であるとはいえない」というテキストに、合致する文型パターン①および③を適用したときに抽出される内容表現を表3に示す。

表3が示すように、文型パターンに部分木の抽象部分を含まない時(文型パターン①との合致を利用するとき)、抽出された内容表現は、当該文の内容語となる形態素にほぼ相当する部分となり、形態素解析を使わずとも、すなわち、形態素解析辞書を使わず、機能表現を中心とする辞書のみを使って内容語抽出ができることを示している。

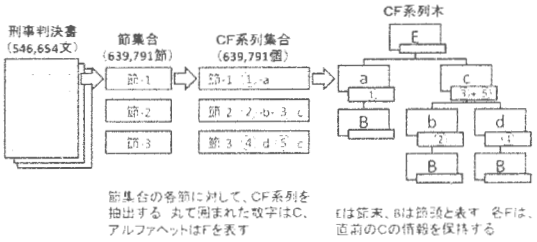


図6：CF系列木までの処理の流れ

表1:刑事判決書の情報

	文数	節数(CF系列数)
全判決書	546,654	639,791
CF系列木生成に用いた判決書	546,654	639,791
文型パターン抽出に用いた判決書	874	1,030

表2:抽出された文型パターンの例

節	文型パターン
被害者の告訴能力の有無の判断に直接影響するものではない。	①end-ではない-に-の-の-の-begin ②end-ではない-の-の-begin ③end-＊-の-の-の-begin
これに基づいて被告人の処罰を求めていると認められる。	①end-られる-ている-と-を-の-て-に-begin ②end-ている-と-を-の-て-に-begin ③end-ている-と-＊-に-begin
被告人に逆らえなくなったというCの供述が不自然であるとはいえない。	①end-ない-とは-が-の-た-という-なく-に-begin ②end-ない-が-の-た-という-なく-begin ③end-ない-とは-が-＊-に-begin

さらに、本手法では、上記の内容語抽出が、文型パターン①との連言条件で成立することを保証するように、内容語抽出の抽出条件を文型パターンの選択により制御することが可能である。表3の文型パターンに部分木の抽象部分を含む場合では、抽出できる内容表現相当部分を形態素に相当する単位だけではなく、波括弧内のような句レベルの文字列を内容表現として抽出可能であると同時に、このような多様な表現を抽象化して検索することが可能となる。

このような文型パターンの抽象化は、制約条件を無制限に緩和しすぎてしまうと内容表現を限定することが困難であるが、本稿の提案手法では、あらかじめ、当該文章集合の可能性あるCF系列の大部分はCF系列木により保存しており、内容表現の抽出制約を当該文書の特質に即して制御可能である。

また、同一文に複数の文型パターンを対応付けることが容易に可能である点を利用すれば、文を限定すれば、当該の文に合致する文型パターンを利用して「同じような書かれ方をする文」を検索することが可能であり、文型パターンと内容表現を複合的に用いた類似検索が検討可能と考えている。

4.4 今後の課題

本稿の提案の主眼は、自然言語文章をデータベース化する際のデータ構造に文型パターンを意識したCF系列によるデータ保存を行う点にある。現在の実装では、このCF系列のCRFに基づく解析とCF系列木の構築に関しては一定の効率的な実装を行えている

が、保存したデータの参照効率には課題が残る。特に、CF 系列からの文型パターンの抽出に関して非常に時間が掛かっており、多くの文章に一般的に出現する CF 系列を符号化しておき、今回の判決書などの専門性の高い、特定分野の文章集合に特徴的な CF 系列をあらかじめ区別して処理する工夫が必要である

具体的には、4.1節と4.2節の実験的検討により、少なくとも刑事裁判判決書の特徴な文型パターンの機能表現数は 3 または 4 で調整することが有効であることが予想され、この知見を生かして、CF 系列木を効率的に保存・参照する工夫を行っていきたい。

今回の検討では、文型パターンの抽象化に関して統計的な検討を行わなかったが、これは、上記の CF 系列の参照高速化の課題に依存性をもつ。また、内容表現を抽象化するための情報も、最大文字数、文字種といった限定的な情報を保持するに留まっており、文型パターンの抽象化について限定的なアルゴリズムを提案している点に課題が残る。

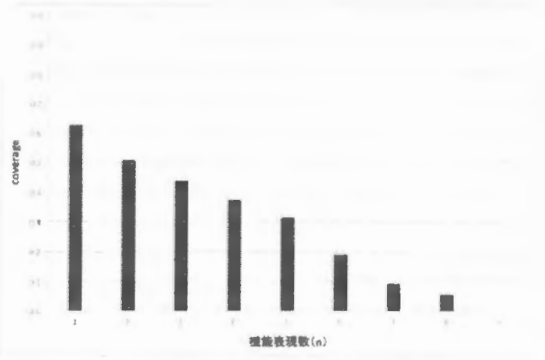


図 7：機能表現数毎の文型パターンの被覆率

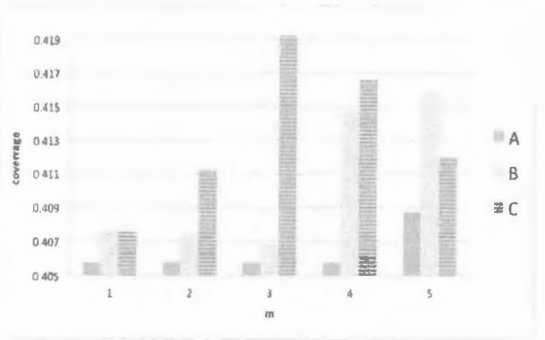


図 8：提案手法毎の被覆率の変化

表 3：文型パターンを用いて抽出した内容表現

文型パターン	内容表現
部分木抽象化なし	• 被告人 • 逆らえ • なっ • C • 供述 • 不自然である
部分木抽象化あり	• 被告人 **{逆らえなくなったというC}** • 供述 • 不自然である

5. まとめ

本稿では、刑事裁判例を計量的に分析するための基礎技術として、自然言語で記述された刑事判決書を収集し、その文書集合から、機能表現を中心とした辞書と CRF を用いることにより、機能表現と内容表現情報の組の繰り返しからなるCF系列を抽出しCF系列木として構造的にデータ保存をする方法を提案した。そして、CF 系列木に基づいて文型パターンを抽出するアルゴリズムを複数検討し、文型パターンといった表層的な統語構造を基準に、当該文章の内容表現の抽出を制御可能なことを実験的に確認した。

現状においては、比較的軽量な処理を実現できた自然言語文章の浅い統語解析に基づいて、構造的な情報を全文書に関して保存しておくことは可能であるが、その利用アルゴリズムを多様化する上での高速な参照を実現する実装が課題である。今後は、より高速な CF 系列木の実装に基づいて、複雑な類似判例の検索・クラスタリングといった応用処理を検討していきたい。

謝辞

本研究は、財団法人日本証券奨学財団の研究調査助成金「裁判員裁判において当事者の主張提示が量刑および判決に与える影響の検討」(研究代表・三島聡)による研究成果の一部である。同財団には謝意を表したい。

参考文献

1) 工藤拓, 松本裕治. Cabocha - Yet Another Japanese Dependency Structure Analyzer - Google Project Hosting (オンライン). 入手先 (http://code.google.com/p/cabocha/).

2) 野田奏. 分野非依存辞書に基づく多様な文章に対する文節分析とその応用. 大阪電気通信大学工学研究科情報工学専攻, 修士論文. 2012.

3) 国立国語研究所:特定領域研究 日本語コーパス 形態素解析辞書 UniDic(オンライン). 入手先(http://www.tokuteicorpus.jp/dist/).
4) 松吉俊, 佐藤理史, 宇津呂武仁. 日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123-146. 2007.
5) 森山良行, 松木正恵. 日本語表現文型:用例中心・複合辞の意味と用法. 株式会社アルク. 1989.
6) 遠藤織枝, 小林賢次, 三井昭子, 村木新次郎, 吉沢靖. 使い方の分かる類語例解辞典 新装版. 株式会社小学館. 2003.
7) J.Lafferty, A.McCallum, and F.Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the 18th International Conference on Machine Learning. 2001.
8) 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析. 情報処理学 会研究報告. 自然言語処理研究会報告, Vol. 2004, No. 47, pp. 89-96. 2004.
9) 東藍, 浅原正幸, 松本裕治. 条件付確率場による日本語未知語処理, 情報処理学会研究報告, 自然言語処理研究会報告, Vol. 2006, No. 53, pp. 67-74. 2006.
10) 福島健一, 鍛治伸裕, 喜連川優. コーパスからの固有表現辞書の自動構築, 知識ベースシステム研究会, Vol. 79, pp. 19-24. 2007.
11) 齋藤邦子, 今村賢治. タグ信頼度に基づく半自動自己更新型固有表現抽出, 自然言語処理, Vol. 17, No. 4, pp. 3-21. 2010.