構文情報タグ付き法律文コーパスにおける並列表現の分析とタグ付け誤りの修正

加藤　竜太[†]　　小川　泰弘[††]　　外山　勝彦[††]
名古屋大学工学部電気電子・情報工学科 †
名古屋大学大学院情報科学研究科 ††
{ryuta,yasuhiro,toyama}@kl.i.is.nagoya-u.ac.jp

1　はじめに

裁判員制度の施行やインターネット犯罪の増加などによって、法律に対する国民の意識は高まっている。法律文は通常の文に比べ複雑であるため、その内容を的確に理解するためには構文情報が有用であると考えられる。それに対して、構文情報を付与した法律文コーパスの作成基準が提案されている [1]。

本稿は、構文情報の中でも、特に法律文における並列表現を取り上げる。法律文における並列表現には、慣習に従った独特の言い回しが用いられており、従来の構文解析器 CaboCha は、そのままでは法律文の並列表現が持つ階層構造に対応できない。そこで、法律文の並列表現をパターン化し、そのうち CaboCha で正しくタグ付けできないパターンに対する修正方法を提案する。

本稿の構成は以下のとおりである。次の 2 章で法律文コーパス上で定義される係り受け関係を説明する。3 章でパターン分類の方法を述べ、続く 4〜7 章で各パターンを挙げ、最後にまとめと今後の課題を述べる。

2　法律文コーパスの係り受け基準

京都大学コーパス [2] では、係り受け関係を次の 3 種類に分類している。

1. 通常の係り受け関係
   主語・述語、修飾・被修飾、接続・被接続の関係
2. 並列関係
   意味上同等な語、句、節間の関係
3. 同格関係
   単に併置され、互いに意味を限定し合う関係

また、京都大学コーパスでは、通常の係り受け関係にはタグ D、並列関係にはタグ P、同格関係にはタグ A をそれぞれ付与している。

本稿では、名詞の並列関係に限って述べ、動詞の並列関係については、紙面の都合により割愛する。他の品詞の並列関係については、通常の文では多く見られるものの、今回調査した法律文中にはその例を見つけられなかったため、対象としていない。

法律文では、並列関係を表すために用いる接続詞として、主に「又は」「若しくは」「及び」「並びに」の 4 つが用いられる [3]。これらは法律文で並列関係を表す際に特徴的な役割を果たすため、本稿ではこれらを用いた並列表現を対象とする。すなわち、法律文においてはこれらの接続詞の使い分けによって階層構造を表現する。併合的接続の「及び」「並びに」については、最も内側の並列に「及び」を、それより外側には「並びに」を用いる。選択的接続の「又は」「若しくは」については、最も外側の並列に「又は」を、それより内側には「若しくは」を用いる。例えば、

例）国会若しくは県議会又はその委託を受けた議会

という並列表現は、

例）((国会若しくは県議会) 又はその委託を受けた議会)

のように括弧で示された階層構造をもつ。

さらに、同等のものを 3 つ以上並列させる場合には、

例）都庁、道庁、各府庁及び各県庁

のように最後の 2 つの間にのみ接続詞を用い、その他は読点を用いて連結する。

ところが、京都大学コーパスの基準では、並列関係を表すタグとしては、P 一つしか定められていないため、先に述べた階層構造をもつ並列関係を表現することができない。それに対して、法律文コーパス作成基準 [1] では、並列関係に対して P1、P2、P3、…というタグを付与し、数字が大きくなるほど外側の並列を表すこととしている。

このようなタグを用いると、並列関係は下記のように表せる。

例） 国会/若しくは/県議会/又は/その/委託を/受けた/議会

構文情報付き法律文コーパスを作成する際には、はじめに、対象とする各法律文に対して、CaboCha によって係り受けタグを付与する。このタグ付けを、作成基準 [1] に従って人手で修正することによって、最終的なコーパスを得る。

この修正を人手で行うためには、CaboCha の解析結果を 1 文ごとに確認しなければならない。しかし、並列表現のように表現方法が制限されているものについては、同様の修正を毎回施すこととなるため、修正作業を機械的に施すことが可能だと考えられる。

本稿では法律文の並列表現をパターン化し、同時にそれらに対する修正方法を提案する。この修正方法を実装することによって、人手による修正作業を軽減することが期待できる。

3　並列表現の分類と修正方法

本章では、法律文に現れるさまざまな並列表現を体系化し、パターンに分類する。

なお、以下では、含まれている自立語が名詞（または動詞）である文節を名詞節（または動詞節）という。

また、形容詞、形容動詞、連体詞などからなる修飾節については、それらが係り元や係り先となる並列関係は無いものと仮定し、パターン中の任意の場所に挿入されてもよいものとする。

さらに、句読点や接続詞は以下のように扱う。

句読点の扱い

句読点単独で文節は成さず、直前の文節に付属して 1 文節とする。

接続詞の扱い

CaboCha の解析では、接続詞は直前の名詞とともに 1 文節とするが、本稿では、法律文コーパス作成基準に従い、名詞からなる節と接続詞からなる節とに分割する。分割した節の係り先は分割前の係り先とし、名詞節からの係り受けにはタグ P1 を、接続詞節からの係り受けにはタグ D を付与する。例えば、CaboCha が

例） 弁護士又は/検察官

と解析した並列表現は、この分割処理によって、

例） 弁護士/又は/検察官

と修正する。なお、この分割処理は、以降で述べるパターンマッチングより前に行う。

次に、並列表現のパターンの記述方法について述べる。本章以降で述べる各パターンは、文節の並びと係り受け情報によって記述する。以下は文節の並びを表すために用いる記号である。

n	名詞節 1 つ
v	動詞節 1 つ
c	「又は」「若しくは」「及び」「並びに」のうち 1 つ
.	句点「。」または読点「、」 または箇条書き等における文末
、	読点「、」。これと上の「.」は、直前の文節に付属して 1 文節となる
$\vert$	この記号を挟むもののうちどちらか
+	直前の 1 回以上の繰り返し（欲張りマッチ）
*	直前の 0 回以上の繰り返し（欲張りマッチ）
S	$(n\vert v)*$ の略記
/	文節区切り

次の図はパターンの記述例である。

$$n_1 / c / S / n_2 / S / n_3 / S.$$

矢印で示される係り受け関係は CaboCha の解析によるものである。パターン中では、関係のない係り受け情報は省略する。例えば上の例では、n_3 からの係り受けの情報はパターンに含まれないので省かれている。

並列関係を表すタグについて、CaboCha では P が付与されるが、パターン記述上は P1 を用いる。また、以降では通常の係り受け関係に対するタグ D は記述を省略する。上述の例では、$c \to n_2$ にはタグ D が付与されていることとなる。

係り受け関係を示す矢印のうち、× 印があるものは、その関係がないことを示している。例では、$n_2 \to n_3$ に D という係り受けがないことを示している。

なお、パターンの図とともに、修正後の係り受け関係の図も示すが、紙面の都合上、一部を省いている。

パターンマッチングについては、文頭から走査し、並列関係の係り元と考えられる文節を見つけるごとにマッチングを行う。マッチした箇所に、次章で提案する修正を施し、文の残りの部分に対して、引き続きマッチングを行う。

以降の章では、まず 4 章で階層がない名詞節 2 つの並列、5 章で階層がない名詞節 3 つ以上の並列、6 章で階層がある名詞節の並列、7 章で人手の修正が必要な並列をそれぞれ扱う。最終的なパターンの数は、機械的な修正を施さないものを除くと、4 章で **3 つ**、5 章で **2 つ**、6 章で **4 つ**となった。

4　階層がない名詞節 2 つの並列

1-1) 名詞節が動詞節に係る場合-1

P1 $n_1/c/v^*/n_2/S/v_1/S.$ $\xrightarrow{\text{修正}}$ P1 $n_1/c/v^*/n_2/S/v_1/S.$

修正方法：n_1 と c の係り先を n_2 とする

例）故意/又は/過失によって/行い、

(例文中の点線の矢印は、修正前の係り受け関係を示す)

名詞節と動詞節が並列関係となる状況は、通常考えられず誤りである。パターン 1-1 は名詞節が係り先の動詞節より前に現れる場合で、この場合は n_1 と c の係り先を、接続詞の後に初めて現れる名詞節 n_2 へ修正する。

1-2) 名詞節が動詞節に係る場合-2

P1 $n_1/c/v_1/S/n_2/S.$ $\xrightarrow{\text{修正}}$ P1 $n_1/c/v_1/S/n_2/S.$

修正方法：n_1 と c の係り先を n_2 とする

例）第九条/又は/関連する/法案。

パターン 1-1 とは異なり、名詞節が係り先の動詞節より前に現れない場合は、v_1 が係る名詞節 n_2 を、n_1 と c の新たな係り先とする。

1-3) 名詞節が読点以後の文節に係る場合

P1 $n_1/c/v^*/n_2/S$、 $\xrightarrow{\text{修正}}$ P1 $n_1/c/v^*/n_2/S$、

修正方法：n_1 と c の係り先を n_2 とする

例）知事/又は/副知事について、/これを/任命する。

名詞節が接続詞の後の読点を跨いで並列関係となる場合は、後述の階層構造を除き誤りであるので修正する。

5　階層がない名詞節 3 つ以上の並列

本章では、「n_1、/n_2、/n_3/c/n_4」という表現の中の「n_1、」「n_2、」が並列関係の係り元となる場合を扱う。「n_3」以降は、そこだけを見れば接続詞を用いた 2 つの名詞節の並列であり、前節のパターンのどれかとなる。

なお、「名詞節＋読点」は必ずしも並列の一部ではなく、例えば「～場合、」は条件節にあたる。このように並列の一部とみなされなかった場合、CaboCha ではタグ D やタグ A が付与される。そこで、係り元「名詞節＋読点」に対して、CaboCha がタグ P を付与している場合を並列関係とみなしてパターン化する。

2-1) 読点直前の名詞節どうしの並列となっていない場合

P1 n_1、$/S/n_2$、 $\xrightarrow{\text{修正}}$ P1 n_1、$/S/n_2$、

修正方法：n_1、→ n_2、に修正し、P1 を付与する

例）議員、/議員の/家族、

読点直前の名詞節どうしが並列となるように修正する。

2-2) 読点直前の名詞節と接続詞直前の名詞節が並列となっていない場合

P1 n_1、$/S/n_2/c/S/(n/c/S)^*.$ $\xrightarrow{\text{修正}}$ P1 n_1、$/S/n_2/c/S/(n/c/S)^*.$

修正方法：n_1、→ n_2 に修正し、P1 を付与する

例）宝石、/それに/並ぶ/もの/又は/現金。

読点直前の名詞節と接続詞直前の名詞節が並列となるように修正する。

6　階層構造がある名詞節の並列

階層構造がある並列では、並列の階層をタグ P1、P2、P3、…によって表す必要があるが、この処理は 2 段階に分けて行う。まず、並列関係を見つけた場合、係り先の位置を確定し、暫定的に P1 を付与する。階層に関する処理は、4～7 章で述べているすべてのパターンに対するマッチングを経た後で行う。

階層構造における「及び」「並びに」などの接続詞の使い分け（2 章参照）は、係り先を確定する際に区別するため、階層に関する処理では接続詞の区別は必要ない。

以下、6.1 節で階層構造がある名詞節の並列のパターンを挙げ、続く 6.2 節で階層に関する処理について示す。

6.1　並列のパターン

3-1) 複数の接続詞を跨がないで名詞節と動詞節の並列となる場合

P1 n_1/若しくは$/S/v_1/S/$(若しくは | 又は)
P1 n_1/(及び | 並びに)$/S/v_1/S/$並びに

修正方法：パターン 1-1、もしくはパターン 1-2 と同じ

例）容疑者/及び/関係する/人物/並びに/それらの/親族。

名詞節と動詞節が並列関係となることは誤りであるので、v_1 の直前の S に名詞節が含まれていればパターン 1-1、そうでなければパターン 1-2 と同様の修正を施す。

3-2) 複数の接続詞を跨いでいて直後が「若しくは」や「及び」である場合

P1 n_1/若しくは$/v^*/n_2/S/$(若しくは | 又は)
P1 n_1/及び$/v^*/n_2/S/$並びに

修正方法：n_1 とその直後の接続詞の係り先を n_2 とする

例）身体上/若しくは/精神上の/障害/又は/疾病に/より、

2 つめの接続詞以降の文節に係るならば、それが名詞節、動詞節のどちらであっても誤りである。n_1 とその直後の接続詞の係り先を、接続詞の後に初めて現れる名詞節 n_2 へ修正する。

3-3) 複数の接続詞を跨いで名詞節と動詞節が並列となる場合

P1
n_1 / 又は /((S、)* / S / 若しくは)+ / S / v_1 / S.
P1
n_1 / 並びに /((S、)* / S / 並びに)* /(S、)* / S / 及び / S / v_1 / S.

修正方法：パターン 1-1、もしくはパターン 1-2 と同じ

P1
例） 施設の / 設置 / 又は / 物の / 製造 / 若しくは / 製造した / 物の / 販売。

名詞節と動詞節が並列関係となることは誤りであるので、v_1 の直前の S に名詞節が含まれていればパターン 1-1、そうでなければパターン 1-2 と同様の修正を施す。

3-4) 接続詞の跨ぎ方が不適切で直後が「又は」や「並びに」の場合

P1
n_1 / 又は /((S、)* / S / 若しくは)+ / v* / n_2.
P1
n_1 / 並びに /((S、)* / S / 並びに)* /(S、)* / S / 及び / v* / n_2.

修正方法：n_1 とその直後の接続詞の係り先を n_2 とする

P1
例） 虚偽の / 報告 / 又は / 検査の / 拒否、/ 妨害 / 若しくは / 忌避。

係り先を最後に現れた接続詞の後に初めて現れる名詞節 n_2 へ修正する。

6.2 階層に関する処理

本節では階層に関する処理について述べる。前述の通り、この処理の適用は全パターンのマッチングとそれに関する修正の後である。

1 つの文節が複数の並列関係の係り先となるとき

1 つの文節にタグ P1 を付与された複数の係り節があるときは、より離れた係り元との間の並列関係タグ Pi の数字 i が大きくなるように修正する。

~~P1~~ P2 P1
A / 又は / B / 若しくは / C

例えば、この表現では「A」→「C」のタグを P2 に修正する。

並列関係の係り先が別の並列関係の係り元となるとき

並列関係 Pi の係り元となっている文節 α に注目し、その文節の係り先の文節 β が別の並列関係 Pj の係り元となっていて、かつ α の直後が接続詞ならば、j を $i+1$ に修正する。例えば、

P1 ~~P1~~ P2
A / 若しくは / B / 又は / C

という表現では、α を「A」、β を「B」と見れば、「B」→「C」のタグ P1 は P2 に修正される。

一方、β が別の並列関係 Pj の係り元となっていて、かつ α が読点を含むならば、j を i に修正する。例えば、

P1 ~~P1~~ P2 ~~P1~~ P2
A / 若しくは / B、/ C / 又は / D

という表現では、まず先の修正によって、「B、」→「C」のタグが P2 に修正され、ここでさらに α を「B、」、β を「C」と見れば、「C」→「D」のタグ P1 は P2 に修正される。

7 人手の修正が必要な並列表現

CaboCha が係り受けを付与した下記の 2 文を考える。

P1
例） 辞退 / 又は / 内容の / 変更。
P1
例） 刑法 / 又は / 民法の / 適用。

前者については、「辞退」と「変更。」が、後者については、「刑法」と「適用。」が、それぞれ並列関係としてタグが付与されている。しかし、後者については、その意味を考えれば、「刑法」と「民法の」が並列関係であり、

P1
例） 刑法 / 又は / 民法の / 適用。

という係り受けタグを付与するのが正しい。このように、文節の並びが同じであっても、文の意味を吟味して係り先を判断しなければならない場合がある。この例では、後者の修正に際して、意味解析が必要となる。

本稿では、このような場合には機械的な修正をせず、人手による修正にまかせることとする。

8 まとめと今後の課題

本稿では、法律文コーパス作成基準に従って、さまざまな並列表現に対する CaboCha のタグ付けをパターン化した。また、各パターンに対する修正方法を提案した。今後はこれらのパターンのマッチング処理を実装し、実際に CaboCha のタグ付けを修正するツールを構成する。

人手による修正が必要となる並列表現に対して、修正作業を軽減するための方法を提案することは今後の課題である。

参考文献

[1] 山田将之, 小川泰弘, 外山勝彦：構文情報付き法律文コーパスの設計と構築, 言語処理学会第 14 回年次大会講演論文集, pp.604-607 (2008).

[2] 黒橋禎夫, 居蔵由衣子, 坂口昌子：形態素・構文タグ付きコーパス作成の作業基準 version 1.8, http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html.

[3] 法令解釈の基礎, 長谷川彰一, ぎょうせい (2002).