平成21年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	加 藤　竜 太
卒業研究題目	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	

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山田らによって提案されている。

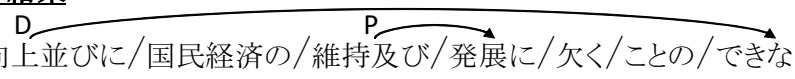
法律文コーパスを作成する際には、はじめに、対象とする各法律文に対して、構文解析器によって係り受けタグを付与する。山田らが提案した作成基準においては、構文解析器に CaboCha を用いる。CaboCha が付与する係り受けタグには誤りが含まれることがあるため、これを作成基準に従って人手で修正することによって、最終的なコーパスを得る。

本研究は、CaboCha のタグ付け誤りの中でも、特に並列表現に関する誤りに着目する。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、CaboCha では、法律文の並列表現が持つ階層構造に対応できない。その理由として、CaboCha を含む従来の構文解析器においては、係り受け基準の中に、階層の深さを表す方法が定められていないことが挙げられる。例えば CaboCha の係り受け基準では、並列関係に付与するタグとして、P 一つしか用意していないため、階層の深さを表すことができない。これに対して、法律文コーパス作成基準では、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。

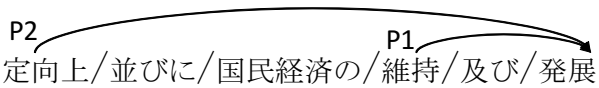
本研究は、CaboCha が付与した係り受けタグを機械的に修正することにより、法律文コーパス作成時における人手による修正作業を軽減することを目的とする。そのために、法律文の並列表現をパターン化し、そのうち CaboCha で正しくタグ付けできないパターンに対する修正方法を提案する。例えば、

例）国民生活の安定向上並びに国民経済の維持及び発展に欠くことのできない
という並列表現に対しては、　(エネルギー政策基本法第一条)

CaboChaによる解析結果



修正後



のように修正を施す。なお、タグ D は主語・述語関係などの通常の係り受け関係を表す。

また、提案手法を実装し、すでにコーパスが人手によって作成されている法令 8 本計 1,378 文に対して適用を試みた。その結果、本研究において対象とした並列表現全 625 個のうち、69 %にあたる 432 個に対して、本手法の適用によって正しい係り受けタグが付与されていた。これにより、本稿において提案した手法によって、人手による修正作業を軽減できることを確認した。

発表実績

- 加藤竜太, 小川泰弘, 外山勝彦：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, 言語処理学会第 16 回年次大会 (2010) (発表予定).