平成25年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	河 地　玄
卒業研究題目	法情報発信のための「法令のあらまし」の文書構造化	

近年の世界のグローバル化に伴い、国際取引の円滑化や在留外国人への対応などのため、日本の法情報を迅速に世界へ発信し、日本法の理解を促進する必要性が増大している。そのための取り組みとして、法令の内容を簡潔に説明した「法令のあらまし」の英訳を機械翻訳によって支援する試みが進められている。しかし、翻訳対象となる「あらまし」や、それを英訳した文書は、現時点では単なるテキストデータとしてしか得ることができない。日本法の理解促進という観点から考えると、「あらまし」を可能な限り伝わりやすい形で発信することは必要不可欠であると考えられ、付加的な要素が一切無いテキストデータはその意味で十分な形式とは言えない。伝わりやすい「あらまし」を実現する方法の一つとして、「あらまし」がもつ文書構造の情報を活用することが考えられる。

「法令のあらまし」はもともと論理的な文書構造を持っており、それによってその内容を体系付けることができる。その中の要素も、見出しや規定文・表など、いくつかの種類に分類することができる。そこで、このような論理構造のメタ情報を付け加えた「あらまし」を提供することができれば、その構造を利用して「あらまし」の参照、利用を容易にすることができる。さらに、構造情報を含んだ「あらまし」を提供することにより、その情報を再利用した様々な応用が期待できる。

本研究では、「法令のあらまし」が持つ文書構造を XML 文書型定義 (DTD) の形で定義し、それに基づいて実際の文書をマークアップすることにより、「法令のあらまし」の文書構造化を行った。

「あらまし」の文書構造を定義するため、本研究では「あらまし」文書の各行頭に現れる記号に着目して、その要素を調査した。「あらまし」では「一」、「[1]」、「(イ)」など、数字やカナ・括弧を組み合わせた見出し番号によって階層構造を表現し、規定文間の序列を提示している。また、表や注なども特定の記法で表現されており、それらを目印に各構造要素を見いだすことができる。本研究では、記号や数字・カナから始まる行を調査し、階層構造の見出し番号に用いられる 11 種類の表現を獲得した。さらに、それらの見出し番号を含んだ「あらまし」を人手で調査した。その結果、それらの見出し番号が現れる順序は、上位の階層では比較的厳密に決まっているが、下位の構造になるほど曖昧さが増し、例外的な記法が多くみられることを確認した。これらの知見により、階層構造を示す要素や、各階層での文や表・題名に関わる要素など、合計 43 個の構造要素からなる「あらまし」の DTD を作成した。

一方、毎年 100 本前後制定される新たな法律の「あらまし」をすべて人手でマークアップするためには莫大なコストを要する。このコストを軽減するため、「あらまし」の自動マークアップシステムを作成した。その際、Ruby で実装された構文解析器 Racc を用いて、行単位で「あらまし」の構造を解析した。また、番号の現れる順序に着目して、それぞれの「あらまし」における番号の付与ポリシーを特定するアルゴリズムを作成し、見出し番号の表記揺れに対応した。このシステムに対して、無作為に選んだ実際の「あらまし」20 本を入力として与えたところ、80%の文書を正しくマークアップすることができた。