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	坂 本　聡 美
卒業研究題目	法令用語対訳辞書拡充のための日本語見出し語抽出	

近年の国際社会のグローバル化により、日本法の迅速な国際的情報発信の需要は増大している。我が国の法令の外国語訳は、英語訳を中心として所管府省や民間などにより個別に取り組まれてきた。しかし、翻訳された法令の量と質、及び利便性に問題があった。これに対応するため、法務省は2009年4月より日本法令外国語訳データベースシステムを開設し、翻訳した法令と法令用語日英対訳辞書（標準対訳辞書）を公開している。標準対訳辞書では、統一的で信頼できる法令の英語訳が継続的に行われることを目指し、翻訳の基本的なスタンスを示すとともに主要な用語・言い回しについての日英対訳を示している。しかし、辞書の収録語は法令の専門用語に限られ、見出し語の量は十分ではない。対訳が示されていないものは翻訳者それぞれが翻訳を考える必要がある。そのため、訳語の不統一や翻訳者の負担増による翻訳の遅れといった問題の原因となっている。

本研究では、標準対訳辞書の拡充を目指し、辞書の見出し語となる日本語表現を抽出する。本研究における見出し語とは、法令文に出現する全ての表現である。また本研究では、これまで法令用語の抽出に使用されてこなかった英文官報に着目する。英文官報とは戦後占領期に発行されていた官報の英訳版で、公布される法令を含め日本語版と全く同じ内容が掲載されている。英文官報の中で実際に使用された表現は、見出し語に対してより適切な訳語を決定する際の候補として有効である可能性がある。そのため、見出し語の抽出に英文官報に対応する日本語法令を使用する。見出し語の抽出手法には、チャンカーYamChaによる抽出（チャンキング）と、文書頻度による指標のひとつであるdf_2/dfに基づく抽出を検討する。前者は、人手により形態素単位でチャンクタグを付与したコーパスからYamChaで学習を行い、浅い構文解析によりキーワードを抽出する。後者は、ある文書において一度出現した文字列が、もう一度繰り返し出現する度合い(df_2/df)をあらかじめコーパスから計算し、その分布から得られるキーワードの特徴を用いて文分割を行い、キーワードを抽出する。

本研究では、法令1,624本を対象に見出し語を抽出する実験をした。前処理として、法令コーパス中の文語文だけで構成される行と表組の行を排除した。これはチャンキングに用いる形態素解析器が文語文に対応していないこと、及び表組の行が抽出精度に悪影響を及ぼす危険性があることによる。YamChaの学習には、16本の法令にあらかじめ形態素解析を施し、形態素毎に名詞句・動詞句を示すタグを人手で付与したデータを使用した。文字列のdf_2/dfを計算する元のコーパスは、法令1,624本とWikipedia日本語版の200,000記事とした。

実験の結果、チャンキングでは90,709個（名詞句80,951個、動詞句9,758個）を抽出し、df_2/dfに基づく手法では69,657個を抽出した。標準対訳辞書の収録語のうち対象とした法令コーパス中に出現する見出し語2,242語と抽出した表現とを比較したところ、チャンキングでは約98%、df_2/dfに基づく手法では約71%をカバーした。またdf_2/dfに基づく手法では、辞書の見出し語となり得る慣用句が抽出できることを確認した。