平成29年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	植 原　リ サ
卒 業 研 究 題 目	単語の分散表現を用いた法令用語間の関係の獲得	

法令文には、一般的に馴染みのない用語が使われており、一般的に使われている用語でも、法令文特有の意味として使われることがある。例えば、法令文では「昇任」と「昇格」という二つの用語には明確な使い分けがなされているが、その知識がなければ法令文の意味を正確に理解することは容易ではない。それらの用語が、法令文を読みにくくする原因の一つとして挙げられる。

こうした用語間の関係を獲得することは、法令用語の意味を理解するための方法の一つである。単語間の関係を獲得するには、一般的にシソーラスや意味ネットワークが用いられる。シソーラスとは、単語の意味あるいは概念の関係を体系化した辞書のことである。意味ネットワークとは、単語の意味、概念間の連想関係、知識などをネットワーク形式で図式化して表したもののことである。しかし、既存のシソーラスや意味ネットワークには一般的･基本的な用語のみしか収録されておらず、法令用語や複合語は収録されていない。

そこで本研究では、分散表現を用いて法令用語間の関係を獲得することを提案する。分散表現とは、単語を高次元の実数ベクトルで表現したものであり、その一つに Mikolov らが提案した Word2Vec がある。$v(\text{king}), v(\text{man}), v(\text{woman}), v(\text{queen})$ をそれぞれ king, man, woman, queen の単語ベクトルとすると、Word2Vec では、$v(\text{king}) - v(\text{man}) + v(\text{woman}) \simeq v(\text{queen})$ のように、単語ベクトルの足し引きが意味の足し引きと同等に扱える性質である加法構成性を示すことができる。この加法構成性が単語ベクトル間で成り立つ用語の四つ組を得ることにより、法令用語間の関係の獲得を試みた。

本研究では特に、法令文中において、法令用語を定義する定義文に注目し、定義語と語釈語の関係を獲得した。ここで定義語とは、定義文で定義されている語句のことをいい、語釈語とは、定義語の語釈文の中で名詞句となっている語句のことをいう。その際、形態素解析によって 2 語以上となる定義語も存在するため、定義語を 1 語として扱えるようフレーズ処理を行った。フレーズ処理とは、二つの単語からなる一般的なフレーズをテキストから自動的に検出するものである。この処理を繰り返すことによって、複数の単語からなる定義語も 1 語として扱えるようになる。

本研究では、8,565 法令 (2,510,570 文) にフレーズ処理を施し、法令用語を分散表現で表した。単語の分散表現化モデルには、Word2Vec の CBOW モデル、Skip-gram モデルの 2 種類を用いた。次に、定義文から定義語 2,575 語 (異なり) とその語釈語延べ 35,449 語を抽出し、定義語と語釈語の単語ベクトルの差を計算した。その後、階層型クラスタリングを行い、同クラスタ内で加法構成性が成り立つ用語の四つ組を獲得した。なお、階層型クラスタリングには、Python で実装された ward 法を用いた。

その結果、CBOW モデルで 471 組、Skip-gram モデルで 423 組の用語の組を獲得した。そのうち、両方のモデルで獲得した用語の組は 85 組であった。また、獲得した二つの定義語の組は、291 組であった。実際に、「昇任」−「階級」+「級」≃「昇格」のような演算を構成する用語の組が獲得できた。これより、法令用語間の関係が取得できていることが確認できた。