平成28年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	藤岡 和弥
卒業研究題目	分散表現を用いた類似例規検索	

地方自治体は多くの条例や規則（例規）を制定している。例規は各自治体で策定されるが、その際には他の自治体の例規の中で類似したものを参照することがある。しかし、現在約 1,700 団体存在する自治体がそれぞれ多くの例規を定めている。その中から参考になる例規を人手で探すことは容易ではない。

このような背景のもと、角田らは、類似例規の参照作業を効率化することを目的に、eLen 条例データベース（以下、eLen）を開発している。eLen には自治体より提供された例規が XML 形式で収録されているとともに、収録された各例規の間の類似度が記録されている。この類似度は、単語単位での編集距離、すなわち、二つの単語列を一致させるために必要な単語の挿入・削除・置換といった編集操作の数によって計算されている。

しかし、編集距離を用いた類似度計算には二つの問題点がある。一つ目は、計算量が大きいことである。単語数がそれぞれ m, n である二つの単語列の間での編集距離の計算量は $O(mn)$ となる。eLen ではこの欠点を補うために、検索の度に類似度を計算するのではなく、スーパーコンピュータを用いて事前に類似度を計算しておくことにより、検索処理の高速化を図っている。二つ目は、単語の一致判定を表層形によって行うため、表記の揺れを吸収できない点である。例えば「基づく」という単語は「基く」と書かれることもあるが、これらは単語として異なるものとして扱われる。例規は各自治体が独自に制定しているため、このように表記が揺れていることが想定され編集距離では類似度を精度よく推定できない可能性がある。

そこで本研究では、分散表現を用いた類似例規検索を提案する。分散表現とは、単語や文を高次元の実数ベクトルで表現したもので、単語や文の間の類似性がベクトル空間上での二つのベクトルの距離（コサイン類似度）として得られる。分散表現への変換は、大量の例規データによって学習させたモデルを事前に生成しておけば、任意の例規文書に対して短時間で行うことができる。さらに、コサイン類似度の計算量は小さいため、検索の度にオンラインで類似度を計算しても、検索結果を高速に出力することができる。

本手法の有効性を検証するため、実際の例規データに対して本手法による類似例規検索を実施し、編集距離による類似例規検索の結果と比較する実験を行った。文書の分散表現化モデルには、Mikolov らの提案した Distributed Memory Model of Paragraph Vectors（PV-DM）を Python で実装したものを用いた。また、このモデルの学習では、例規 1,159,422 本分のテキストを形態素解析ツール MeCab で単語分割したものを入力として用いた。検索対象は、あるクエリに対して、それに類似した例規 30 本と類似していない例規 70 本からなるデータセットとした。本実験では、両手法でクエリと検索対象の各例規との間で類似度を計算し、これが閾値 α を超えている例規の集合をその手法における検索結果とした。

いくつかの α について、本手法と編集距離を用いた手法による実験を行った結果、本手法での最大の F 値は 0.931（$\alpha = 0.2$）となり、編集距離を用いた手法での最大の F 値 0.784（$\alpha = 0.1$）より大きい値となった。また計算時間については、例規の分散表現への変換が平均で 0.394 s、分散表現の間での類似度計算が平均で 5.18×10^{-5} s と、共に小さいことが確認できた。以上の点から、本手法の有効性が確認できた。