パブリックコメント投稿者を支援するための行政関係文書の分割

鈴木 愛加

近年，行政の透明化や情報共有，経済の活性化の観点から，公共データを広く公開することによるオープンデータへの取り組みが行われている．その中でも，行政関係文書は，電子化されていたとしても，堅苦しく長い文書で書かれていることから，パブリックコメント等を投稿する際にそのまま参考とすることは難しい．そこで，本研究では，行政関係文書を読む負荷を軽減し，より容易に内容の理解を促すことを目的とし，パブリックコメントを募集する際に提示される参考資料をトピックごとにパッセージ単位に分割する手法を提案する．

本研究での提案手法では，行政関係文書をトピックごとに分割する手法として，TopicTilingを採用する．TopicTilingは，従来手法である語彙的結束性に着目したTextTilingと比べ，LDA（潜在的ディリクレ配分）を活用し，単語に対してトピックIDを付与することにより，話題に焦点を当てた分割が可能となる．本研究では，行政関係文書を正しく分割するために，LDAによってトピックを抽出する．そのために，LDAの訓練データとして，政治行政分野の話題を全体的に把握でき，オープンデータとしても活用可能な白書データを使用する．また，白書の本文中の複数のトピックを区別するために，段落単位で区切られたパッセージを利用する．

本研究で採用したテキスト分割手法が有効である状況を明らかにするため，TopicTilingによる提案手法とTextTilingによる従来手法との比較実験を行った．分野ごとの差異をみるために，「環境」，「観光」，「情報・技術」の3つの分野を選択し，それぞれについて訓練用データセットを構築した．全ての分野に対する，本提案手法のTopicTiingによる分割手法，全ての品詞を手がかり語としたTextTilingによる分割手法，名詞を手がかり語としたTextTilingによる分割手法それぞれの最良のF値の平均をとった結果，それぞれ0.61，0.35，0.41となった．以上のように，LDAのトピック数とトピックに対する出力単語数を調整することで，本研究で採用したTopicTilingの手法の有効性を確認することができた．特にTopicTilingでは，話題が共通している単語に同じトピックIDを付与することができれば有効ということが分かった．一方で，分割されるべき境界の直前の文に，境界の直後に含まれる単語と共通のトピックIDが付与された単語があると，同じ話題と見なされ，境界を認識できない失敗例も見られた．

今後の課題として，行政分野を追加した比較実験を行う予定である．また，トピックIDを正確に付与するために，教師ありLDAを活用することによって，TopicTilingの改善を進めていきたい．さらには，パブリックコメント投稿者の情報要求に応えられるように，分割したパッセージを検索できるようなインターフェースを実現していくことを検討している．

（指導教員　関 洋平）