トピックモデルを考慮したセミマルコフ過程に基づくテキスト分割

北原　美穂

近年、インターネットの発達により、日々のネットワーク上のテキストデータの生成量は増加し、膨大な量が蓄積されている。特に、インターネット上のテキストデータは意味的に構造化されておらず、文書内に複数の異なる話題が混在していることが多い。このようなデータにユーザが効率的にアクセスできるようにするには、テキストを意味的なまとまりで区分するテキストセグメンテーションの技術が有効である。意味的な分割を全て人手で行うことは困難であるため、コンピュータにより自動的に行うことが望ましい。

コンピュータによる自動的なテキストセグメンテーションは、テキスト中に出現する単語の分布を手がかりにする方法で実現できる。しかし、単語の出現分布を用いた手法では分割対象の文書の語彙に近い学習用文書が必要になる上、学習する分布のベクトルの次元が高いことにより数も大量に求められてしまう。

この問題を克服する方法としてトピックモデルを用いた手法がある。単語の分布の代わりにトピックの分布に基づいて分割を行うことによって、学習に用いる文書の数が少なくても、十分に話題の区切りを推定できる。既存のトピックモデルを用いた従来手法であるTopicTilingは、学習用文書を使用せず、また少ない計算時間で高い精度を出すことができる。しかし、分割される正解のセグメント長に大きなばらつきがある場合に分割精度が悪化してしまう問題がある。そのため、本研究ではセグメント長のばらつきに依存せずセグメントを検出できる、隠れセミマルコフモデルによる動的計画法を採用した。

一方、動的計画法に基づくテキスト分割手法では、事前に与えるトピックモデルの学習単位の切れ目で分割してしまうために、外部の学習文書を利用せずにトピックモデルを素直に適用することが難しいという問題がある。トピックモデルで学習をする際は本来１文書を単位として学習を行うが、複数の文書が連なりその切れ目がわからないようなデータを対象とする場合、学習の単位について考える必要がある。しかしその問題について触れた研究は存在しない。そこで、本研究では分割結果をトピックモデルの学習単位に反映させて動的に変化させる方法を提案する。それにより、分割対象の文書のみで十分な学習を行うことが可能になる。

実験として動的な学習単位の変更による効果の検証と、それを取り入れた隠れセミマルコフモデルと他手法であるTopicTilingとの比較評価を行った。動的な学習単位の変更を取り入れることにより、正解セグメント中のトピック分布の局所的な変化を減らすことができ、精度の向上に成功した。また、TopicTilingとの比較では、特に正解セグメント長のばらつきの高いデータに対して分割精度を向上させることができた。

（指導教員　若林啓）