Reconstruction and Bayesian Nonparametrics Based Multi-Document Summarization

審 査 の 結 果 の 要 旨

論 文 提 出 者 氏 名　　馬　騰 飛

インターネットなどを介して膨大な文書が提供されている現在、ひとつのトピックに対してですら全ての文書を読むことは困難である。したがって複数文書の要約は与えられたトピックに関する多数の情報源から得た文書の全容を短時間で把握するために必須の技術である。単一文書の要約は定式化、評価方法とも確立しているが、多数の文書をまとめて要約することは技術的にも難しい。とくに文書集合が複雑かつ多様なトピック構造を持つ場合、その複雑さ多様さが要約長に反映するであろうという直観を数理モデル化し、要約アルゴリズムとして実現することが重要な課題である。本論文では、ノンパラメトリック・ベイズ手法を複数文書要約に適用することで解決を図っている。

本論文は「Reconstruction and Bayesian Nonparametrics Based Multi-Document Summarization」(再構成とノンパラメトリック・ベイズ手法に基づく複数文書要約)と題し、６章からなる。

第１章「Introduction」(序論)では、文書要約の定義と解決すべき問題を提示し、本論文で提案する解決方法の概要を示している。

第２章「Document Summarization Overview」(文書要約の概要)では、文章構成型、文抽出型など種々の形態の文書要約を概括し、要約アルゴリズム、評価方法などについて説明している。さらに本論文で対象とする複数文書を対象にする文書要約について詳細に記述している。

第３章「Multi-document Summarization using Minimum Distortion」(最小歪みによる複数文書要約)は、歪みの最少化を目的とする最適化問題として複数文書要約を定式化している。そこでは文の出現場所および線形順序に基づく再構成によって歪み最少化を図っている。

第４章「Topic Models and Bayesian Nonparametrics」(トピックモデルとベイジアン・ノンパラメトリックス) では、まず文書集合において多文書にわたるトピックを抽出するベイジアン・ノンパラメトリックモデルの有用性を説明している。次に階層型入れ子ハイブリッドディリクレ過程を提案し、この数理モデルの実装により既に提案されているベイジアン・ノンパラメトリックモデルよりも高いパープレキシティを達成できることを実験的に示している。

第５章「Bayesian Nonparametric Summarization and Summary Length Determination」(ベイジアンノンパラメトリックな要約と要約長決定)は、ベイジアン・ノンパラメトリックモデルを用いて複数文書要約における要約長を推定する手法を提案している。この手法はトピック構造の多様性に応じて要約長を自動的に推定する数理モデル化点である。自動推定された長さの要約が人手で与えた正解要約と高い類似度を持つことを実験的に示し、本論文における主要な貢献となっている。

第６章「Conclusion」(結論)は、本論文のまとめである。

以上を要するに、本論文は複数文書要約において扱われているトピックの多様性に適応した要約長を新規提案したベイジアン・ノンパラメトリックモデルによって精度よく推定できることを示すことによって、数理情報学分野の技術発展に寄与した。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。