

氏 名 (本 籍)	中 村 明 (愛知県)
学 位 の 種 類	博 士 (工学)
学 位 授 与 番 号	甲第 404 号
学 位 授 与 日 付	平成 23 年 3 月 25 日
専 攻	電子情報システム工学専攻
学 位 論 文 題 目	トピック適応言語モデルの高精度化と応用に関する研究 (Improvement of topic-based language model and its application)
学位論文審査委員	(主査) 教授 鎌 部 浩 (副査) 教授 速 水 悟 准教授 横 田 康 成

論文内容の要旨

近年、情報機器の普及とインターネットの飛躍的な発展に伴い、オフィス業務から日常生活に至るまで多くの文書が電子的に記録、保存、流通されるようになってきている。しかし電子文書の急速な普及は一方で、新たな情報格差を引き起こしかねないという問題点を孕んでいる。これは、電子文書を有効に活用するためにはアクセシビリティの向上（電子文書が容易に利用できること）が必要不可欠であるためである。

電子文書のアクセシビリティ向上には入力時と利用時の2つの側面があるが、本研究では主として入力時のアクセシビリティ、すなわちテキストの入力しやすさに焦点を当てる。そして、自然言語処理技術を活用して万人に使いやすいテキスト入力システムを実現することを目的として、入力語予測を用いたテキスト入力支援の高精度化を目指す。

自然言語処理の研究は、かつては文法規則や語彙等の言語知識を手で蓄積してシステムを構築していくアプローチが主流であった。しかしこの方法は拡張性に難があり、実用的なシステムの開発には多大な人的コストを要するという問題があった。これに対し近年では、自然言語を確率的現象ととらえ大量のデータから自動的もしくは半自動的にモデルを構築しようとする、統計的自然言語処理と呼ばれる研究が盛んである。

統計的自然言語処理の中核をなす要素技術のひとつである確率的言語モデルは、自然言語の統計的性質を確率論の枠組みでモデル化したものであり、今なお新しいモデルの提案や改良が試みられ発展しつつある研究分野である。近年では、従来からある N グラムモデルではモデル化し得ない大域的な依存関係をトピック（話題）としてモデル化する言語モデル、すなわちトピック適応言語モデルが各種、提案されている。トピック適応言語モデルでは、単語の生起確率が話題によって異なると考え、文脈から話題を推定しこれに基づいて適応的に単語の生起確率を算出する。

トピック適応言語モデルは単語間の大域的な依存関係をモデル化できる点で優れているが、以下の2点が課題として挙げられる。

- ・トピック数を増やすとモデルの記述能力が高まるが、ある程度のトピック数で性能が飽和する
- ・トピック適応の際に用いる単語列の長さ（文脈長）によって推定精度が変動するが、最適な文脈長の決定が困難

そこで本研究では、代表的なトピック適応言語モデルである LDA (Latent Dirichlet Allocation) を主な対象として、上記2つの課題に取り組みモデルを高精度化する。そして入力予測によるテキスト入力支援への応用を想定した評価を行う。

上記2つの課題の内、1点目は言語モデルに限らず有限個の事例に基づいてモデルのパラメータ推定を行う場合に不可避である過適応の問題である。既存のトピック適応言語モデルでは、通常、最も高精度なモデルが得られるトピック数を採用することが行われているが、これはモデルの精度を向上するものではない。これに対して本研究では、独立に学習した複数のモデルから得られた推定結果を集団学習の枠組みで統合することによって、モデルの高精度化・安定化を図る。そして新聞記事コーパスを用いた評価実験を通して、モデル規模が同程度の従来方式と比較して常に性能が向上することを示す。さらにテキスト入力支援に適した言語モデル評価指標 i -PP を新たに提案する。 i -PP は任意の入力読み文字数における平均単語分岐数を表し、入力支援システムにおいて操作系の仕様の影響を受けずに言語モデルの能力を測定することができる。この指標を用いた評価を通してテキスト入力支援における提案手法の有効性を示す。

2点目の課題はテキストから話題変化点を抽出するタスクとも関連する問題である。文脈長によってモデルの精度が変化する問題は従来から知られていたが、最適な文脈長を推定することが容易ではないため、先行研究の多くでは、文脈長を一定単語数とするか文書先頭以降の全単語を用いる方式が採られてきた。

話題変化点を確率的に推定して様々な文脈長からの予測を混合する方式も提案されているが、必ずしも十分な改善効果が得られるものではなかった。これに対して本研究では、話題変化を追跡することなく単語の予測に最適な文脈長をより直接的に推定する。提案方式では、現在の文の既知部分を最も精度良く推定できる文脈長を逐次求め、この結果に基づいてトピック適応を行い単語を予測する。そして、既存方式との比較を通して提案手法が予測精度と安定性の両面で優れていることを示す。

本研究で提案する方式は、テキスト入力支援に限らず、連続音声認識やテキストストリームからのキーワード抽出など他のオンラインアプリケーションにも応用が可能である。今後は提案方式のさらなる高精度化とともに応用範囲の拡大を図り、電子文書のアクセシビリティ向上を通して社会に貢献していきたい。

論文審査結果の要旨

本論文は、統計的自然言語処理の中核をなす要素技術のひとつであるトピック適応言語モデルを高精度化し、入力語予測を用いたテキスト入力支援に応用したものである。

自然言語処理の研究は、かつては文法規則や語彙等の言語知識を手で蓄積してシステムを構築していくアプローチが主流であった。しかしこの方法は拡張性に難があり、実用的なシステムの開発には多大な人的コストを要するという問題があった。これに対し近年では、自然言語を確率的現象ととらえ大量のデータから自動的もしくは半自動的にモデルを構築しようとする、統計的自然言語処理と呼ばれる研究が盛んである。

統計的自然言語処理の中核をなす要素技術のひとつである確率的言語モデルは、自然言語の統計的性質を確率論の枠組みでモデル化したものであり、今なお新しいモデルの提案や改良が試みられ発展しつつある研究分野である。近年では、従来からある N グラムモデルではモデル化し得ない大域的な依存関係をトピック（話題）としてモデル化する言語モデル、すなわちトピック適応言語モデルが各種、提案されている。トピック適応言語モデルでは、単語の生起確率が話題によって異なると考え、文脈から話題を推定しこれに基づいて適応的に単語の生起確率を算出する。本研究では、代表的なトピック適応言語モデルである LDA(Latent Dirichlet Allocation)を主な対象とした。

トピック適応言語モデルは単語間の大域的な依存関係をモデル化できる点で優れているが、2つの課題がある。第一に、トピック数を増やすとモデルの記述能力が高まるが、ある程度のトピック数で性能が飽和することである。第二に、トピック適応の際に用いる単語列の長さ（文脈長）によって推定精度が変動するが、最適な文脈長の決定が困難であることである。

上記の内、第一の課題は言語モデルに限らず有限個の事例に基づいてモデルのパラメータ推定を行う場合に不可避である過適応の問題である。既存のトピック適応言語モデルでは、通常、最も高精度なモデルが得られるトピック数を採用することが行われているが、これはモデルの精度を向上するものではない。これに対して本研究では、独立に学習した複数のモデルから得られた推定結果を集団学習の枠組みで統合することによって、モデルの高精度化・安定化を行った。そして新聞記事コーパスを用いた評価実験を通して、モデル規模が同程度の従来方式と比較して常に性能が向上することを示した。さらにテキスト入力支援に適した言語モデル評価指標 i-PP を新たに提案した。i-PP は任意の入力読み文字数における平均単語分岐数を表し、入力支援システムにおいて操作系の仕様の影響を受けずに言語モデルの能力を測定することができる。この指標を用いた評価を通してテキスト入力支援における提案手法の有効性を示した。

第二の課題はテキストから話題変化点を抽出するタスクとも関連する問題である。文脈長によってモデルの精度が変化する問題は従来から知られていたが、最適な文脈長を推定することが容易ではないため、先行研究の多くでは、文脈長を一定単語数とするか文書先頭以降の全単語を用いる方式が採られてきた。話題変化点を確率的に推定して様々な文脈長からの予測を混合する方式も提案されているが、必ずしも十分な改善効果が得られるものではなかった。これに対して本研究では、話題変化を追跡することなく単語の予測に最適な文脈長をより直接的に推定した。提案方式では、現在の文の既知部分を最も精度良く推定できる文脈長を逐次求め、この結果に基づいてトピック適応を行って、単語を予測する。そして、既存方式との比較を通して提案手法が予測精度と安定性の両面で優れていることを示した。

最終試験結果の要旨

本論文は、入力語予測を用いたテキスト入力支援に応用するために、統計的自然言語処理技術の中核をなすトピック適応言語モデルの高精度化に関するものである。トピック適応言語モデルでは、単語の生起確率が話題によって異なると考え、文脈から話題を推定しこれに基づいて適応的に単語の生起確率を算出する。トピック数を増やすとモデルの記述能力が高まるがある程度のトピック数で性能が飽和する点と、トピック適応の際に用いる単語列の長さの最適な決定が困難である点が課題であった。これに対して、独立に学習した複数のモデルから得られた推定結果を集団学習の枠組みで統合することによって、モデルを高精度化・安定化する方式を提案した。また話題変化を追跡することなく単語の予測に最適な文脈長をより直接的に推定する方式を提案した。

これらの研究成果は、学術論文、国際会議論文集に掲載されており、その内容は博士論文としてふさわしいものと考えられる。以上の点から、論文提出者は学位授与の基準を満たしていると判断し、最終試験の結果を合格とした。