公開

研 究 終 了 報 告 書

「談話構造に基づく教師なし生成型要約」

研究期間： 2019 年 10 月～2022 年 3 月
研 究 者： 磯沼 大
加速フェーズ期間： 2022 年 4 月～2023 年 3 月

1. 研究のねらい

近年、Web の発達と電子化された情報の飛躍的な増大により、自動文書要約のニーズが高まっている。例えば EC サイト上に投稿された大量の商品レビューを要約することで、消費者の選択や製品企画に有用な情報を与えられる。自動要約は大量の文書からの知識抽出を容易にし、人の情報収集・意思決定の効率化・質的向上に貢献する技術である。

自動要約のアプローチは、要約に相応しい文や節を抽出する抽出型要約と、単語や句の言い換え・一般化を行う生成型要約に分けられる。生成型要約はより人手に近い自動要約を実現でき、その確立は自動要約研究の大きな目標である。一方で、生成型要約は見本となる要約（参照要約）を大量に要し、現実の文書の多くは参照要約の数が少なく、それらの用意に多大な労力を要することから、実用上の大きな障害となっている。

そこで、本研究は教師なし生成型要約手法を開発することによって、見本の要約が不要な汎用文書要約技術の実現に挑む。教師なしアプローチでは、要約の潜在表現を参照要約なしにいかに獲得するかが鍵となる。本研究では文書の潜在構造、特にトピック構造に着目し、各トピックに関する要約文（トピック文）の潜在表現を得ることで、複数文で構成された要約を教師なしに生成する手法を開発する。例えば**エラー! 参照元が見つかりません。**に示したあるレビューの要約は、food、place、service といった各観点について、様々な粒度の評価を述べている。トピック木構造を捉えた上でトピック文を生成することで、多様な粒度のトピックで構成された要約を生成できると考えた。

教師なし生成型要約は萌芽的な研究であり、本研究課題の申請時点にて申請者の研究を含め 2 例のみが報告されている。そのうち、本研究は文書に潜在するトピック構造に着目することで要約を生成する初の試みである。生成された要約と得られたトピック木構造との評価により、文書に潜在するトピック構造を捉えることが教師なし要約生成に有用であることを明らかにする。

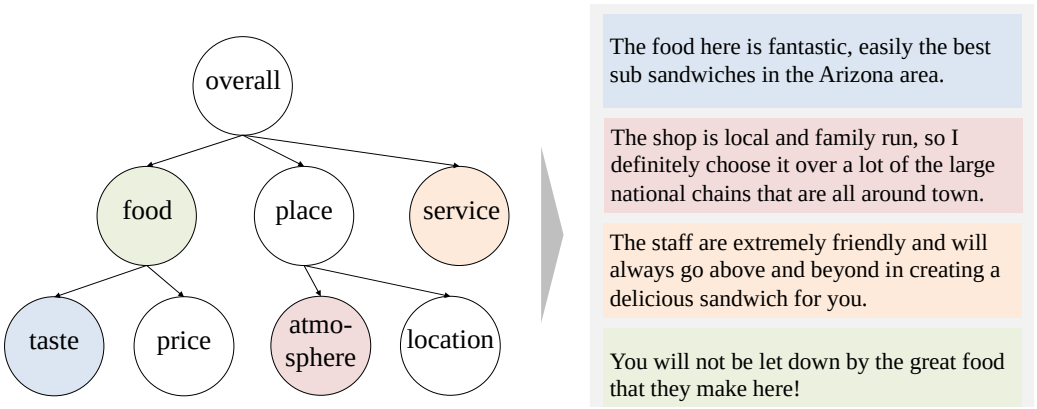


図 1 あるレストランレビューの要約例と対応するトピック木構造

加速フェーズでは、前述の研究成果を踏まえ、応用面/理論面にわたる自動文章要約技術の深化に取り組む。応用面では、従来自動文章要約が適用されてこなかった領域の開拓に取り組み、その一環としてサーベイ論文の自動生成に取り組む。サーベイ論文生成では、(1) サーベイ論文の対象分野における学術文献のトピック分類・構造化 (2) 各トピックにおける学術文献の要約が必要となる。(1)に関しては、従来の時系列トピックモデルを発展させた時系列構造化トピックモデルを開発し、学術文献のトピック分類及び学術トピックの分岐・統合の把握を可能にする。(2)に関しては、実際のサーベイ論文を用いて作成した教師データをもとに、サーベイ論文生成モデルを学習し、人が作成したサーベイ論文との比較評価を行う。これら一連の取り組みにより、サーベイ論文自動生成の可能性・課題について明らかにする。

理論面では、文章要約を非可逆データ圧縮として見做すことで、人手で作成された見本の要約に頼らない、文章要約の新たな定式化に取り組む。画像や動画、音声では教師データを用いずに汎用的なデータ圧縮を実現しており、文章要約もまた一種の非可逆データ圧縮と見做すことができる。そこで、本研究では (3)非可逆データ圧縮の基礎理論であるレート歪み理論に基づく文章要約の定式化に取り組む。本研究で提案する定式化に基づき学習されたモデルが文章要約を解くことができるか明らかにするとともに、前年度までに開発した手法をはじめとした一連の教師なし文章要約技術が、レート歪み理論においてどのように解釈できるか議論する。

2. 研究成果

(1)概要

本研究課題では以下の過程を経て、トピック構造に基づく教師なし生成型要約手法の開発を行った。

前半では大規模文書に適用可能な木構造トピックモデルの開発に取り組んだ。本研究課題では木構造トピックモデルを用いて要約を生成するが、要約生成の学習には大量の文書が必要な一方、既存の木構造トピックモデルは大規模な文書に適用が困難である。そこで複数の文書を並列に学習できる木構造ニューラルトピックモデルを開発することで学習時間を短縮し、要約生成など大量の文書を要するタスクでの利用を可能にした。提案法は従来法と同等の解釈性を持つトピックとその木構造を得ながら、学習時間を約 15 倍短縮し、要約生成などのニューラルモデルとの一体的な学習を可能にするなど、木構造トピックモデルの応用可能性を広げた。本研究を取り纏めた論文は計算言語学分野のトップ国際会議 ACL2020 に採択された。

後半では、トピック構造に基づく教師なし要約生成手法の開発に取り組んだ。前半で開発した木構造トピックモデルにより、木構造上の各トピックに関する要約文の潜在表現を獲得し、要約を生成する手法を開発した。得られた要約と人手で作成した要約を比較評価した結果、既存の教師なし要約手法より元文書の内容を網羅し、かつ一貫性の高い要約が得られることが確認された。また、文の詳細度合いはその潜在分布の分散の大きさに依存し、潜在分布の分散が大きいほどより一般的な文が生成されるという、要約のみならず文生成タスク全体に有用な知見が得られた。以上の成果を取り纏めた論文は、計算言語学のトップジャーナル TACL に採択された他、言語処理学会第 27 回年次大会で若手奨励賞を、情報処理学会第 246 回自然言語処理研究会で優秀研究賞および山下記念研究賞を受賞した。

加速フェーズでは、(1)時系列構造化トピックモデルの開発、(2)サーベイ論文生成用の教師データ作成及びサーベイ論文自動生成モデルの評価に取り組んでおり、後述するように既存手法では捉えられなかった学術トピックの分岐・統合や、サーベイ論文生成のモデル化が可能になった。また理論面について、(3) レート歪み理論に基づく文章要約の定式化では、要約モデルの学習過程でレートと歪みが小さくなることは確認できたものの、逆(レートと歪みの最小化により文章要約モデルが学習できること)を示すには至らず、文章要約の定式化において、レート・歪み以外の要素が必要であることが明らかになった。本研究は今後も継続的に取り組む予定である。

(2)詳細

- **前半:大規模文書に適用可能な木構造トピックモデルの開発**

本研究では木構造トピックモデルを用いて要約を生成するが、要約生成の学習には大量の文書が必要な一方、既存の木構造トピックモデルは大規模な文書に適用が困難である。そこで前半では複数の文書を並列に学習できる木構造トピックモデルを開発することで学習時間を短縮し、要約生成など大量の文書を要するタスクでの利用を可能にした。

従来の木構造トピックモデルにて事後分布推定に用いられている崩壊型ギブスサンプリングや平均場近似は、学習性能や並列化の困難さから、大規模な文書への適用が困難である。また、要約生成とトピックの事後分布を一体的に学習することで、要約として有用なトピックとその木構造が得られるとより望ましい。

そこで本研究では、文書からトピック分布への写像をニューラルネットワークにより構成した木構造ニューラルトピックモデルを提案した。提案法は variational autoencoder(VAE)の枠組みによる学習が可能であり、前述の問題を解決可能である。

既存研究は LDA などのフラットなトピックモデルに VAE を適用しているが、無限木上のトピック分布への関数を有限のパラメータでどう構築するかは自明でない。そこで本研究は、親子間と兄弟間それぞれに再帰的な構造を持つ doubly-recurrent neural networks (Alvarez-Melis and Jaakkola, 2017) を用いて木構造上の stick-breaking process を表現することで、文書から無限木上のトピック分布への関数を構築した。

評価実験にて、提案法は既存の木構造トピックモデル(nCRP; Blei et al., 2010)とほぼ同等の一貫性を持つトピックと木構造を得た(図 2)。一方、学習時間は約 15 倍短縮され、大規模な文書に適用できることが示されたほか(図 3)、要約生成といった下流タスクのニューラルモデルとの一体的な学習が可能になるなど、木構造トピックモデルの応用可能性を広げた。

研究成果を国内会議にて発表したほか**エラー! 参照元が見つかりません。**、本研究を取り纏めた論文は計算言語学分野の国際会議 ACL2020 に採択された**エラー! 参照元が見つか**

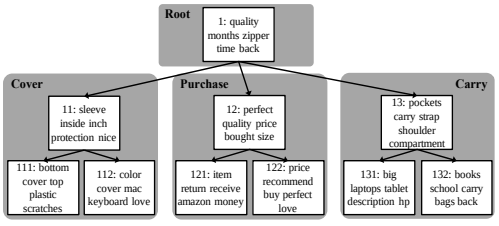


図 3 得られたトピック木構造

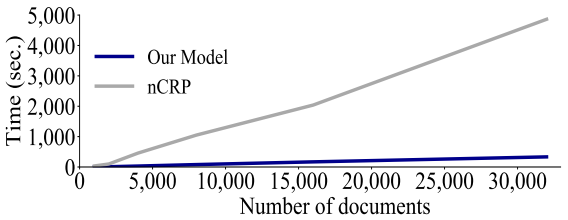


図 2 文書数(横軸)に対する学習時間(縦軸)

りません。。

・　**後半:トピック構造に基づく教師なし要約生成手法の開発**

本研究では、開発した木構造トピックモデルにより文書のトピック木構造を推定し、各トピックの要約文を生成する手法を開発した。トピック木構造では、文書から推定したトピックが木構造に配置され、子が親のサブトピックとなる構造を持つ。それらのトピックから要約として相応しい詳細度合いのトピックを選択し、各トピックに関する要約文を生成することで、意見文書の要約が教師なしに得られることを示した。

トピック文生成の文脈では、文書中の文の潜在分布を混合ガウス分布で表現することで、その構成要素である各単峰ガウス分布がトピック文の潜在分布として機能することを明らかにした既存研究が存在する(Wang et al., 2019)。一方、本研究のように多様な詳細度合いを持つトピック文を生成するためには、文の詳細度合いを潜在空間上でモデル化する必要があり、その方法は明らかでない。

そこで本研究では、トピック文の潜在表現をガウス分布で表現する際に、子の分布の分散が親よりも小さくなるようにモデルを構築することで、根からは抽象的な文を、葉に近づくにつれより詳細な文を生成することを試みた。単語の潜在分布としてガウス分布を用いるGaussian word embeddingでは、「犬」のような具体的な単語は、「動物」といった一般的な単語よりも小さい分散を持つことが示されている(Vilnis et al., 2015)。文においても同様に、具体的な文は意味の分散が小さいため、その潜在分布は分散が小さいと考えられる。

子の分布の分散を親よりも小さくするために、本研究では再帰的混合ガウス分布(再帰的GMM)を文書中の文の潜在表現の事前分布として導入した。再帰的 GMM は、木構造上の各トピックに対応するガウス分布で構成され、子の事前分布に親の事後分布を設定して構築される。これにより、根の潜在分布の分散は大きく、葉に近づくにつれ分散は小さくなる。

評価実験では、提案法の要約性能は最新の教師なし生成型要約手法 (Bražinskas et al.,

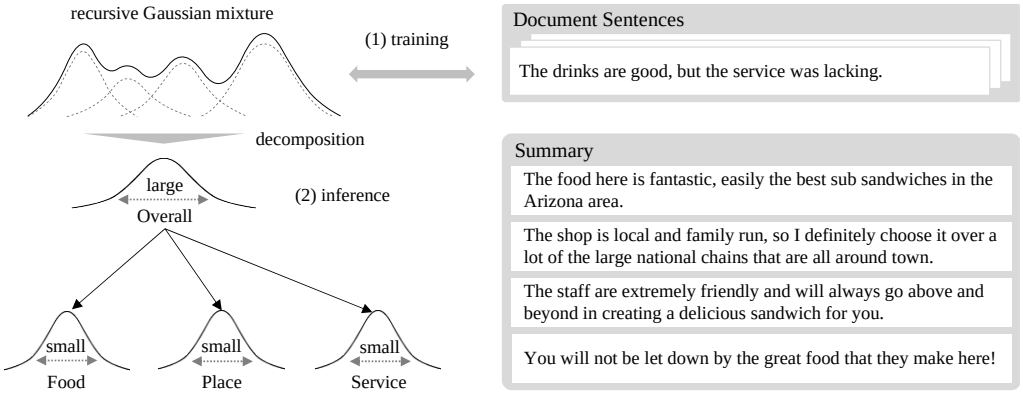


図 4 提案法の概要図。(1) 学習フェーズでは、文書の文の潜在変数が混合ガウス分布に従うと仮定し、そのパラメータを学習する。(2) 要約生成フェーズでは、混合ガウス分布を構成する各単峰ガウス分布からトピック文を生成し、その集合を要約として出力する。潜在分布の分散は葉に近づくほど小さくなるようモデル化されており、葉に近づくにつれより具体的なトピックに関する要約文が出力される。

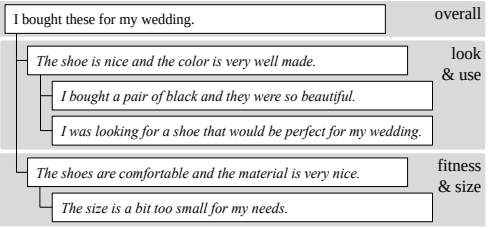
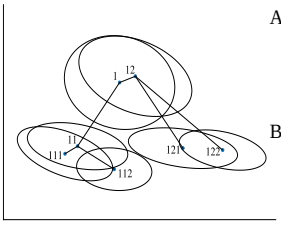


図 6 冠婚葬祭用シューズのレビューから得られた要約（トピック文の集合）の例



A)

LogDetCov	Yelp	Amazon
First level	28.83	27.21
Second level	26.22	26.54
Third level	23.28	24.55

B)

Topic Specialization	Yelp	Amazon
First level	1.68	1.60
Second level	1.99	1.63
Third level	2.16	1.84

図 5 トピック文の潜在分布の可視化（PCA）。葉に近づくにつれ分布の分散が小さくなる一方（表 A）、トピック文の内容は詳細になる（表 B）

2020）と競合することを確認した。図 5 に示すように、提案法は“look & use”や“fitness & size”といった多様なトピックを捉えており、人手評価では要約の informativeness や元の文書の被覆率といった観点で既存手法を上回ることを確認した。また、トピック文の詳細度合いはその潜在分布の分散の大きさに依存し、根の文の潜在分布は分散が大きく一般的な文が生成される一方、葉に近づくにつれ分散が小さくなり具体的な文が生成されるといった特性を確認した（図 6）。これは単語の潜在表現にガウス分布を用いた Gaussian word embedding にて報告された特性と類似しており、要約のみならず、質問応答や対話生成などの文の詳細度合いを考慮する他タスクにも有用な知見である。

以上の成果を取り纏めた論文は、計算言語学のトップジャーナル TACL に採択された他**エラー! 参照元が見つかりません。**、言語処理学会第 27 回年次大会で若手奨励賞を**エラー! 参照元が見つかりません。エラー! 参照元が見つかりません。**、情報処理学会第 246 回自然言語処理研究会で優秀研究賞および山下記念研究賞を受賞した**エラー! 参照元が見つかりません。エラー! 参照元が見つかりません。エラー! 参照元が見つかりません。**。

・ **加速フェーズ（1）トピックの分岐・統合を捉える時系列構造化トピックモデルの開発**

本研究では、文書のトピック分類を行うとともに、トピックの分岐・統合を捉えることのできる時系列構造化トピックモデルを提案した。

トピックの時系列変化を捉えることのできる著名なトピックモデルとして時系列トピックモデルが存在するが、時系列トピックモデルはトピック間の依存関係を捉えることができず、トピックの分岐・統合をモデル化することができない。そこで本研究では、新しいトピックが過去のどのトピックに依存して生成されたかを self-attention 機構を用いて表現することで、トピック間の依存関係を捉えられる時系列構造化トピックモデルを提案した（**エラー! 参照元が見つかりません。**）。

評価実験では、パープレキシティやトー貫性の観点で既存の時系列トピックモデルを上回る性能を確認した他、学術トピックの分岐・統合が捉えられていることを確認した。本研究はトピック推定・構造化に限らず、将来出現するトピック予測などの応用可能性が期待される。

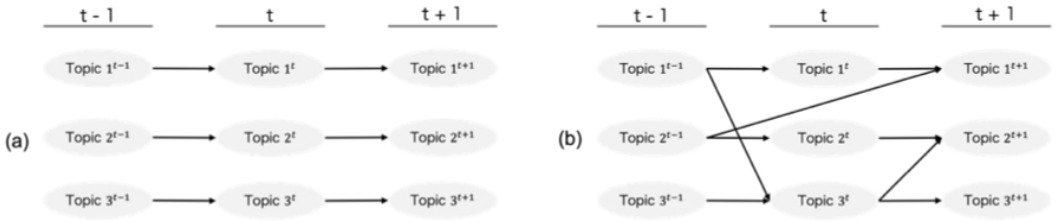


図 7 （a）既存の時系列トピックモデルと（b）提案する時系列構造化トピックモデル

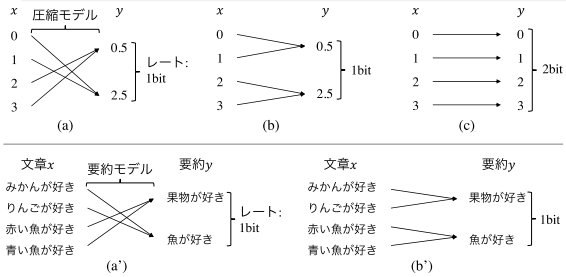


図 8 レート歪み理論から見た要約モデル

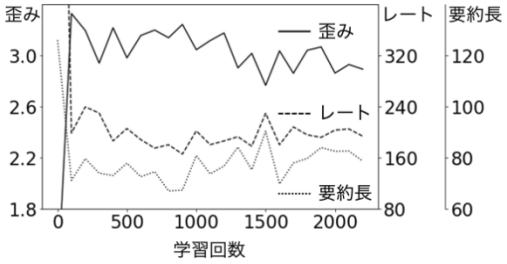


図 9 歪み・レート・要約長の推移

- **加速フェーズ（2）サーベイ論文生成用の教師データ作成及びサーベイ論文自動生成モデルの評価**

本研究では、サーベイ論文生成用の教師データを作成するとともに、サーベイ論文自動生成モデルと人が作成したサーベイ論文の比較評価に取り組んでいる。

これまで自動文章要約研究においてサーベイ論文生成は取り組まれていない。その大きな理由の一つは、サーベイ論文生成モデルの学習に必要な学習データが存在しないためである。そこで本研究では学術論文 1,200 万本以上の本文データが格納された S2ORC データセットを用いて、サーベイ論文の本文を取得し、サーベイ論文生成モデル学習用データセットを作成した。作成したデータセットは計算機科学分野のみで 1 万本以上のサーベイ論文から構成され、既存の要約モデルの入出力長でも扱えるように各論文を章ごとに分割することで、要約モデルを学習した。

サーベイ論文生成モデルと人が作成したサーベイ論文の比較評価を人手により行った結果、初期的には流暢性や網羅性の観点から高い評価が得られた。現在、専門家による人手評価を追加実施中である。

- **加速フェーズ（3）レート歪み理論に基づく文章要約の定式化**

本研究では、文章要約を非可逆データ圧縮として見做すことで、人手で作成された見本の要約に頼らない、文章要約の新たな定式化に取り組んでいる。本研究で提案する定式化に基づき学習されたモデルが文章要約を解くことができるか明らかにするとともに、一連の教師なし文章要約技術が、レート歪み理論においてどのように解釈できるか議論した。

レート歪み理論とは、非可逆データ圧縮において、圧縮後の情報量（レート）を一定にしたもとで、どの程度まで圧縮前後の誤差（歪み）を小さくできるかについて論じた、画像や音声圧縮の基礎

理論である。データ圧縮と同様に、文章要約においても、要約らしさを捉えられる適切な歪み関数を設計し、一定のレート（要約長）のもとで歪みを最小化するように要約モデルを学習することで、文章要約を解くことができると考えた（図 8）。

図 9 に示すように、文章要約モデルの教師あり学習を進めると、レートが一定のまま歪みが小さくなることが確認され、レートと要約長は強く相関することが確認できた。しかし、逆（レートと歪みの最小化により文章要約モデルが学習できること）を示すには至らず、より適切な歪み関数の設計、及びレート・歪み以外の要素が文章要約の定式化において必要であることが明らかになった。一方、既存の教師なし要約モデルの多くはレート歪み理論により一般化することができ、一般化に際し要約モデルとして重要な要素が欠落してしまっているのか今後議論が望まれる。

3. 今後の展開

これまで 2 年間の研究計画では、トピック構造に基づく教師なし文書要約技術の確立を目指しており、その目標は達成されたものと認識している。

しかし、教師なし文書要約の最終的な目標は汎用的な自動文書要約の実現であり、これまで評価実験で用いてきた商品レビュー以外にも適用できるのかという工学的な関心が残る。特にこれまで自動文書要約技術が適用されてこなかった領域を、トピック構造を捉える本アプローチによって開拓することを目指したい。

一方、計算言語学的な観点からみると、トピック構造を捉えることが要約の質にどのように寄与するのか、実験的には明らかにしたものの理論的には明らかにできていない。要約に求められる性質（非冗長性、原文書との関連性、情報量の多さ）を定式化した上で、それらとトピック構造を捉えることの数理的なつながりを議論したい。

加速フェーズでは、自動文書要約技術が適用されてこなかった領域のひとつとして、サーベイ論文の自動生成に挑戦した。サーベイ論文生成タスクを (1) サーベイ論文の対象分野における学術文献のトピック分類・構造化 と (2) 各トピックにおける学術文献の要約 に分割してそれぞれ検討したが、今後はこれらを一体化した手法の開発に取り組みたい。一方、(3) レート歪み理論に基づく文章要約の定式化 については、文章要約の定式化においてレート歪み理論が重要であることは示唆されつつも、更なる検討が望まれる。本課題については学術振興会海外特別研究員の研究課題として引き続き取り組む予定である。

4. 自己評価

- 研究目的の達成状況

  2 年間の研究計画では、トピック構造に基づく教師なし文書要約技術の確立を目指しており、その目標は達成されたと認識している。

- 研究の進め方（研究実施体制及び研究費執行状況）

  研究費は研究補助を担う学生の謝金や雇用経費に多く充てられており、リヴァプール大学ダヌシカ・ボレガラ教授や当該学生など、研究室内外の研究者との共同研究によって、研究が大きく加速した。

- 研究成果の科学技術及び社会・経済への波及効果

  本研究で開発した教師なし要約生成は、自動要約技術の実用化に重要な技術であり、社会・経済的にも重要な意義を持つと認識している。これまで自動文書要約技術が適用されてこなかった領域を今後開拓することによって、研究成果の社会的意義をより直接的に発信していきたい。

- その他

  ACT-X 開始以前は研究者としての個を確立していくという意識が欠けていたが、領域内の研究者やアドバイザーからの触発を通じて、いかに個を確立していくか常に自問するようになったことは、ACT-X で通じて得られた最も大きな価値の一つであった。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：2件

[1] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Tree-Structured Neural Topic Model. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp.800-806, 2020.
概要 本研究では複数の文書を並列に学習できる木構造ニューラルトピックモデルを開発することで学習時間を短縮し、大量の文書に対するトピック構造の推定を可能にした。提案法は従来法と同等の解釈性を持つトピックとその木構造を得ながら、学習時間を約 15 倍短縮し、ニューラルモデルを用いた要約生成など下流タスクとの一体的な学習を可能にするなど、木構造トピックモデルの応用可能性を広げた。
[2] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance. Transactions of the Association for Computational Linguistics (TACL), vol.9, pp.945-961, 2021.
概要 本研究では開発した木構造トピックモデルにより文書のトピック木構造を推定し、各トピックの要約文を生成する手法を開発した。トピック木構造では、文書から推定したトピックが木構造に配置され、子が親のサブトピックとなる構造を持つ。それらのトピックから要約として相応しい粒度のトピックを選択し、各トピックに関する要約文を生成することで、意見文書の要約が教師なしに得られることを示した。

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

国際会議発表

[3] Naomi Sato, Masaru Isonuma, Kimitaka Asatani, Shoya Ishizuka, Aori Shimizu and Ichiro Sakata. Lexical Entailment with Hierarchy Representations by Deep Metric Learning.

Findings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022. (※加速フェーズ 実施の成果)

国内会議発表

[4] 磯沼 大, 森 純一郎, ボレガラ ダヌシカ, 坂田 一郎. 木構造ニューラルトピックモデル. 言語処理学会第 26 回年次大会, 2020.

[5] 磯沼 大, 森 純一郎, ボレガラ ダヌシカ, 坂田 一郎. 潜在的なトピック構造を捉えた生成型教師なし意見要約. 情報処理学会第 246 回自然言語処理研究会, 2020.

[6] 磯沼 大, 森 純一郎, ボレガラ ダヌシカ, 坂田 一郎. トピック文生成による教師なし意見要約. 言語処理学会第 27 回年次大会, 2021.

[7] 向井 穂乃花, 磯沼 大, 森 純一郎, 坂田 一郎. Homophily に基づくサイレントマジョリティの意見推定. 言語処理学会第 28 回年次大会, 2022.

[8] 宮本 望, 磯沼 大, 森 純一郎, 坂田 一郎. Self-attention 機構に基づく Dynamic Structured Neural Topic Model. 人工知能学会第 36 回全国大会, 2022. (※加速フェーズ 実施の成果)

[9] 笠西 哲, 磯沼 大, 森 純一郎, 坂田 一郎. Transformer Encoder-Decoder モデルによるサーベイ論文の自動生成. 人工知能学会第 36 回全国大会, 2022. (※加速フェーズ 実施の成果)

受賞

[10] 優秀研究賞, 情報処理学会第 246 回自然言語処理研究会, 2020.

[11] 若手奨励賞, 言語処理学会第 27 回年次大会, 2021.

[12] 山下記念研究賞, 情報処理学会, 2021.

[13] 工学系研究科長賞, 東京大学, 2021.