様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月　 ６ 日現在

機関番号：１２６０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１２０７６

研究課題名（和文）大規模学術文献データのネットワーク構造を考慮した事前学習言語モデルに関する研究

研究課題名（英文）Pre-trained language models using the network structure of large-scale scholarly data

研究代表者
森　純一郎（Mori, Junichiro）

東京大学・大学院情報理工学系研究科・准教授

研究者番号：３０５０８９２４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：膨大な学術文献データから新発見や課題解決に繋がる多様な学術知を抽出することの重要性が認識されてきている。本研究では、大規模な学術文献データから有用な知識の抽出と発見を支援することを目的に、学術文献データのネットワーク構造を考慮した大規模ハイパーテキストデータからの事前学習言語モデルの構築に関する基本的な方法論の研究を行った。研究成果として、大規模学術文献データの文献間の引用関係に基づくハイパーテキストデータから事前学習言語モデルを構築する技術ならびに事前学習言語モデルを用いて大規模な学術文献データから有用な知識の抽出と発見を支援するための技術の開発を行なった。

研究成果の学術的意義や社会的意義
まず、COVID-19に関する科学的エビデンスや重要な技術などの情報を抽出しその解析結果を広く一般に公開した。次に、 引用ネットワーク構造を考慮した文献コーパスからの事前学習言語モデル構築のための予測問題の設計と実装に取り組んだ。また、事前学習言語モデルにより獲得された分散表現を用いた引用ネットワークのリンク予測および ノード分類タスクによる評価に取り組んだ。 最後に、期間中に研究開発を行った手法を応用し、萌芽的な学術論文の発見、サーベイ論文の自動生成、研究トピックの抽出と時系列変化の可視化など、複数の新たなタスクに取り組んだ。これらの研究成果を複数の学会で発表した。

研究成果の概要（英文）：The importance of extracting diverse academic knowledge from vast amounts of academic literature data that leads to new discoveries and problem-solving has been recognized. In this study, with the aim of supporting the extraction and discovery of useful knowledge from large-scale academic literature data, we conducted research on the fundamental methodology for constructing pre-trained language models from large-scale hypertext data that considers the network structure of academic literature data. As research results, we developed a technology for constructing pre-trained language models from large-scale academic literature data based on the citation relationships between documents in the form of hypertext data, as well as a technology for supporting the extraction and discovery of useful knowledge from large-scale academic literature data using pre-trained language models.

研究分野：知能情報学

キーワード：学術文献データ　事前学習言語モデル　引用ネットワーク　表現学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

学問領域の細分化と科学技術知識の深化に伴い、学術文献情報が大規模に蓄積されて来ており、発表される学術文献の数は、年々爆発的に増加している。この背景の元、大規模な学術文献データのオープン化が現在進められている。学術論文誌の主要な出版社は API による学術文献データへのアクセスを可能にし、また、ウェブ上には arXiv などの膨大な学術文献レポジトリが公開されており、これらを自動的に収集し整理した大規模な学術文献データを研究開発用途に公開、共有する試みが複数の研究グループで進められている。

特に近年、学術文献データのさまざまなエンティティ（論文、著者、機関など）とそれらの関係で構成されるアカデミックグラフと呼ばれる大規模な学術文献データを元にした知識グラフがオープンデータとして公開されてきている。Microsoft、清華大学、Allen Institute for AI などが現在公開しているそれらのオープンデータは 2 億におよぶ学術文献データを元に構築された大規模なアカデミックグラフとなっている。

アカデミックグラフの核をなすものは、論文とその引用関係から成る膨大なテキストとネットワークのデータであり、科学技術知識の集積である学術文献データから情報技術を用いて高次の知識を横断的に抽出・発見しようという研究の取り組みが、国内外で盛んになされている。

大規模な学術文献テキストデータに対する言語処理技術の応用の背景には近年の目覚ましい言語処理技術の革新がある。中でも、大規模なテキストデータの解析においては BERT や XLNet のような事前学習言語モデルが様々なタスクにおいて有効であることが近年示されてきている。学術文献データについても大規模な学術文献テキストデータから構築した事前学習言語モデルである SCIBERT や BIOBERT などが提案され、学術文献テキストからの固有表現抽出や関係抽出、テキストの分類などのタスクを高精度で解くことが示されている。

一般に事前学習言語モデルの構築においては、文中のマスクされた単語の予測や文の並びの予測などの予測問題を大規模なテキストコーパスを元に解くことで自己注意機構により文脈の汎用的な表現を学習している。大規模な学術文献テキストデータからの事前学習言語モデルの構築においても学術文献テキストコーパスを対象にこれらの予測問題を通して学術ドメインに特有の言語表現を学習している。

２．研究の目的

本研究では、大規模学術文献データの持つテキストとそのネットワークに着目し、文献同士が引用関係で結ばれた大規模ハイパーテキストデータからの事前学習言語モデルの構築に関する基本的な方法論の研究を行う。画像処理分野において事前学習モデルとその転移学習がさまざまなブレークスルーをもたらしたように、言語処理分野における事前学習モデルも今後基盤的な技術となることが予見され、現在盛んに研究が行われている最中である。一方、直近の国内外の研究動向を見ても、ハイパーテキストデータのようなつながり、ネットワーク構造、を持つテキストを対象とした事前学習モデルの構築に関する方法論の研究開発は途上にある段階であり、未だ十分な知見が得られていない。特に、互いにつながりのある膨大なテキストデータをどのように分析し統合するかを明らかにすることは、アカデミックグラフのような大規模な学術文献データの解析において重要な課題である。

テキストコーパスからの従来の事前学習言語モデル構築では単語出現予測や文の並び予測などの予測問題の設計がなされている。本研究の学術的独自性・創造性は、大規模な学術文献テキストコーパスからの事前学習言語モデルの構築において、テキストに関する予測問題に加え

てテキスト間の関係性、ネットワーク構造を考慮した予測問題の設計を行った上で事前学習言語モデルを構築することことである。これにより、学術領域全体の知識構造を考慮した言語表現を学習することで、その事前学習言語モデルを学術文献データからの知識抽出や発見の支援に活用することが期待できる。

３．研究の方法

本研究では、次の主たる技術(1)および(2)を大目標として研究開発を行う。各技術は具体的には以下に示す 6 つの研究項目を行うことで実現を目指す。

(1)大規模学術文献データの文献間の引用関係に基づくハイパーテキストデータから事前学習言語モデルを構築する技術

① 大規模な学術文献データから引用ネットワークを構築する手法の設計と実装
② 引用ネットワーク構造を考慮した文献テキストコーパスからの事前学習言語モデル構築のための予測問題の設計と実装
③ 事前学習言語モデルの自己注意機構の設計と実装

(2)事前学習言語モデルを用いて大規模な学術文献データから有用な知識の抽出と発見を支援するための技術

① 事前学習言語モデルを用いた文献テキストからの固有表現抽出、関係抽出およびテキスト分類タスクによる評価
② 事前学習言語モデルにより獲得された分散表現を用いた引用ネットワークのリンク予測およびノード分類タスクによる評価
③ データセットおよび事前言語学習モデルの公開

(1)について、学術文献のようにテキスト同士が関係で結ばれ全体としてネットワーク構造を持ったテキストコーパスからの事前学習言語モデルを構築するための知見を明らかにする。
本研究に用いるデータとしては、オープンデータとして現在利用可能なアカデミックグラフを用いる。また、すでに連携体制を構築している文献データベースプロバイダと協力し評価用のデータセットを独自に構築する。これらの学術文献データからの引用ネットワークの構築については、提案者のこれまでの研究の知見と成果を活用する。

事前学習言語モデルの構築には従来文献のアブストラクトのような単一テキスト内に閉じた予測問題が用いられているが、本研究ではネットワーク構造を持ったテキスト同士の横断的な予測問題の設計を行う。具体的には引用関係を考慮することで左図のように複数文献のアブストラクトを横断した単語の出現予測や文のエンテールメント予測などが考えられる。また、これらの予測問題に適した自己注意機構の設計を行う。(1)でえられた知見は、学術文献データのほかウェブやソーシャルメディアなどハイパーテキスト構造を持つデータにも活用できる。

次に、(2)について、事前学習言語モデルを用いた学術文献データから知識抽出・発見の応用に関する知見を明らかにする。具体的には、文献テキストからの固有表現や関係抽出などのタスクで事前学習言語モデルの評価を行う。これらは学術文献を対象とした既存の事前言語学習モデル、SCIBERT や BIOBERT、の適用タスクでもありこれらをベースラインとした評価を行う。さらに、事前学習言語モデルにより獲得された分散表現を用いて引用ネットワークのリンク予測やノード分類などのタスクでの評価を行う。本研究の成果は学術論文として投稿するとともに、評価用データセット、事前学習言語モデル、コードなどについては広くその研究成果を公開する。

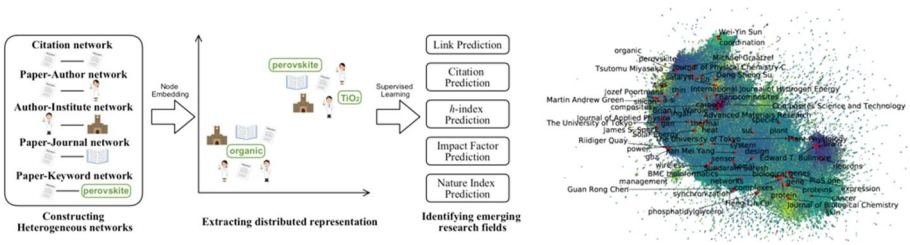
そのために、提案者らが研究開発を進めている大規模な学術文献データを分析するシステムである「学術産業技術俯瞰システム」を活用する。

４．研究成果

まず、COVID-19 のパンデミックの発生以降において、その学術研究は急速に増加していた。そこで、関連する学術文献データセットに対 して引用ネットワーク解析を行い、COVID-19 に関する科学的エビデンスや重要な技術などの情報を抽出した。その解析結果を広く一般に公開した。

学術産業技術俯瞰システムによる COVID-19 関連文献引用ネットワークの解析

次に、大規模学術文献データの文献間の引用関係に基づくハイパーテキストデータから事前学習言語モデルを構築する技術として、引用ネットワーク構造を考慮した文献テキストコーパスからの事前学習言語モデル構築のための予測問題の設計と実装に取り組んだ。また、事前学習言語モデルを用いて大規模な学術文献データから有用な知識の抽出と発見を支援するための技術として、事前学習言語モデルにより獲得された分散表現を用いた引用ネットワークのリンク予測およびノード分類タスクによる評価に取り組んだ。



多層ネットワークの表現学習に基づく研究指標の予測

最後に、期間中に研究開発を行った手法を応用し、萌芽的な学術論文の発見、サーベイ論文の自動生成、研究トピックの抽出と時系列変化の可視化など、複数の新たなタスクに取り組んだ。

サーベイ論文の自動生成について、自動文章要約の応用として、学術文献データからのサーベイ論文自動生成のためのベンチマークデータセット構築と評価を行った。近年、大規模言語モデルによりサーベイ論文生成の試みがなされているが、大規模データセットの欠如がその進歩の足枷となっている。本研究では、1 万本超のサーベイ 論文と 69 万本超の被引用論文で構成されたサーベ イ論文生成データセットを構築した。本データセッ トをもとに、近年の Transformer 要

約モデルをサー ベイ論文生成用に改良し、サーベイ論文生成の評価実験を行なった。その結果、人手評価により、モデルにより生成された要約の一部は人が作成したサーベイ論文と遜色ないことが示された一方で、自動サーベイ論文生成の課題が明らかになった。

時系列構造化ニューラルトピックモデルに関する研究について、トピックモデルにおいてトピック間の依存関係を捉えながら、その時系列的発展を扱うことができる新たな時系列構造化ニューラルトピックモデルを新たに開発した。提案モデルにより、トピックの依存関係を self-attention 機構に基づいてモデル化することで、トピックの分化・統合過程を捉えることが可能になった。さらに、アテンションの重みが文書間の引用関係を反映するように、引用正則化項を新たに導入した。これにより、提案モデルは Perplexity や Coherence において、 既存の時系列トピックモデルを上回る性能を達成した。また、実際のデータセットとして学術文献データにモデルを適用し、トピックの遷移過程を捉えられることを検証した。

これらの研究成果は言語処理に関するトップジャーナルや複数のトップ学会に採択されるに至り、また国内学会での優秀論文受賞に至った。

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　2件／うちオープンアクセス　3件）

１．著者名 Masanao Ochi, Masanori Shiro, Junichiro Mori, Ichiro Sakata	４．巻 -
２．論文標題 Classification of the Top-cited Literature by Fusing Linguistic and Citation Information with the Transformer Model	５．発行年 2022年
３．雑誌名 Proceedings of the 18th International Conference on Web Information Systems and Technologies	６．最初と最後の頁 286-293
掲載論文のDOI（デジタルオブジェクト識別子） 10.5220/0011542200003318	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ochi Masanao、Shiro Masanori、Mori Jun’ichiro、Sakata Ichiro	４．巻 17
２．論文標題 Predictive analysis of multiple future scientific impacts by embedding a heterogeneous network	５．発行年 2022年
３．雑誌名 PLOS ONE	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0274253	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Masaru Isonuma, Junichiro Mori, Danushka Bollegala, Ichiro Sakata	４．巻 9
２．論文標題 Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance	５．発行年 2021年
３．雑誌名 Transactions of the Association for Computational Linguistics	６．最初と最後の頁 945-961
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/tacl_a_00406	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Isonuma Masaru、Mori Junichiro、Bollegala Danushka、Sakata Ichiro	４．巻 1
２．論文標題 Tree-Structured Neural Topic Model	５．発行年 2020年
３．雑誌名 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)	６．最初と最後の頁 995-1005
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/2020.acl-main.73	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Chou Jen Shiau、Masanao Ochi、Takeshi Sakaki、Ken Nagahama、Kanji Sakai、Junichiro Mori、Ichiro Sakata	４．巻 1
２．論文標題 Constructive Approach for Early Extraction of Viral Spreading Social Issues from Twitter	５．発行年 2020年
３．雑誌名 Proceedings of ACM Web Science 2020 (WebSci2020)	６．最初と最後の頁 96-105
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3394231.3397899	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ochi Masanao、Shiro Masanori、Mori Jun’ichiro、Sakata Ichiro	４．巻 -
２．論文標題 Which Is More Helpful in Finding Scientific Papers to Be Top-cited in the Future: Content or Citations? Case Analysis in the Field of Solar Cells 2009	５．発行年 2021年
３．雑誌名 Proceedings of the 17th International Conference on Web Information Systems and Technologies	６．最初と最後の頁 360-364
掲載論文のDOI（デジタルオブジェクト識別子） 10.5220/0010689100003058	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 宮本望，磯沼大，高瀬翔，森純一郎，坂田一郎
２．発表標題 時系列構造化ニューラルトピックモデル
３．学会等名 言語処理学会第29回年次大会
４．発表年 2023年

１．発表者名 笠西哲，磯沼大，森純一郎，坂田一郎
２．発表標題 サーベイ論文自動生成に向けた大規模ベンチマークデータセットの構築
３．学会等名 言語処理学会第29回年次大会
４．発表年 2023年

１．発表者名
大知正直、城真範、森純一郎、坂田一郎

２．発表標題
Transformerモデルを用いた学術文献の言語情報と引用情報の融合

３．学会等名
2022年度人工知能学会全国大会

４．発表年
2022年

１．発表者名
宮本望、磯沼大、森純一郎、坂田一郎

２．発表標題
Self-attention機構に基づくDynamic Structured Neural Topic Model

３．学会等名
2022年度人工知能学会全国大会

４．発表年
2022年

１．発表者名
笠西哲、磯沼大、森純一郎、坂田一郎

２．発表標題
Transformer Encoder-Decoderモデルによるサーベイ論文の自動生成

３．学会等名
2022年度人工知能学会全国大会

４．発表年
2022年

１．発表者名
大知 正直，城 真範，森 純一郎，坂田 一郎

２．発表標題
科学研究のインパクト予測に向けた学術文献情報から抽出した分散表現による特定可能性分析

３．学会等名
2021年度 人工知能学会全国大会（第35回）

４．発表年
2021年

１．発表者名
Masaru Isonuma, Junichiro Mori, Danushka Bollegala, Ichiro Sakata

２．発表標題
Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree Structured Topic Guidance

３．学会等名
The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021)（国際学会）

４．発表年
2021年

１．発表者名
向井穂乃花、磯沼大、森純一郎、坂田一郎

２．発表標題
Homophilyに基づくサイレントマジョリティの意見推定

３．学会等名
言語処理学会第28回年次大会

４．発表年
2022年

１．発表者名
Junichiro Mori

２．発表標題
Citation Network Analysis of the COVID-19 Open Research Dataset

３．学会等名
Second International Workshop on SCIentific DOCument Analysis (SCIDOCA 2020)（国際学会）

４．発表年
2020年

１．発表者名
磯沼大、森純一郎、坂田一郎

２．発表標題
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

３．学会等名
情報処理学会 第246回自然言語処理研究会

４．発表年
2020年

１．発表者名
磯沼大、森純一郎、坂田一郎

２．発表標題
トピック文生成による教師なし意見要約

３．学会等名
言語処理学会第27回年次大会

４．発表年
2020年

１．発表者名
蕭喬仁、大知正直、長濱憲、榊剛史、森純一郎、阪井完二、坂田一郎

２．発表標題
構築主義的アプローチに基づく情報拡散型社会問題の早期抽出

３．学会等名
2020年度人工知能学会全国大会

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

COVID-19関連論文の引用解析
https://academic-landscape.com/analysis/36093

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関