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　６ 年　 ６ 月　 ７ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2021～2023
課題番号：２１Ｈ０３４９５

研究課題名（和文）テキスト間の関係を考慮した複数テキスト生成型ニューラル要約に関する研究

研究課題名（英文）Abstractive Neural Multi-document Summarization Considering Cross Document Structure

研究代表者
奥村　学（Okumura, Manabu）

東京工業大学・科学技術創成研究院・教授

研究者番号：６０２１４０７９

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：文間の関係を解析する文書構造解析器では，大規模言語モデル（LLM）を活用して，プロンプトを介してシフト還元動作を LLM で模倣する手法を提案し，評価実験の結果，提案法は世界最高の解析性能を達成した．テキスト要約の方では，この文書構造解析結果を活用したニューラル文書要約モデルを提案し，要約の性能向上に寄与することを確認した．また，事前学習済み言語モデル(PLM) を追学習する時に，エンコーダに要約長を予測させることで要約タスクに固有の情報を理解させた上で，デコーダには予測した要約長の要約を生成させるモデルを提案し，要約の性能向上に寄与することを確認した．

研究成果の学術的意義や社会的意義
文間の関係を解析する文書構造解析器は，我々のグループが世界最高性能を達成していたが，引き続き研究開発を継続し，新しい手法を提案することで，現在も世界最高性能を維持している．テキスト要約において要約長を予測するというアイデアはこれまでに提唱されておらず，そういう意味で斬新なアイデアに基づいており，しかも，要約長を予測するよう要約モデルを学習することで性能向上に寄与することを示しており，学術的な意義は大きい．

研究成果の概要（英文）：In document structure analysis, that analyzes the relationships between sentences, by utilizing large language models (LLMs), we proposed a method to imitate shift-reduce operations through prompts. As a result of evaluation experiments, the proposed method achieved the state-of-the art performance. In text summarization, we proposed a neural model that utilizes the results of this document structure analysis. We confirmed that this contributes to improving the performance of the summarization. We further proposed a method for enabling the model to understand the summarization-specific information by predicting the summary length in the encoder and generating a summary of the predicted length in the decoder in fine-tuning. We confirmed that this also contributes to improving the performance.

研究分野：知能情報学

キーワード：自然言語処理　複数テキスト要約　ニューラルモデル　生成型要約　文書横断文間関係

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

世の中のテキストの量の増大は留まることを知らず，もはや要約とはいえ，1 つ 1 つのテキストを読み手が読んではいられない現在の状況では，複数テキスト要約技術は不可欠と言わざるを得ない．1 テキストを対象とする単一テキスト要約技術と比べて，同一トピックに関するテキスト集合を対象に，それらを 1 つのテキストにまとめる複数テキスト要約技術は，利用可能なデータセットの少なさ，小ささから研究が近年においてさえ発展途上の段階である．テキスト要約技術は，元テキストの単語，句のみを用いて作成する抽出型と，元テキストにない単語，句も用いて作成する生成型の 2 つの手法に大別される．複数テキスト要約の場合，単一テキスト要約と比べて，はるかに低い要約率（元テキストの長さと比べた要約の長さ）で要約を作成する必要があり，抽出型ではなく，より短い要約を作成できる可能性がある生成型手法を採ることが望ましいと考えられる．

しかし，テキスト要約研究では長らく生成型手法が成功せず，その実現可能性の高さから抽出型手法が大多数である期間が続いた．近年になり，Rush ら[1]が文要約で初めてニューラルモデルを用いて生成型要約を実現して以後，文要約にとどまらず単一テキスト要約，複数テキスト要約においても，生成型のニューラルモデルの研究が盛んになりつつある．しかし，複数テキストを対象とした生成型ニューラル要約の研究は依然，文要約や単一テキスト要約と同様のモデルである encoder-decoder モデルを単に用いて，複数テキストを encode し，要約テキストを生成する手法がほとんどであり，必ずしも質の高い要約が作成できていない．

一方，複数テキスト要約研究では，1) 共通のトピックの複数テキスト中には類似する内容の文が多数存在する可能性があり，単一テキスト要約と異なり，生成する要約中の冗長性をいかに除去するかを十分に考慮する必要がある，2) 元テキストの順序で要約文を並べれば問題があまりない，単一テキスト要約の場合と異なり，複数テキスト要約では要約文をどのような順序で並べれば良いかは明らかでない，ことが知られている．また，後述するように，我々のこれまでの研究から，3) テキスト集合中に共通して出現する固有名などを検出（文書横断共参照解析）したり，それらのテキスト集合中の文間の関係を解析（文書横断構造解析）したりした結果を利用することにより，テキスト要約の性能を向上できる可能性がある．

[1] Alexander M. Rush, Sumit Chopra, Jason Weston, A Neural Attention Model for Abstractive Sentence Summarization, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 379–389, 2015.

２．研究の目的

そこで本研究課題では，複数テキストを対象とした生成型要約において，以下の特性を有するニューラルモデルを研究開発することを目的とする．

a) 文書横断共参照解析や文書横断構造解析の結果を encode 時に活用する，
b) 適切な要約文の出力順序を決定できる，
c) 要約テキスト中の冗長性を適切に除去する機構を保持する．

より具体的には，ニューラル要約モデルを 2 段階の連結モデルとして構成し，1) 文書横断共参照解析や文書横断構造解析の解析結果を考慮した上で，要約文集合をその順序とともに生成するニューラルモデル，2) 冗長性の度合いや文の順序の首尾一貫性の度合いを元に，順序付き要約文集合をリランキングし，最適な順序付き要約文集合を出力するニューラルモデルを研究開発し，より高品質の複数テキスト要約作成を実現する．

本研究課題では，複数テキストを対象とした生成型要約を実現するためニューラルモデルを採用するが，a), b)の特性を有するニューラルモデルはこれまでになく新規性は高い．c)に対してこれまでのニューラル要約モデルは，coverage 機構[2]など，一文ずつ出力する際にこれまでの出力文との重複を考慮する「greedy」な手法を採用していた．このため，要約テキスト全体として冗長性が十分に除去できていない可能性がある．そこで本研究課題では，ニューラル要約モデルを 2 段階の連結モデルとして構成し，1 段階目で生成した順序付き候補文集合を，2 段階目で冗長性の度合いを元にリランキングする手法を採用することで，冗長性をより十分に除去した最適な要約文集合を出力できる手法を研究開発する．

ニューラルモデルはこれまで，タスクの入出力のペアを訓練データとして与え，end-to-end で学習する手法がほとんどであったが，近年になって，タスクをいくつかに分割し，分割したサブタスクごとに学習を行い，それらを統合する形でタスクを実行するニューラルモデルが性能向上を実現している．単一テキスト要約では，テキスト全体のベクトルとの類似度が最大の要約文集合が最適であるという仮説を元に，抽出型ではあるが，2 段階で要約を行うニューラルモデルが近年世界最高性能を達成している[3]．しかし，冗長性を最小化するという観点で，2 段階で大域的最適化を行うニューラルモデルはこれまでにない．

[2] Abigail See, Peter J. Liu, and Christopher D. Manning, Get to the point: Summarization with pointer-generator networks, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1073–1083, 2017.
[3] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, Xuanjing Huang, Extractive Summarization as Text Matching, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6197–6208, 2020.

３．研究の方法

2 で述べたように, 本研究課題では, ニューラル要約モデルを 2 段階の連結モデルとして構成し, 1) 文書横断共参照解析や文書横断構造解析の解析結果を考慮した上で, 要約文集合をその順序とともに生成するニューラルモデル, 2) 冗長性の度合いや文の順序の首尾一貫性の度合いを元に, 順序付き要約文集合をリランキングし, 最適な順序付き要約文集合を出力するニューラルモデルを研究開発する. 1)の研究開発は, 文書横断共参照解析および文書横断構造解析技術の研究開発と, それらの解析結果を encode して要約文集合を生成するニューラルモデルの研究開発に細分化できるので, 結果的に本研究課題は以下で述べる 3 つのコア技術に分解し研究開発を行なうことになる.

・文書横断共参照解析および文書横断構造解析技術の研究開発
・文書横断共参照解析および文書横断構造解析の解析結果を encode して, 要約文集合をその順序も考慮して生成するニューラルモデルの研究開発
・冗長性の度合い, 文の順序の首尾一貫性の度合いの尺度の研究開発及び, その尺度を元に最適な要約文集合を選択可能な 2 段階の連結型ニューラル要約モデルの研究開発

４．研究成果

テキスト内の文間の関係を解析する文書構造解析器は, 我々のグループが世界最高性能を達成していたが, 引き続き研究開発を継続し, 新しい手法を提案することで, 現在も世界最高性能を維持している.

初年度には, ニューラル機械翻訳で提案された逆翻訳による疑似正解データの活用にヒントを得て, 既存の文書構造解析器を用いて自動的に作成された大規模な疑似正解データを用いて解析器を事前学習し, 本来の訓練データを用いて追学習することで性能を改善する枠組みを提案した. また, 疑似正解データを大量かつ高品質に獲得するために, 複数の解析器が出力する木の間で重複する部分木を疑似正解データとして効率よく抽出するアルゴリズムを提案した.

2 年目では, この分野の技術の進歩を明確にできるよう, 既存の上向き, 下向きの解析戦略と最新の事前学習済み言語モデルを組み合わせることで強いベースライン解析器を構築した. この解析器を評価した結果, 解析戦略には大きな差がなく, トークンではなくスパンのマスキングを採用した事前学習済み言語モデルが有効であることが分かった. 特に, DeBERTa を用いると世界最高性能を達成した. また, 十分な量の学習データを確保できないことにより, 文間の文書構造解析の性能は文内と比較して大幅に低く, 下流タスクにとって大きな問題となっている. これを解決するため, 学習データを逆翻訳することで得た擬似正解データを用いて解析器を事前学習し, 正解データを用いて追加学習する手法を提案した. シフト還元法による上向き解析器, スパン分割による下向き解析器に提案法を適用し, 標準的ベンチマークデータセットである RST-DT, Instr-DT を用いて評価した結果, 疑似正解データを用いることで Standard-ParsEval のスコアが約 1-2 ポイント向上することを確認した.

デコーダのみからなる大規模言語モデル (LLM) の発展は目覚ましく, 様々な自然言語処理タスクにおいて良好な結果を残している. 一方, 文書構造解析におけるそれらの有効性はこれまで議論されていない. そこで, 3 年目では, 今後の文書構造解析の研究において LLM を活用すべきかどうかを探ることを目的として, プロンプトを介してシフト還元動作を LLM で模倣する手法を提案し, その有効性を議論した. 評価実験の結果, 提案法は世界最高の解析性能を達成し, テキストドメインの汎化性においても優れていた. つまり, 文書構造解析においても LLM に注力すべきことが強調される結果を得た.

テキスト要約の方では, このテキスト内での文書構造解析結果を活用したニューラル文書要約モデルを提案し, 要約の性能向上に寄与することを確認した. また, 事前学習済み言語モデル (PLM) を追学習することで実用的な生成型要約モデルを獲得できることが明らかになっているが, 目標の要約長など要約タスクに固有の情報を PLM の事前学習時に十分考慮できているとは言えない. そこで追学習時に, エンコーダに要約長を予測させることで要約タスクに固有の情報を理解させた上で, デコーダには予測した要約長の要約を生成させるモデルを提案した. WikiHow, NYT, CNN/DM データセットを用いた実験により, BART よりも ROUGE スコアを向上させること, WikiHow データセットでは, GSum よりも ROUGE-1, -2, -L をそれぞれ約 3.0, 1,5, 3.1 ポイント向上させることを確認した. テキスト要約において要約長を予測するというアイデアはこれまでに提唱されておらず, そういう意味で斬新なアイデアに基づいており, しかも, 要約長を予測するよう要約モデルを学習することで性能向上に寄与することを示しており, 学術的な意義は大きい.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura and Masaaki Nagata, Improving Neural RST Parsing Model with Silver Agreement Subtrees, Proc. of NAACL-HLT 2021, 2021.
Ying Zhang, Hidetaka Kamigaito and Manabu Okumura, A Language Model-based Generative Classifier for Sentence-level Discourse Parsing, Proc. of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), 2021
Jingun Kwon, Naoki Kobayashi, Hidetaka Kamigaito and Manabu Okumura, Considering Nested Tree Structure in Sentence Extractive Summarization with Pre-trained Transformer, Proc. of The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), 2021.
Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura and Masaaki Nagata, A Simple and Strong Baseline for End-to-End Neural RST-style Discourse Parsing, Findings of The 2022 Conference on Empirical Methods in Natural Language Processing EMNLP 2022, 2022.
Jingun Kwon, Hidetaka Kamigaito and Manabu Okumura, Abstractive Document Summarization with Summary-length Prediction, Findings of EACL 2023, 2023.
Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, Can we obtain significant success in RST discourse parsing by using Large Language Models?, Proc. of The 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2024), 2024.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 小林尚輝，平尾努，上垣外英剛，奥村学，永田昌明	４．巻 29
２．論文標題 疑似正解データを 活用したニューラル修辞構造解析	５．発行年 2022年
３．雑誌名 自然言語処理	６．最初と最後の頁 10-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.5715/jnlp.29.875	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7件（うち招待講演　0件／うち国際学会　5件）

１．発表者名 Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura and Masaaki Nagata
２．発表標題 A Simple and Strong Baseline for End-to-End Neural RST-style Discourse Parsing
３．学会等名 The 2022 Conference on Empirical Methods in Natural Language Processing EMNLP 2022（国際学会）
４．発表年 2022年

１．発表者名 前川在，小林尚輝，平尾努，上垣外英剛，奥村学
２．発表標題 逆翻訳を利用したデータ拡 張による文間の修辞構造解析の改善
３．学会等名 言語処理学会第29回年次大会(NLP2023)
４．発表年 2022年

１．発表者名 Jingun Kwon, Naoki Kobayashi, Hidetaka Kamigaito and Manabu Okumura
２．発表標題 Considering Nested Tree Structure in Sentence Extractive Summarization with Pre-trained Transformer
３．学会等名 The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)（国際学会）
４．発表年 2021年

１．発表者名
Ying Zhang, Hidetaka Kamigaito and Manabu Okumura

２．発表標題
A Language Model-based Generative Classifier for Sentence-level Discourse Parsing

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

４．発表年
2021年

１．発表者名
Jingyi You, Chenlong Hu, Hidetaka Kamigaito, Hiroya Takamura and Manabu Okumura

２．発表標題
Abstractive Document Summarization with Word Embedding Reconstruction

３．学会等名
RANLP 2021（国際学会）

４．発表年
2021年

１．発表者名
Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura and Masaaki Nagata

２．発表標題
Improving Neural RST Parsing Model with Silver Agreement Subtrees

３．学会等名
NAACL-HLT 2021（国際学会）

４．発表年
2021年

１．発表者名
小林尚輝，平尾努，上垣外英剛，奥村学，永田昌明

２．発表標題
言語モデルと解析戦略の 観点からの修辞構造解析器の比較

３．学会等名
言語処理学会第28回年次大会(NLP2022)

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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
研究分担者	上垣外　英剛 (Kamigaito Hidetaka) (40817649)	奈良先端科学技術大学院大学・先端科学技術研究科・准教授 (14603)	

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