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　３ 年　 ６ 月 １７ 日現在

機関番号：１２６０８
研究種目：若手研究
研究期間：2019～2020
課題番号：１９Ｋ２０３３９

研究課題名（和文）段階的な抽出と書き換えに基づく生成型要約手法の研究

研究課題名（英文）Generative Summarization Based on Stepwise Extraction and Rewriting

研究代表者
上垣外　英剛（Kamigaito, Hidetaka）

東京工業大学・科学技術創成研究院・助教

研究者番号：４０８１７６４９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：ニューラルネットワークに基づく既存の文書要約手法において、人間の様に文の抽出、圧縮、書き換えを伴う段階的な要約を実現するために、既存の要約手法を援用可能な様々なドメインで動作可能な頑健な文圧縮器を作成した。この文圧縮器の作成過程で、事前学習された単語ベクトルの利用が性能向上に寄与することが判明したため、単語ベクトルを外部知識を用いて補強する際に必要となる知識グラフの埋め込みについても調査し、学習時に適した損失関数を選択するための理論的な背景を示した。最終的に作成した文圧縮器を既存の文書要約手法に組み込んだ結果、文抽出要約の設定において、自動評価の観点から性能の向上が確認された。

研究成果の学術的意義や社会的意義
文書の自動要約はデジタル文書が増加するインターネット社会において、読者が情報の取捨選択を行う際に重要な技術であると考えられる。本研究では要約生成時の動作が隠蔽されている既存のニューラルネットワークに基づく文書要約手法とは異なり、実際に要約が生成される過程が明確であるため、獲得したい要約結果の調整が容易であるという点で有用である。また文圧縮過程において使用される単語情報に外部知識を反映可能であるため、既存の文書圧縮手法に比べより多くのドメインでの動作が期待できる。これはニュース記事のみならずブログ記事やレビュー投稿等も対象とすることが可能である点で適用範囲が広く有用である。

研究成果の概要（英文）：In order to achieve human-like stepwise summarization with sentence extraction, compression, and rewriting in existing document summarization methods based on neural networks, we have developed a robust sentence compressor that can work with the conventional document summarization method in various domains. Through the investigation of the sentence compressor, we found that pre-trained word vectors contribute to performance improvement. We also investigated the knowledge graph embedding, which is necessary when we enhance word vectors by external knowledge. We provided a theoretical background for selecting a suitable loss function to support the training for knowledge graph embedding. Eventually, we incorporated our sentence compressor into the conventional document summarization method. We observed a performance improvement of automatic evaluation in the sentence extraction summarization setting.

研究分野：自然言語処理

キーワード：自動要約　文抽出　文圧縮　自然言語生成

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

研究開始当初、文書を入力とする生成的な要約手法はニューラルネットワークに基づくものが主流となっていた。これらのニューラルネットワークに基づく手法の多くはエンコーダ・デコーダの改良モデルを使用することで要約を生成している。しかし、エンコーダ・デコーダは当初、文から文への翻訳を実現するために提案された手法であり、文書のような長い入力を扱うことは想定されてはいない。そのため、エンコーダ・デコーダが文書のような長い入力を扱うための構造として必ずしも適切であるとは限らない。その理由の一つとして、エンコーダ・デコーダは、入力された情報に対して、入力が行われるたびに逐次的な出力を行うということはせず、文書の情報が全て入力されてからデコーダを通した出力を行う、という点が挙げられる。この動作のために、エンコーダ・デコーダが入力に対して適切な出力を行うには、一度文書中の全情報をネットワーク中に保持しなければならないことになる。さらに、既存のエンコーダ・デコーダでは、一度出力した結果を書き換えるということは行わないため、一度の出力で誤りのない出力を生成しなければならない。このためには、デコーダが出力を開始する際、つまり、エンコーダが入力の情報を全て読み取った時点で、これから行う出力の先読みを行う必要がある。このように、既存のエンコーダ・デコーダでは入力情報の保持、出力結果の先読みをネットワーク上の同じ箇所で同時に行わなければならず、入力情報の増大に対して頑健ではないという欠点が存在した。

２．研究の目的

本研究の目的は文書から適切な要約を生成することである。その実現のために人間が行っているような、段階的な要約の生成過程を考慮したニューラルネットワークに基づく手法を提案する。当時主流となっていたニューラルネットワークに基づく要約手法では、抽出的な要約の手順を踏まえた上で生成を行う手法は少数しか存在していなかった。従来手法では、抽出型の要約手法を用いて、文書中の重要文を選択した上で、選択された文を元に、最終的な要約を生成するといった二段階の手順からなる方法が用いられていた。しかし、このような手法は、文への選択に誤りが生じた場合、もしくは生成時に重要箇所が欠落するなどの誤りが生じた際に対処を行うことが困難である。本研究では文の抽出のみには基づかない、文選択の誤りを後段の処理で緩和可能な手法の提案により要約精度の向上を目指した。具体的にはこの目的を達成するために、計画を通して実際に段階的な要約の生成過程を考慮したニューラルネットワークに基づく文書要約器を実装し、長い文書を正しく要約することが可能となったかの確認を行う。

３．研究の方法

当初の計画は次に示すようなものであった。まず、初年度に段階的な要約の生成を考慮したエンコーダ・デコーダを作成する。このモデルは、**エンコーダ**、**デコーダ**、**出力部**によって構成される。このうち、エンコーダとデコーダはブロック構造を再帰的に積み上げることによって構成されている。このモデルにおいて要約の生成は次の手順で行われる。まず、**エンコーダ**において入力された単語は、各単語ごとに異なるベクトルへと変換され、入力文書中での単語の位置を表すベクトルと結合された後に、自己注意層へと入力される。次の自己注意層では単語ベクトル間の内積に基づいて、各単語ベクトルの重み付き和を計算する。全結合層では各ベクトルの異なる次元間の関係を考慮した変換を行う。そしてプーリング操作によって各単語ごとに存在するベクトルは入力文書中の各文を代表する固定長のベクトルへと変換され、単語ベクトルに行われたものと同様の操作を経て、文書を代表する固定長の文書ベクトルへと変換される。文選択層と単語選択層は、この文書ベクトルを用いて、重要な文と単語への重み付けを行う。単語選択層及び文選択層の出力は各単語ベクトル及び文ベクトルと共に再帰的に自己注意層へと入力され、前回と同様の処理が行われる。**デコーダ**では、出力部で出力された単語列を、エンコーダと同様の手順でベクトルへと変換し、階層注意層においてエンコーダの出力と組み合わせることで、入力文書中の重要な単語への重み付けを行う。その後の計算についてもエンコーダと同様である。**出力部**はデコーダの階層注意の結果に基づいて入力文書中の単語をそのまま出力しようとするコピー出力層と、デコーダの最終層の出力に基づいて生成する単語を決定するソフトマックス層によって構成される。コピー出力を行うことにより提案モデルは広範な語を出力することができる。最終的な出力単語はこの二つの出力の混合分布によって決定される。

二年目では、このモデルに対して GPU 計算機を用いて、既存の生成的な要約用の配布データセットで学習・予測をし、自動評価を行うことで、段階的な生成を仮定することによる要約精度向上の確認を行う。また、階層的注意層の出力分布を可視化することにより、モデルの内部ではどのような生成手順が仮定されているかを調査し、必要に応じて提案モデルを修正する。そして、最終的に構成されたモデルに対して自動及び人手評価を実施する。人手評価では冗長性、流暢性、重要情報が残されているか、の三つの尺度を用いる。これらの比較を通じて、ニューラルネットワークに適した要約の生成過程を明らかにするための知見を得る。

４．研究成果

初年度である 2019 年度には前年度に登場した強力な事前学習モデルである BERT の要約分野への浸透により、当初作成する必要があったエンコーダ・デコーダと類似した構造を持つモデルが一般的に公開されることとなった。そのため公開されているエンコーダ・デコーダに追加する形で、最終的に目指している階層的な要約を実現することを計画し、その上で重要となる文圧縮器

の実現に注力した。エンコーダ・デコーダに基づく文圧縮器の弱点である，左から右への単一方向のデコードにより、将来時刻において読み込む単語を事前に考慮することが難しいという問題に対処するために、依存構造木の情報を用いてデコーダからエンコーダへの注意を教師あり学習する手法を提案した。Google Sentence Compression データセットを用いた人手評価の結果、提案手法により出力された圧縮文は、文法性を低下させることなく情報性を向上させることが判明した。(表 1)

表 1 出力された圧縮文に対する人手評価の結果

手法	文法性	情報性
Tagger	3.90	3.79
Base	3.86	3.80
Parent w/ syn	3.82	3.77
Child w/ syn	3.94	3.85
提案手法	3.91	3.90

また Broadcast News Corpus を用いて、ドメイン外のデータセットにおける提案手法の性能を評価したところ、従来のエンコーダ・デコーダに基づく文圧縮器の出力と比較して良好な結果を示し、依存構造情報がドメイン外データにおいても有効であることを示した。(表 2)

表 2 ドメイン外のデータセットにおける各手法の性能

	F_1	正解の圧縮率との差分
Tagger	54.7	-39.1
LSTM	54.8	-39.2
LSTM-Dep	55.1	-38.8
Attn	54.1	-39.6
Base	55.4	-38.6
Parent w/ syn	54.2	-39.1
Parent w/o syn	54.0	-40.1
提案手法 (文法情報あり)	57.7	-35.9
提案手法 (文法情報なし)	54.6	-39.5

さらに、開発データにおいて、事前学習モデルを比較した結果、文脈情報のみではなく、単語埋め込みの持つ情報が効果的であることが判明した。(表 3)

表 3 開発データにおける事前学習手法の有効性の比較 (✔:その手法を使用)

Glove	✔	✔		✔	✔		
ELMo	✔	✔	✔			✔	
BERT	✔		✔	✔			✔
F_1	86.2	86.0	85.9	85.4	85.5	85.9	84.8

本研究の成果を情報処理学会第 243 回自然言語処理研究会で発表[1]し、優秀賞及び 2020 年度山下記念研究賞を受賞した。また、この研究内容を発展させたものが人工知能分野のトップ国際会議 Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)にて採択された。[2]

次年度には前年度に得た結果に基づき、要約で重要な単語の埋め込みをより詳細に扱うために知識グラフの埋め込みに関する研究と、前年度に作成された文圧縮器の文書要約への適用を進めた。知識グラフの埋め込みに関する研究においては、学習時に使用される負例サンプリングとソフトマックス交差誤差関数の関連及び差異について理論的に探求し、両者が同一の傾向を示す条件を導いた。この内容を言語処理学会第 27 回年次大会（NLP2021）にて発表し[3]委員特別賞を受賞した。またこの内容を発展させた研究を自然言語処理分野のトップ国際会議である The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing に投稿し、採択された。[4]

最終的な目的である段階的な要約過程を考慮した文書要約の実現のために文圧縮器の BERT に基づく文書要約手法への適用を行った。文圧縮器の適用においては、BERT の性能をより引き出すために、前年度に使用した注意の教師あり学習ではなく、ニューラルネットワークに基づくグラフ埋め込み法を適用することで、さらなる性能向上を果たした。この構造の改良により、Google Sentence Compression および Broadcast News Corpus の両データセットにおいて、さらなる性

能の向上を確認している。この文圧縮器を利用して　BERT で段階的な文書要約を行う際に必要となるデータセットを作成し、このデータを利用して　BERT で段階的な文書要約を行うことにより、抽出型の設定では CNN/Daily Mail データセットにおいて既存の手法よりも高い ROUGE スコアを示すことを確認した。現在、この成果を The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)に投稿することを計画している。

[1] 上垣外英剛, 奥村学. 2019. "階層的な注意機構に基づき統語的な先読みを行う文圧縮", 情報処理学会第 243 回自然言語処理研究会 (NL 研). 優秀賞, 2020 年度山下記念研究賞
[2] Hidetaka Kamigaito and Manabu Okumura. 2019. "Syntactically Look-Ahead Attention Network for Sentence Compression", In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), pp. 8051-8057, (2020).
[3] 上垣外英剛,林克彦. 2021. "知識グラフ埋め込み学習における損失関数の統一的解釈 ”,言語処理学会第 27 回年次大会(NLP2021). 委員特別賞
[4] Hidetaka Kamigaito and Katsuhiko Hayashi. 2021. " Unified Interpretation of Softmax Cross-Entropy and Negative Sampling: With a Case Study for Knowledge Graph Embedding ", In Proceedings of The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), (2021) (To be appeared).

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　2件）

１．発表者名
上垣外英剛,林克彦

２．発表標題
知識グラフ埋め込み学習における損失関数の統一的解釈

３．学会等名
言語処理学会第27回年次大会(NLP2021)

４．発表年
2021年

１．発表者名
Hidetaka Kamigaito and Katsuhiko Hayashi

２．発表標題
Unified Interpretation of Softmax Cross-Entropy and Negative Sampling: With a Case Study for Knowledge Graph Embedding

３．学会等名
The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)（国際学会）

４．発表年
2021年

１．発表者名
上垣外 英剛, 奥村 学

２．発表標題
階層的な注意機構に基づき統語的な先読みを行う文圧縮

３．学会等名
情報処理学会第243回自然言語処理研究会

４．発表年
2020年

１．発表者名
Hidetaka Kamigaito, Manabu Okumura

２．発表標題
Syntactically Look-Ahead Attention Network for Sentence Compression

３．学会等名
Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)（国際学会）

４．発表年
2020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

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