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　２ 年　 ６ 月 １０ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2017～2019
課題番号：１７Ｈ０１７８６

研究課題名（和文）ニューラルモデルと離散最適化技術を用いた高品質な要約作成

研究課題名（英文）Text Summarization Based on the Combination of Neural Models and Optimization Technologies

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）　　13,500,000円

研究成果の概要（和文）：ニューラルモデルを用いた要約手法の性能向上を図るため，1) 構文，談話情報を組み込んだニューラル要約モデルの提案，2) 1)のモデルを構築するのに不可欠な談話構造解析の新しい手法の提案を行った．1)では，構文木の情報を考慮した文圧縮手法，談話構造解析木の情報を考慮した文選択手法を提案した．一方2)では，教師あり，教師なしの2つの談話構造解析手法を提案したが，提案した教師あり手法は，現在の世界最高性能を達成している．

研究成果の学術的意義や社会的意義
Sequence-to-Sequence（Seq2Seq)モデルに基づく文圧縮では，すでに圧縮文に採用された単語列とこれから圧縮文に採用しようとする単語との間の文法的な依存関係を明示的に捉える事が難しい為，デコード時に階層的な注意機構に基づき構文的な先読みを行う事が可能なモデルを提案した．Seq2Seqモデルに基づく抽出型手法が単一文書要約において良い性能を示しているが，文間の談話構造は明示的には利用しない．談話構造に関する情報の欠如は，重要度スコア決定における性能劣化や出力要約の一貫性の低下を引き起こす為，原文書の談話構造と文の重要度スコアリング器を同時に学習する新たな枠組みを提案した．

研究成果の概要（英文）：In order to improve the performance of summarization methods using a neural model, 1) we proposed neural summarization models incorporating syntactic and discourse information. 2) We proposed a new method for discourse structure analysis to construct the model for 1). In 1), we proposed a sentence compression method that considers the information of syntactic trees and a sentence selection method that considers the information of discourse structure trees. In 2), we proposed two
discourse structure analysis methods, with and without supervision.
The proposed supervised method achieves the current state-of-the-art performance.

研究分野：知能化工学

キーワード：自然言語処理

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年我々人間の周囲には，さまざまなメディアを通した情報が満ち溢れ，情報洪水という言葉が使われるようになってからかなりの歳月を経ている．そのため，テキスト情報の重要な部分のみを提示するテキスト要約技術が注目を集め研究が活発になるとともに，WWW 上のニュースサイトなどでは，要約機能が実際に一般に提供され始めている．テキスト要約の分野は歴史も古く，すでに 60 年ほどの期間研究されているが，要約対象のテキスト（集合）（以下，原文書と呼ぶ）から表現を抽出する抽出型要約が主流であった．

一方で，テキスト要約の分野では，いくつかのブレイクスルーとも呼べる画期的な成果が近年得られている．第一には，近年の深層学習技術の成功により，原文書中の表現を用いないで要約を生成できる，非抽出型の文要約を高精度に実現するモデルが提案され始めている．研究成果欄で述べるように，我々は，ニューラルモデルの 1 つである encoder-decoder モデルによる非抽出型の文要約において出力文の長さを制御する手法を世界に先駆けて開発し[1]，モデルを実際にテキスト要約に利用できる道筋をつけることに成功した．第二には，テキスト要約に離散最適化技術を適用することで，最適な表現集合を選択することができ，それにより，テキスト要約の性能を画期的に向上できることが分かってきている．我々も，テキスト要約を最大被覆問題などとしてモデル化することで，単一テキスト要約，複数テキスト要約に対する，離散最適化技術を用いた要約手法を提案している（たとえば，[2]）．

２．研究の目的

そこで本研究課題では，我々がこれまでに開発してきたテキスト要約の要素技術を援用し，さらに高度化するとともに，後述する a), b), c)の 3 つのテキストの自動評価手法を開発することで，単一テキスト要約，複数テキスト要約どちらにも適用可能な，高品質な要約を作成できるテキスト要約手法を確立することを目的とする．以下目的を達成するための副目標について述べる．

1) 我々の開発している文要約モデルを，長さだけでなく，原文書中の表現を用いる度合い（抽出度）も制御できるように拡張することで，テキスト要約手法の中で利用可能な，抽出型，非抽出型どちらも実行可能な文要約モデルを完成する．

2) 離散最適化技術を用いたテキスト要約では，候補となる文集合の中から，定義した評価関数を元に，最適な要約テキスト（文集合）を選択し出力する．本研究課題で開発するテキスト要約の枠組みにおいては，1)で開発する文要約モデルを用いて原文書中のすべての文から要約文の候補を生成し，それらの組み合わせを要約テキストの候補とする．テキスト要約の要素技術としては，（抽出型，非抽出型両方の）文要約，文選択，文融合の 3 つが挙げられているが[3]，上述した枠組みでは，まだ文融合が実現できていない．そこで，これまで実現が困難とされていた文融合の要約技術を新たに研究開発するとともに，それを上述したテキスト要約の枠組みに統合する．さらに，3)で述べる評価基準の 1 つである一貫性の高いテキスト作成のために必要な，要約テキストの推敲モジュールを研究開発する．

3) これまで，要約テキストは，(1) 文章としての質（テキストとして読み易いか），(2) 内容の情報量（原文書の重要な情報をカバーしているか），の大きく 2 つの評価基準で評価することが一般的であった（たとえば，[4, 5]）．内容の情報量を自動評価する尺度としては，ROUGE が以前から離散最適化技術を用いた要約でも用いられている．しかし，文章としての質に関しては，自動評価尺度がこれまでなく，テキスト要約の枠組みに取り込むことができないことから，高品質な要約を作成する上でボトルネックとなっていた．そこで本研究課題では，a)文の文法性，b) テキストの一貫性，c) テキストの misleadingness（誤解しやすさ）の 3 つの観点から文章としての質を自動評価する尺度を新たに研究開発し，（要約）テキストの質の自動定量評価手法を確立した上で，2)のテキスト要約の枠組みで評価関数として利用する．

３．研究の方法

研究目的欄で述べた 3 つの副目標を達成するためには，以下の 5 つが研究の柱となる．

a．我々の開発している文要約モデルを，長さだけでなく，原文書中の表現を用いる度合い（抽出度）も制御できるように拡張することで，テキスト要約手法の中で利用可能な，抽出型，非抽出型どちらも実行可能な文要約モデルを完成する．

b．これまで実現が困難とされていた文融合の要約技術を新たに研究開発する．

c．一貫性の高いテキスト作成のために必要な，要約テキストの推敲モジュールを研究開発する．

d．文の文法性，テキストの一貫性，テキストの誤解しやすさの 3 つの観点から文章としての質を自動評価する尺度を新たに研究開発する．

e．a，b，c，d で研究開発したテキスト要約の要素技術を，離散最適化技術を元に 1 つのテキスト要約の枠組みに統合した上で，単一テキスト要約，複数テキスト要約どちらにも適用可能なテキスト要約手法を完成し，実際のテストセットを用いて評価実験を行う．

４．研究成果

柔軟に抽出度を制御できる文要約モデルの研究開発においては，我々がこれまで開発してきている，出力文の長さを制御できる encoder-decoder モデルによる文要約手法を拡張することで，原文書中の表現を用いる度合い（抽出度）も制御できるようにした．

テキストの自動評価尺度の研究開発は，良いテキスト（たとえば，文法的な文）のモデルを大量のデータから教師なしで学習し獲得する方法と，良いテキストかどうかを識別する分類器を教師有り学習を用いて学習する方法に大別できる．要約文の自動評価への受容可能性尺度の適用可能性を吟味した上で，その問題点を検討し，要約文の自動評価に適した形に受容可能性尺度を改良した．テキストの誤解しやすさに関しては，大量のデータを安価で入手するのは困難であることから，要約が元テキストの内容を誤解するようなものになっているかどうかを自動的に判別できる分類器を研究開発した．要約は，一般にテキストであり複数の文から構成される．テキスト要約手法には一般に，研究目的欄で述べたように，文選択，文要約，文融合の 3 つの手法が主に用いられている．そのため，要約作成時において誤解を招くようなテキストを作成してしまう可能性としては，文選択で，選択した複数の文により要約テキストを構成する段階と，文要約あるいは文融合で，要約文を構成する段階が考えられることになる．それぞれの段階において得られる要約テキストがどのような場合に誤解を招くものとなり，どのような場合にそうならないかを明らかにするため，一定量の要約テキストに対して，誤解を招くテキストかどうかのラベルを付与したデータを作成した．

また，我々が世界に先駆けて開発した，ニューラルモデルの 1 つである encoder-decoder モデルによる非抽出型の文要約において出力文の長さを制御する手法[1]を受け，出力長を考慮した文要約モデルを評価するためのコーパス JApanese MUlti-Length Headline Corpus (JAMUL)を提案，開発した．

次に，ニューラルモデルを用いた要約手法の性能向上を図るため，1）構文，談話情報を組み込んだニューラル要約モデルの提案，2）1)のモデルを構築するのに不可欠な談話構造解析の新しい手法の提案を行った．1)では，構文木の情報を考慮した文圧縮手法，談話構造解析木の情報を考慮した文選択手法を提案した．Sequence-to-Sequence (Seq2Seq)モデルに基づく文圧縮では，すでに圧縮文に採用された単語列とこれから圧縮文に採用しようとする単語との間の文法的な依存関係を明示的に捉えることが難しいため，デコード時に階層的な注意機構に基づき構文的な先読みを行うことが可能な Seq2Seq モデルを提案した．

また，原文書の談話構造を考慮する新たな抽出型ニューラル要約モデルを提案した．修辞構造理論に代表される談話構造を表現する枠組みは，文書中の文や単語などの間に内在する意味的なつながりに着目する．リカレントニューラルネットワーク (RNN)に基づく抽出型手法が 2016 年以降，単一文書要約において良い性能を示している．この手法は原文書を文の系列とみなしベクトル化し文の重要度を決定し，文間の談話構造は明示的には利用しない．談話構造に関する情報の欠如は，重要度スコア決定における性能劣化や出力要約の一貫性の低下を引き起こす可能性がある．そこで，談話構造解析器の解析誤りによる影響を抑えながら，RNN を用いた要約モデルの性能における利点を活用するため，原文書の談話構造と文の重要度スコアリング器を同時に学習する新たな枠組みを提案した．DailyMail データセットを用いた評価実験において，提案手法がベースラインよりも ROUGE 値および人手評価において良い評価値を得た．さらに，既存の性能の良い手法と同等もしくは，より良い結果を得た．

一方 2)では，教師あり，教師なしの 2 つの談話構造解析手法を提案したが，提案した教師あり手法は，現在の世界最高性能を達成している．従来の談話構造解析手法の多くは葉ノードである EDU から開始し，それらをボトムアップに組み上げていくことで談話構造木を構築している．しかし，ボトムアップな解析手法は出力結果が葉ノード近辺の解析結果に依存しやすくなってしまう傾向があり，応用タスクにおいて利用価値が高い情報が根ノード近辺に存在していることから望ましくない．この問題を踏まえトップダウンに談話構造解析を行う手法を提案した．

[1] Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura and Manabu Okumura, Controlling Output Length in Neural Encoder-Decoders, EMNLP 2016, 2016.
[2] 高村大也，奥村 学，最大被覆問題とその変種による文書要約モデル，人工知能学会論文誌，Vol.23，No.6，pp.505-513，2008.
[3] Hongyan Jing and Kathleen R. McKeown, Cut and Paste Based Text Summarization, Proc. of the 1st Annual Meeting of the North American chapter of the Association for Computational Linguistics, pp.178-185, 2000.
[4] 奥村学，難波英嗣,テキスト自動要約,オーム社，2005.
[5] Ani Nenkova and Kathleen McKeown, Automatic Summarization, Foundations and Trends in Information Retrieval, Vol 5, No 2-3, pp. 103-233, 2011.

５．主な発表論文等

〔雑誌論文〕　計8件（うち査読付論文　8件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Hidetaka Kamigaito and Manabu Okumura	４．巻 -
２．論文標題 Syntactically Look-Ahead Attention Network for Sentence Compression	５．発行年 2020年
３．雑誌名 Proc. of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)	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） なし	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 Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura and Masaaki Nagata	４．巻 -
２．論文標題 Top-down RST Parsing Utilizing Granularity Levels in Documents	５．発行年 2020年
３．雑誌名 Proc. of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)	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） なし	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 Naoki Kobayashi, Tsutomu Hirao, Kengo Nakamura, Hidetaka Kamigaito, Manabu Okumura and Masaaki Nagata	４．巻 -
２．論文標題 Split or Merge: Which is Better for Unsupervised RST Parsing?	５．発行年 2019年
３．雑誌名 Proc. of EMNLP-IJCNLP 2019, 2019	６．最初と最後の頁 5797-5802
掲載論文のDOI（デジタルオブジェクト識別子） 10.18653/v1/D19-1587	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 Tatsuya Ishigaki, Hidetaka Kamigaito, Hiroya Takamura and Manabu Okumura	４．巻 -
２．論文標題 Discourse-aware Hierarchical Attention Network for Extractive Single-Document Summarization	５．発行年 2019年
３．雑誌名 Proc. of RANLP 2019, 2019	６．最初と最後の頁 497-506
掲載論文のDOI（デジタルオブジェクト識別子） 10.26615/978-954-452-056-4_059	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 Hirao, T., Kamigaito, H. and Nagata, M.	4．巻 -
2．論文標題 Automatic Pyramid Evaluation Exploiting EDU-based Extractive Reference Summaries	5．発行年 2018年
3．雑誌名 Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing	6．最初と最後の頁 4177-4186
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 小林尚輝，平尾努，中村健吾，上垣外英剛，奥村学，永田昌明	4．巻 -
2．論文標題 テキストセグメンテーションによる教師なし修辞構造 解析	5．発行年 2019年
3．雑誌名 言語処理学会第25回年次大会発表論文集	6．最初と最後の頁 998-1001
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 小林尚輝，平尾努，上垣外英剛，奥村学，永田昌明	4．巻 -
2．論文標題 階層構造を考慮したトップダウン談話構造解析	5．発行年 2019年
3．雑誌名 言語処理学会第25回年次大会発表論文集	6．最初と最後の頁 1002-1005
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 人見雄太，田口雄哉，田森秀明，菊田洸，西鳥羽二郎，岡崎直観，乾健太郎，奥村学	4．巻 -
2．論文標題 出力長制御を考慮した見出し生成モデルのための大規模コーパス	5．発行年 2019年
3．雑誌名 言語処理学会第25回年次大会発表論文集	6．最初と最後の頁 6-11
掲載論文のDOI（デジタルオブジェクト識別子） なし	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　4件）

1．発表者名
長谷川駿，上垣外英剛，奥村学

2．発表標題
生成型文要約のための抽出性に着目したデータ選択

3．学会等名
情報処理学会 第241回自然言語処理研究会

4．発表年
2019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手法

3．学会等名
第243回自然言語処理研究会

4．発表年
2020年

1．発表者名
石垣達也，上垣外英剛，高村大也，奥村学

2．発表標題
談話構造を考慮する階層的注意機構による抽出型ニューラル単一文書要約

3．学会等名
言語処理学会第25回年次大会

4．発表年
2019年

1．発表者名
Hidetaka, K., Hayashi, K., Hirao, T., Takamura, H., Okumura, M. and Nagata, M.

2．発表標題
Supervised Attention for Sequence-to-sequence Constituency Parsing

3．学会等名
the 8th International Joint Conference on Natural Language Processing (IJCNLP)（国際学会）

4．発表年
2017年

１．発表者名
Tatsuya Ishigaki, Hiroya Takamura and Manabu Okumura

２．発表標題
Summarizing Lengthy Questions

３．学会等名
the 8th International Joint Conference on Natural Language Processing (IJCNLP)（国際学会）

４．発表年
2017年

１．発表者名
Shun Hasegawa, Yuta Kikuchi, Hiroya Takamura and Manabu Okumura

２．発表標題
Japanese Sentence Compression with a Large Training Dataset

３．学会等名
ACL 2017（国際学会）

４．発表年
2017年

１．発表者名
Kamigaito, H., Hayashi, K., Hirao, T. and Nishino, M.

２．発表標題
Higher-order Syntactic Attention Network for Longer Sentence Compression

３．学会等名
the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (to appear).（国際学会）

４．発表年
2018年

１．発表者名
石垣 達也，高村 大也，奥村 学

２．発表標題
「長文質問」のための抽出型及び生成型要約

３．学会等名
情報処理学会第232回自然言語処理研究会

４．発表年
2017年

１．発表者名
上垣外英剛，林克彦，平尾努，永田昌明

２．発表標題
依存構造の連鎖を考慮したニューラル文圧縮

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

１．発表者名
平尾努，上垣外英剛，永田昌明

２．発表標題
抽出型オラクルを利用した要約の自動評価

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

１．発表者名
金澤尚史，高村大也，奥村学

２．発表標題
文書要約のための一貫性モデル

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

１．発表者名
牧野拓哉，岩倉友哉，高村大也，奥村学

２．発表標題
Minimum Risk Training に基づく要約モデルの出力長制御

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

〔図書〕　計0件

〔出願〕　計2件

産業財産権の名称 談話構造解析装置	発明者 平尾努，永田昌明， 小林尚輝，奥村学	権利者 同左
産業財産権の種類、番号 特許、特願2019-028629	出願年 2019年	国内・外国の別 国内

産業財産権の名称 木構造解析装置	発明者 平尾努，永田昌明， 小林尚輝，奥村学	権利者 同左
産業財産権の種類、番号 特許、特願2019-035758	出願年 2019年	国内・外国の別 国内

〔取得〕　計0件

〔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	平尾　努 (Hirao Tsutomu) (40396148)	日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・協創情報研究部・主任研究員 (94305)	
研究分担者	高村　大也 (Takamura Hiroya) (80361773)	東京工業大学・科学技術創成研究院・教授 (12608)	