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　 ７ 月　 ６ 日現在

機関番号：２５３０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００４４１

研究課題名（和文）抽象化を伴う段落タイトルの自動生成に関する研究

研究課題名（英文）Abatractive Generation of Paragraph Titles

研究代表者
菊井　玄一郎（KIKUI, Genichiro）

岡山県立大学・情報工学部・教授

研究者番号：８０３９５０１１

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：文章の概要、特に、論説文における議論の流れを簡潔に提示することを目的として、各段落に簡潔なタイトル（「段落タイトル」）を自動的に付与する手法を検討した。研究用のデータがほとんど存在しないことから、本研究では、小論文123文章の各段落に対して、５名の作業者が１つ以上の段落タイトルを付与したコーパスを新たに構築した。このコーパスに対する単語統計的な検討、および、深層学習に基づく既存の要約手法（タイトル生成手法）の適用結果により、段落タイトルを自動生成するにはabstractiveな要約手法が必要であること、既存手法では不十分な新たな課題を含んでいることが分かった。

研究成果の学術的意義や社会的意義
学術的意義は３点ある。１点目は論説文の議論の流れを簡潔に明示する手段としての段落タイトルに注目し、それらを３つに分類したことである。２点目は約120文章（総段落数786）の各段落に対して5つ以上の段落タイトルを付与したコーパスを構築し、その統計的性質や既存手法の限界などを明らかにしたことである。作成したコーパスは当該分野の研究に寄与できるものと思われる。３点目はタイトルの自動生成に必要な語義の扱い、特に、じでょ未登録語の意味を推定する手段を示したことである。
社会的意義は氾濫するテキスト情報の閲覧を支援する手段として段落タイトルの位置づけとその性質を明らかにしたことである。

研究成果の概要（英文）：This research aims at developing models for generating a title for each paragraph of an English text. A paragraph title is a short linguistic expression which indicates or summarizes information of the given paragraph. A sequence of paragraph titles is useful representation of the text, expressing its argumentation line.
In this work, we created a corpus of paragraph titles, composed by humans. We found that 46% of word tokens in a title do not appear in the corresponding paragraph in average, which means that we need ‘abstractive’ summarization. We, then, applied state-of-the-art title generation models, such as encode-decoder models and transformer models, to our corpus and found that two models produced relatively good performance at 34 rouge-1 score, but rated as ‘does not include main idea’ in average by human evaluators . This means that the corpus can provide a challenging task for abstractive title generation.

研究分野：自然言語処理

キーワード：自動タイトル付与　自動要約　自然言語処理　索引付け

１．研究開始当初の背景

我々の周囲には膨大な電子化テキストが存在しており，これらから効率良く情報を獲得することが求められている．そのための代表的な技術として，検索（テキスト検索）と自動要約が研究開発されている．このうち検索は利用者が望む情報をテキスト単位（文書単位），あるいは，文書中の該当部分をピンポイントで特定することを目指す技術である．テキスト単位の検索の場合は検索結果のテキストから必要な情報を再度検索したり要約したりすることが必要となり，ピンポイントの場合は文脈が無視されるという危険性がある．一方，自動要約は要約率が高いとやはり文脈が失われる可能性があり，低いと冗長になるという問題がある．既存の自動要約技術のうち，修辞構造を利用した方法は議論の骨格を提示できる可能性があるが，長い文章に対して，各段落でどのような議論をしているかを示すような要約を生成する方法は自明ではない．

この問題に対し，報告者らは文章の各段落にその内容を端的に表すタイトル (**段落タイトル**) を付与し，段落タイトルの列によって文章を表現することを考えた．計画当時，報告者らは「ロボットは東大にはいれるか」プロジェクト [1] に参画し，センター試験の英語問題に取り組んでいた ([2] など)．英語センター試験では 2000 年代以降，論説文の読解問題（第 6 問 B）において，文章の各段落に対してその内容を表す短い言語表現を選択肢から選んで答えさせる問題（段落タイトル付与問題）が出題されている．この問題への取り組みから，**段落タイトル を自動生成すること**が文章の議論の流れを簡潔に提示する上で有効であることが分かった．しかしながら，当時は，段落タイトルの候補が選択肢として与えられているにも関わらず，問題の正解率が 43% 程度と低く，段落タイトルを生成するためには，段落内容とこれを抽象化した (abstractive な) 言語表現との関係をモデル化が必要であるとの考えに至った．

２．研究の目的

上述の背景のもと，本研究では論説的テキストの 1 つの段落（形式段落あるいは意味段落）に対して，その内容を簡潔に表すタイトル（見出し表現）を自動的に生成する手法を明らかにする．ここで，タイトルとは数語程度の名詞句あるいは短い文（単文）とする．本研究計画では特に段落テキストの抽象化（abstraction）が必要な場合に焦点をあて，機械学習に基づく手法を検討する．また，その際に必要となるデータの整備を行う．

３．研究の方法

研究開始当初の計画に従って，次の 3 つの項目に分けて進めた．

1. コーパスの構築と（人手による）分析
2. （タイトル生成に必要となる）文章解析ツールの整備
3. 段落タイトル自動生成手法の検討

ただし，それぞれの項目の内容は研究過程で得られた知見や関連分野の研究動向に対応して当初計画を変更した部分がある．以下では実際の取り組み内容を中心に説明する．

３．１　コーパスの構築と分析

文章全体に対するタイトルが多くの文章において存在するのに対して，文章の各段落に対するタイトルは通常存在しない．文章の作成に先立って，各段落で何を述べるかを記述したアウトラインが（著者らによって）作られていれば段落タイトルとみなすことができるが，一般にアウトラインを入手することは困難であるし，完成した文章がアウトラインを忠実に反映したものである保証もない．そこで本研究では新たに次の 2 つのコーパスの作成を計画した．一つ目（コーパス 1）は論説文のタイトルを大量に収集したものであり，二つ目（コーパス 2）は論説文（できれば 2000 文字以内の短めのもの）を収集し，各段落にタイトルを付与したもの

である．本来の段落タイトルはコーパス２であるが，作成にコストがかかる．一方，コーパス１は既存の文章のタイトルを流用するので比較的コストが低いが，文章全体（複数段落）のタイトルであるから段落との対応が取れない．そこで，コーパス１をタイトルの言語的・意味的特徴の分析・推定に利用し，コーパス２を段落タイトル生成手法の評価，および，パラメータ推定に利用することを計画した．

しかし，実際には，コーパス１の構築は行わなかった．これは，英語センター試験の段落タイトル付与問題の分析により，段落タイトルの内容のタイプ分けが提案されたこと [3]，タイトル付きの比較的短い論説文の大規模な収集が困難だったことなどによる．コーパス２については段落タイトル付与対象の「研究利用可能な英語小論文（エッセイ）コーパス」の調査，利用許可取得に時間がかかったため計画よりかなり遅れたものの，長さ 500〜800 単語程度の 3 種類の小論文コーパスの各段落に対して５人の異なる作業者によりそれぞれ１つ以上のタイトルを付与したコーパスを構築した．

３．２　テキスト構造解析基盤の構築

当初計画では，主に既存の技術をカスタマイズすることによりタイトル付与に必要な処理を整備するとし，具体的な処理として，品詞タギング，構文解析，述語項構造解析などの文レベルの解析，共参照解析，修辞構造解析，重要文・重要語抽出を想定していた．しかし、近年の自然言語処理においては単語列を入力とした深層学習技術の利用が進んできたことから，これらの基本ツールについては Stanford 大（英語）や京都大学等（日本語）の公開しているツールを必要に応じて用いることとし，新たなカスタマイズは行わなかった．その代わりに，深層学習モデル利用において重要となる，単語埋め込みについて，辞書未登録語の埋め込みの推定，同義語・反義語判別モデルの検討に注力した．

３．３　タイトル自動生成手法

タイトル自動生成手法については計画段階において抽象度の高いルールの作成・利用を想定していたが，研究開始後に transformer model(転移学習モデル) が発表され，自動翻訳や機械読解における高い性能が明らかになったことから，本研究においてもこれらを利用した自動要約（見出し生成）手法を今回作成したコーパスに適用することとした．なお大学入試センター試験の英語における段落タイトル付与問題については転移学習モデルの一つである BERT をスコアリング関数とする手法により，９割を超える正答率を実現した [4]．このことは，転移学習モデルによって段落原文とそのタイトルの間の関係のモデル化が相当程度できていることを示している．深層学習モデルによる段落タイトル生成の評価においては rouge による自動評価に加えて，人手評価を行った．人手評価については生成される言語表現が短いことを考慮して本研究において新たに設計した評価指標を用いた．

４．研究成果

ここでは，まず，作成したコーパスについて述べ，次に，これを用いたタイトル生成について説明する．最後に，単語の埋め込みについて述べる．

４．１　コーパスの作成 [5]

次の３つの英語論説文コーパスを対象に文章の各段落に対して人手でタイトルを付与した．

1. MOECS: The Corpus of Multilingual Opinion Essays by College Students
2. LOCNESS: The Louvain Corpus of Native English Essays
3. JapanNews: the Japan News の editorial(社説)

タイトル付与作業者は大卒以上で文章の読解に習熟した英語ネイティブである．各文章は５名の作業者に割り当てられ，各作業者は与えられた文章の各段落に１つ以上のタイトルを付与

する．作成されたタイトルコーパスの概略を表 1 に示す．タイトルの平均長は 7.2 であり，もとの段落の長さとタイトルの長さの間に相関関係はみられなかった．

次に，タイトルに出現した単語のうち，元の段落に出現しているものの比率を表 2 に示す．表の「同一」列は当該タイトルが付与される段落に限定した場合,「前後」は前後 2 段落を含めた場合,「全体」はテキスト全体に出現する単語との重なりを表す．「完全一致」の行はタイトルの全体がそのまま対応する段落に出現する割合を表している．表より，同一段落に完全一致する文字列が存在する割合は 0.42% と殆ど存在しないことを示している．unigram でも対応する段落単語 (token) のカバー率は 53%，テキスト全体でも 66% であり，bigram, trigram になると急速にカバー率が下がる．このことから，今回人手で作成したタイトルは本文中の単語の組み合わせでは生成が難しいという意味で abstractive な処理が必要であることが分かる．

表 1　作成したコーパスの基本的な統計量

略称	総段落数	タイトル数	平均長
MOECS	341	1,761	7.3
LOCNESS	136	692	6.3
JapanNews	307	1,562	7.5
全体	784	4,015	7.2

表 2　タイトル表現の ngram が本文に含まれる率 (トークンベース，単位 %)

	同一	前後	全体
完全一致	0.423	0.523	0.697
1gram	53.7	60.9	66.5
2gram	13.7	16.8	20.4
3gram	4.72	5.62	6.83
1gram(CW)	40.9	47.3	53.4

４．２　タイトル生成実験 [5]

今回作成したコーパスに既存の見出し生成手法 (自動要約手法) を適用することにより，タイトルをどの程度うまく自動生成できるか試した．作成した全データをランダムに 3:1:1 に分けてそれぞれ訓練データ，検証データ，評価データとした．適用した自動要約手法は次の 4 つである．

RNN-RNN:　encoder-decoder モデルで encoder 側，decoder 側とも RNN(Recursive Neural Network) を使用し，段落タイトルコーパスの訓練データで学習したもの
BRNN-RNN:　前記と同様であるが encoder 側を双方向 RNN(BRNN) にしたもの
GPT2-pre:　GPT2 の生成器 (次単語予測) で段落を先行文脈として動かしたもの
BERTAbs:　BERT を利用した abstractive 要約モデルを CNN/Dailymail の headline コーパスで finetune したもの

生成したタイトルの品質評価は自動評価法の Rouge-1, 2, 3,L [5] で行い，高評価であった BRNN-RNN および GPT2-pre に対して人手評価を行った．結果を表 3 に示す．

表 3　タイトル生成実験結果 (R1,2,3,L はそれぞれ Rouge-1,2,3.L に対応する。)

	R-1	R-2	R-3	R-L	人手評価
RNN-RNN	29.0	3.66	0.68	34.3	-
BRNN-RNN	35.6	8.21	2.91	41.0	2.66
GPT2-pre	34.9	12.8	5.1	38.2	3.13
BERTAbs	27.2	8.17	3.10	30.27	-

まず自動評価では BRNN-RNN, GPT2-pre が相対的に高い結果となった．特に GPT2-pre は大規模なコーパスから学習した英語の言語モデルのみによって encoder-decoder モデルと同等の性能になっている．段落を先行文脈としてうまく言語表現を推定できる可能性を示している．自動評価値が相対的に良かった BRNN-RNN, GPT2-pre に対して英語の堪能な者 2 名による人手評価を行った．評価尺度の詳細は [5] に譲るが，4 以上が合格レベルである．評価値平均が 3 前後であることから，特に内容的な観点からやや厳しい，すなわち，適切に要

点を捉えていないという結果となった．なお，人手評価と自動評価 (Rouge-L) の相関係数は BRNN-RNN で 0.41, GPT2-pre で 0.21 と極めて低い．rouge のこのコーパスへの適用可能性については再検討の必要がある．

４．３　単語埋め込みの改良

４．３．１　未登録語の単語埋め込みモデル [6]

一般に単語の埋め込みは，あらかじめ各単語に対して作成して辞書に登録しておき，利用時には単に辞書引きするだけである．したがって辞書未登録語の扱いが問題となる．一つの方法は word piece のようなサブワードベースの手法により当該単語の（部分）文字列から推定する方法であり近年広く利用されている．もう一つの方法は本研究のように対象単語の文脈単語から推定する方法である．本研究では未登録語 w の埋め込み v_w を次式で推定する

$$v_w = Au_w$$

ここで，u_w は文脈ベクトル (簡単には文脈語のベクトルの和)，A は変換行列である． 行列 A は次式に示す semantic auto-encoder という考え方を用いて構築する．

$$\arg\min_{A \in R^{d \times d}} \sum_{w \in V} \|u_w - A^* A u_w\|_2^2$$

$$s.t. \quad Au_w = v_w - u_w$$

ここで V は語彙であり，$A^* = A^T$ とする．評価実験の結果，提案手法は従来法よりも少ない用例数により A が推定可能となることが分かった．

４．３．２　同義語と反義語の識別 [7]

単語の埋め込みは当該単語の周辺に出現する単語（文脈語）によって決定される．このため，意味的に近い単語（同義語）も反義語も同じような埋め込み表現となることが問題となっている．この問題に対して，本研究では，２つの単語 w_1, w_2 を前提として，単語 w_1 の共起語を１）w_1, w_2 双方の共起語になりえるもの，２）w_2 の共起語になりえないもの (w_1 のみの特徴に関連するもの)，３）どちらの共起語にもなりえないもの（データ上のノイズ）の３つ（潜在変数）から確率的に出現したものと考えてモデル化した．パラメータを EM 法で推定したところ，本手法は同義語・反義語識別に有効であることが明らかになった．

参考文献

[1] 新井紀子, 東中竜一郎. 人工知能プロジェクト「ロボットは東大に入れるか」. 東京大学出版会, 2018.

[2] 東中竜一郎, 杉山弘晃, 磯崎秀樹, 菊井玄一郎, 堂坂浩二, 平博順, 南泰浩. センタ試験における英語問題の回答手法. 第 21 回言語処理学会年次大会, pp. 187–190, 2015.

[3] 井内健人, 菊井玄一郎, 杉山弘晃, 但馬康宏. 表現の類似性と文書分類を併用したセンター試験英語段落タイトル付与問題の解答手法. 言語処理学会年次大会予稿集, pp. 785–794, 2017.

[4] 杉山弘晃, 成松宏美, 菊井玄一郎, 東中竜一郎, 堂坂浩二, 平博順, 南泰浩, 大和淳司. センター試験を対象とした高性能な英語ソルバーの実現. 第 26 回言語処理学会年次大会, pp. 371–374, 2020.

[5] 菊井玄一郎, 松岩祥平. 英語エッセイテキストに対する段落タイトル付与コーパス. 2020 年人工知能学会全国大会, 2020.

[6] 城内聡志, 菊井玄一郎. Semantic autoencoder を用いた低頻度語の埋め込み生成. 第 25 回言語処理学会年次大会, pp. 507–510, 2019.

[7] 城内聡志, 菊井玄一郎. 潜在変数モデルを用いた同義・反義関係識別. 第 25 回言語処理学会年次大会, pp. 141–144, 2020.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 城内聡志、菊井玄一郎
２．発表標題 潜在変数モデルを用いた同義・反義関係識別
３．学会等名 言語処理学会　第26回年次大会
４．発表年 2020年

１．発表者名 菊井玄一郎、松岩祥平
２．発表標題 英語エッセイテキストに対する段落タイトル付与コーパス
３．学会等名 2020年人工知能学会全国大会
４．発表年 2020年

１．発表者名 城内聡志
２．発表標題 Semantic Autoencoderを用いた低頻度語の埋め込み生成
３．学会等名 言語処理学会第２５回年次大会
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

「段落タイトル付与コーパス」を作成した。このコーパスは英語小論文123文書の各段落に対して５人の英語ネイティブ作業者が独立に１つ以上のタイトルを付与したものである。文献に示したようにこのコーパスは段落タイトル付与の研究に資するものと考えられるので、適切な機関を通じて公開予定である。

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