JSiSE Research Report
vol.38,no.4(2023-11)

# 学術論文を入力とする観点を反映した要約生成システムの構築

田邉 肇比古[*1]，リ ケンコン[*1]，長谷川 忍[*2]
*1 北陸先端科学技術大学院大学 先端科学技術研究科
*2 北陸先端科学技術大学院大学 遠隔教育研究イノベーションセンター

# Construction of a summary generation system that reflects the viewpoints of academic papers

Tanabe Hatsuhiko[*1], Li Jinghong[*1], Hasegawa Shinobu[*2]
*1Division of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
*2 Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology

The purpose of this research is to develop a summary generation system that reflects the viewpoints of academic papers. This is achieved by building a point-of-view classifier for the main text of an article and generating a summary using the sentences classified by the point-of-view classifier as input. The accuracy of the point-of-view classification is important for the final summary generation, and section information is used to improve the accuracy. We verified the possibility of useful summary generation by defining important sentences and learning the degree of importance.

キーワード: 学術論文，観点，自動分類，自動要約，深層学習，汎化性能

## 1. はじめに

研究者や学生が新たな領域で研究を始める際，関連領域の論文サーベイを通して先行研究の動向を理解する必要がある．査読を通過する前の論文を一般に公開するプレプリント・サーバーである arXiv(1)などにより Web 上で膨大な学術論文が日々出版される状況において，多くの論文を読み内容を理解することが重要である．このことから，論文の情報を効率的に収集及び整理し，より効果的な論文サーベイをサポートする要約生成技術が求められている．

論文に対する要約技術は，1 つの論文全体の要約に主眼が置かれている(2)．研究者や学生の理解度や研究の進捗状況に応じて注目すべき研究観点が異なることから，研究背景や方法といった特定の観点で生成することは効果的であると考えられる．しかし，観点を反映した要約を生成する研究は十分ではない．

## 2. 関連研究

### 2.1 論文要約

学術論文サーベイのための研究としては，学術論文の典型的な構造である「序章」「関連研究」「提案手法」「評価実験」「結論」のセグメントに着目して要約を生成するものがある(3)．

要約には論文の主な要点である「目的」「提案手法」「先行研究に対する位置づけ」「実験の設定やその結果」「貢献」などが含まれていることが望ましいと考え，論文をセグメントで分割した後，各セグメントから重要文を抽出し統合したものを要約とするものである．

### 2.2 観点を反映した要約生成

要約を生成する際には膨大な情報の中から必要とする情報を含んだ文章を抽出する手法が用いられる．

学術論文に対する要約生成では，研究活動支援を目的に研究背景などの観点を反映した要約生成システム(ViewPoint Refinement in Automatic Summarization)[4]が開発されている．抽出要約を一種の文章分類タスクとみなし，観点ごとに分類した文章そのものを要約文とするものである．

また，必要とする情報を含む文章を抽出するという研究も行われている．文章の重要度を他の文章と意味的な共通集合を持つ度合いで定義し，この重要度をもとに文書の中核となる文章をマイニングする研究がある[5]．マイニングされた文章を入力として抽象要約を行うことで，もとの文章の内容と乖離した文を生成する可能性を減らしつつ，文の流れが流暢で読みやすい要約生成を目指すものである．

本研究では，事前学習言語モデルを用いて観点分類器を構築し，汎用性を考慮して未知の論文に対する観点分類精度の評価を行う．また，観点毎に分類した文章を入力とする抽象要約を生成することで指定した観点を含む要約の生成を試みる．

## 3. 提案手法

観点毎の要約を生成するために，まず論文の本文に含まれる文章を観点で分類する．分類した文章を入力として要約を生成することにより，任意の観点毎の要約を生成する手法を提案する．全体像を図 1 に示す．

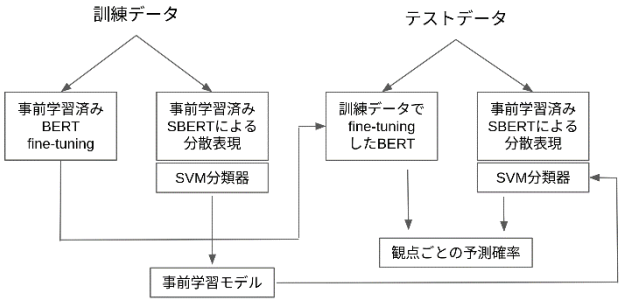


図 1 全体像

### 3.1 前処理

本研究では学術論文の本文を対象に，Apache Tika と呼ばれる Java で開発されたドキュメント分析・抽出ツールを利用し，学会のサイトから取得した論文の本文を抽出した．専門家により文単位で分割した文章に対して「背景」「目的」「方法」「実験・実践」「結果」「知見」「関連研究」からなる観点ラベルと観点ごとの重要度ラベルを付与した．付与されたラベルのうち，観点ラベルは観点分類の教師データとして使用し，重要度ラベルは要約生成における正解要約として使用する．

### 3.2 観点分類

本研究では，図 2 に示すように Transformer[6]を基礎とした BERT[7]と Sentence BERT(SBERT)[8]を用いて観点分類器を構築する．前処理したデータセットを訓練データとテストデータに分割して扱うが，ここで 1 つの論文における文章が訓練データとテストデータの両方に含まれないようにして学習を行う．また，文章を BERT のへ入力するためには文章をトークン化する必要がある．トークン化とは文章を単語レベルに分割した上で BERT に入力できる形式へ変換することであり，形態素解析器の MeCab[9]を用いて単語に分割した後，WordPiece[10]で単語をトークン分割したものを入力として使用する．

BERT に関しては，訓練データにより fine-tuning を行いテストデータで観点毎の予測確率を得る．SBERT に関しては SBERT 自体の fine-tuning は行わず，事前学習モデルにより得られる分散表現を SVM に入力として与え学習することで事前学習モデルを構築する．テストデータに関しては SBERT により得られた分散表現を事前学習モデルに入力することで観点毎の予測確率を得る．

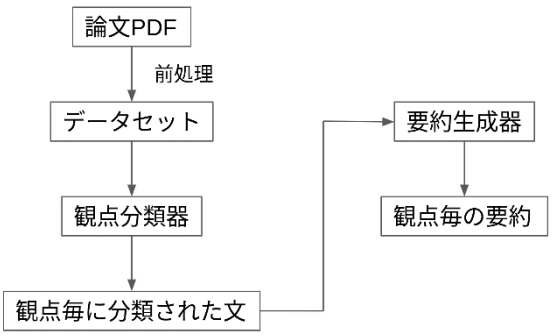


図 2 観点分類モデル

#### 3.2.1 BERT

Generative Pre-trained Transformer(OpenAI GPT)[11]のような fine-tuning アプローチは，一方向性言語モデルにより一般的な言語表現を学習するが，この一方向性により事前学習で学習された表現が制限さ

れている．一方，BERT は双方向から文脈を取り込むという考えに基づき，すべての層で左右両方の文脈を共同で条件づけすることで，ラベル付けされていないテキストから双方向表現を事前学習する Transformer ネットワークである．タスク固有のアーキテクチャを大幅に変更することなく，出力層を追加するだけで質問応答や文のクラス分類，言語推論を含む様々な NLP タスクのモデルを作成可能にする．

3.2.2 Sentence BERT

Sentence BERT は siamese network と triplet network を使用して BERT を fine-tuning することで，コサイン類似度で比較して意味的に重要な文のベクトルを生成するものである[8]．ある文章とそれに類似する文章のペアを学習データとして与えて，これらの文章ペアから生成されるベクトル同士が似たベクトルになるように BERT を fine-tuning することで得られた文のベクトルは，意味の類似性比較やクラスタリング，意味に基づいた情報検索タスクに用いることができる．

### 3.2.3 セクション情報の付与

観点分類精度の試みとして，学術論文が章節項立てで構成されることに着目し，本文を文章単位に分割した後，各文章の先頭に章節項というセクション情報を付与する．概要を図 3 に示す．

「3.4 プログラミング時の活動データ収集」というセクションに「授業は約 100 名の学生を対象に行った」という文章が存在したとき，文章単位に分割した本文「授業は約 100 名の学生を対象に行った」のみを 1 データとしていたが，ここではセクション情報「3.4 プログラミング時の活動データ収集」を本文の先頭へ付与する．このとき，付与するセクション情報のパターンは以下である．

(1)「3.4」等のセクション番号のみ

(2)「プログラミング時の活動データ収集」等のセクション名のみ

(3)「3.4 プログラミング時の活動データ収集」等のセクション情報全体

(1)～(3)それぞれを文章単位に分割したすべての本文に適用し，BERT と SBERT それぞれの分散表現を用いて観点分類を検証する．

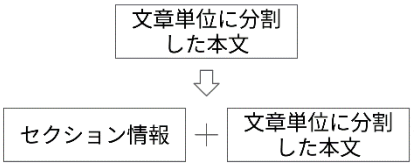


図 3 セクション情報の付与

### 3.3 学習

実応用を考慮すると未知の論文に対して観点分類は行われるものであるため，1 つの論文の文章が訓練とテストデータの両方に含まれていることは適切ではない．そこで，検証においては訓練用とテスト用で論文単位に分けて扱う．BERT の場合，分類タスクのため[CLS]トークンに出力層を接続して fine-tuning 及び推論を行う．SBERT の場合，SBERT を用いるのは入力した文の分散表現を獲得するところまでとし，SBERT 自体の fine-tuning は行わない．つまり，訓練データから SBERT により分散表現を獲得し，得られた分散表現を入力として SVM を学習した後，学習した SVM を用いてテストデータの分類を行う．

### 3.4 要約生成

本研究では，Bidirectional and Auto-Regressive Transformers(BART)[12]を用いて要約生成器を構築する．BART は BERT と同様に大規模データセットで事前学習したものを fine-tuning して目的のタスクに適用する．事前学習を行うにはデータ及び計算リソースが不足していることから，翻訳や要約，言語モデリング等のテキスト生成タスクのためのツールである fairseq の事前学習モデルを使用する．また，文章を BART へ入力するためには文章をトークン化する必要があるが，ここでは形態素解析器の Juman++[13]を用いて単語に分割した後，SentencePiece[14]でトークン化したものを入力として扱う．

3.4.1 BART

BART は BERT を Seq2Seq の形にしたもので，Encoder と Decoder を大量のテキストデータで事前学習したモデルであり，Seq2Seq の仕組みにより機械翻訳や文章の要約に用いられる[12]．また，BERT と同様に fine-tuning することで目的の文章の要約を高精度で生成できる．BERT の双方向 Transformer と

OpenAI GPT の片方向 Transformer を組み合わせた機構であり, BERT を Encoder として入力を処理することに用い, OpenAI GPT の Auto-regressive Decoder で Encoder から出力された情報を受け取って文章を生成するものである.

## 4. 実験

### 4.1 データセット

教育システム情報学会(JSiSE)の 2012 年度の論文 18 編 (JSiSE2012) と 2013 年度の論文 22 編 (JSiSE2013)をデータセットとして用いる. 前処理の手続きに従って論文をそれぞれの文章に分割しラベルを付与した.

### 4.2 観点分類の実験設定

BERT は大規模な文章を用いて汎用的な言語パターンを学習することから 1 から自分で学習することは現実的ではない. そこで, 一般に公開されている東北大学の乾研究室が作成した日本語の学習済みモデルを用いる[15]. また, JSiSE2012 を訓練データ JSiSE2013 をテストデータとして検証する. 学習に際して, BERT に入力する文章を単語に分解するトークナイズにおける最大トークン数は 128 に設定する. また, バッチサイズは 16, 学習率は $1\times10^{-5}$, エポック数は 5 とする.

SBERT は東北大学の乾研究室が作成した日本語の学習済み BERT を fine-tuning して作成された SBERT を用い[16], これにより得られた分散表現を特徴量とし SVM で観点分類を行う. SVM は rbf カーネルを指定し, 誤分類の許容範囲である C を 1.0, カーネル係数を 0.001, その他のパラメータは sklearn のデフォルト値とする.

### 4.3 観点分類結果

先行研究の手法である Word2Vec と PV-DM とこれらの出力結果を組み合わせた Combined-method, BERT および SBERT のセクション情報を付与せず本文のみを使用する場合における観点分類結果を表 1 に示す.

表 1 より, BERT が最も精度が高い結果となった. SBERT においては, 適合率は先行研究を上回っていたが, 再現率と F1 スコアは先行研究を下回る結果であった.

表 1 セクションなしにおける観点分類結果

| 方法 | 適合率 | 再現率 | F1スコア |
|---|---|---|---|
| Word2Vec | 0.36 | 0.30 | 0.30 |
| PV-DM | 0.38 | 0.33 | 0.34 |
| Combined-method | 0.41 | 0.33 | 0.35 |
| BERT(ours) | 0.49 | 0.44 | 0.44 |
| SBERT(ours) | 0.43 | 0.30 | 0.31 |

次に, 論文本文に加えてセクション情報全体を付与した場合における観点分類結果を表 2 に示す.

表 2 セクション情報全体を付与における観点分類結果

| 方法 | 適合率 | 再現率 | F1スコア |
|---|---|---|---|
| Word2Vec | 0.52 | 0.44 | 0.46 |
| PV-DM | 0.45 | 0.39 | 0.40 |
| Combined-method | 0.52 | 0.43 | 0.45 |
| BERT(ours) | 0.66 | 0.69 | 0.67 |
| SBERT(ours) | 0.53 | 0.47 | 0.49 |

表 2 より, 表 1 と同様に BERT が最も精度が高い. また, すべての手法において精度が向上しており, SBERT は先行研究を上回る結果となった. このことから, SBERT はセクション情報といった, 内容に制約のある文字情報が含まれている文章において精度が向上すると考えられる. しかし, SBERT は BART に比べると精度が劣ることから, 観点分類用に SBERT 自体の fine-tuning が必要であると考えられる.

### 4.4 3 行要約データセット

観点毎の要約生成器を構築において, 観点毎の文章を BART に入力として与えて要約を生成するが, BART で要約を生成するためには事前学習モデルを fine-tuning する必要がある. fine-tuning に使用するデータは要約したい文章とその文章の要約文のペアが数千単位で必要になるため, 3 行要約データセットを用いる[17].

3 行要約データセットは，ニュースポータルサイトである livedoor ニュースをスクレイピングして構築する．ニュース記事には本文の他に「ざっくり言うと」という欄が設けられており，記事の内容を説明する文章が 3 つ付与されている．これを記事の要約と見たてて作成したものが 3 行要約データセットである．

### 4.5 重要度の学習

3 行要約データセットで学習したモデルを重要度データセットで fine-tuning することで，論文における各観点においてより重要な文章を含んだ要約を生成することが可能であるか検証する．

### 4.6 要約生成の実験設定

BART も BERT と同様，大規模な文章を用いて汎用的な言語パターンを学習することから 1 から学習することは現実的ではないため，一般に公開されている学習済みモデルを使用する[18]．この学習済みモデルを 3 行要約データセットにより fine-tuning する．また，学習済みの BART はトークン数の上限が 1024，文字数にして 1500 文字程度の制約があるため，入力する記事のうち 1500 文字を超えるものは除外し，学習率は $3\times10^{-5}$，エポック数は 5 として学習する．

### 4.7 要約生成結果

観点が「背景」の本文に対して生成された要約を表 3 に示す．

表 3 「背景」の要約例

| | |
|---|---|
| 本文 | 問題に対して複数の選択肢を提示し，その中から正解を選ばせるといった多肢選択問題は，e ラーニングの教材における典型的な出題形式である．この多肢選択問題における誤選択肢に関しては，(I) 正解を判別できないようにするための誤選択肢（本稿では無意味誤選択肢と呼ぶ），という考え方と，(II) 学習者が間違いうる誤選択肢（本稿では有意味誤選択肢と呼ぶ），という考え方がある．前者の場合，正解の選択肢以外はすべて同等の誤りとなり，解説も正解の解説を用意すればよいことになり，比較的簡単に用意することができる．しかしながら，学習者に対しては正解を教えるだけしかできず，学習者の誤りを活かした学習は行うことができない (1)．後者の場合には，学習者がある誤選択肢を選択したとすると，その誤選択肢が表す誤りを学習者が犯していたということとなり，そのとらえられた誤りに応じたよりきめの細かな支援が可能となる．したがって，学習の効果の面から考えると教材としては有意味誤選択肢で構成された多肢選択問題が望ましいと言える．しかしながら，(A) 学習者が犯しそうな誤りを反映した誤選択肢を用意する必要があり，さらに，(B) 個々の誤選択肢に対する解説の用意が望まれるため，作成に大きなコストがかかることになる．本研究は，(B) の有意味誤選択肢に対する解説文の自動生成を目指したものである．多肢選択問題の自動生成に関する研究はこれまでに盛んに行われている（2）（3）．しかし，これらの手法はいずれも生成される誤選択肢は有意味であるとは言えなかった |
| 生成要約 (3 行要約データセット) | 学習者が間違いうる誤選択肢を紹介している。正解の選択肢以外はすべて同等の誤りとなり，解説も正解の解説を用意すればよいことになる。学習者に対しては正解を教えるだけしかできず，学習者の誤りを活かした学習は行うことができない |
| 生成要約 (3 行要約 + 重要度データセット) | e ラーニングの教材における典型的な出題形式である．この多肢選択問題における誤選択肢に関しては，学習者が間違いうる誤選択肢（本稿では有意味誤選択肢と呼ぶ），という考え方がある．前者の場合，正解の選択肢以外はすべて同等の誤りとなり，解説も正解の解説を用意すればよいことになり，比較的簡単に用意することができる．しかしながら，学習者に |

| | |
|---|---|
| | 対しては正解を教えるだけしかできず，学習者の誤りを活かした学習は行うことができない．後者の場合には，学習者がある誤選択肢を選択したとすると，その誤選択肢が表す誤りを学習者が犯していたということとなり，そのとらえられた誤りに応じたよりきめの細かな支援が可能となる．したがって，学習の効果の面から考えると教材としては有意味誤選択肢で構成された多肢選択問題 |

表3より，3行要約データセットから生成された要約文は，要約対象の本文に対して3行の要約が出力されていることがわかる．これは 3行要約データセットで学習したことによると考えられる．また，要約内容に関しては適切なものが生成できているとは言えない．本文は e ラーニングの教材における出題形式についての内容で，著者は「1. 正解を判定できないようにするための誤選択肢（無意味誤選択肢）」と「2. 学習者が間違いうる誤選択肢（有意味誤選択肢）」の 2つの出題形式を説明し，2の方が学習者にきめ細かい支援が可能となると述べている．しかし，生成された要約を見てみると，1の出題形式の内容に関する説明のみとなっており要約としては不十分であると考えられる．

3 行要約データセットに加えて重要度データセットから生成された要約文は，3 行要約データセットのみで学習させたときよりも生成された要約文が長いことがわかる．これは，重要度データセットにおける重要文の数が 3 行ではないことが原因と考えられる．内容に関しては，著者は本文で「1. 正解を判定できないようにするための誤選択肢（無意味誤選択肢）」と「2. 学習者が間違いうる誤選択肢（有意味誤選択肢）」の 2 つの出題形式を説明していたが，2 の出題形式しか要約に含まれていないが，2 つの出題形式の長所と短所に関しては要約文に含まれており，有意味誤選択肢に着目する内容となった．また，重要文ラベルが付与された太字の文章が部分的にしか含まれておらず，重要文を含むように学習は十分ではないと考えられる．さらに，生成された文章の末尾が途中で途切れる形であることから，正解要約として与えた重要度が付与された文章の形式を工夫する必要があると考えられる．

ここで，重要文を正解要約として「背景」の要約生成例に対する ROUGE スコアを算出する．結果を表 4 に示す．

表 4 より，重要度データセットで fine-tuning することで重要な内容が含まれるようになり ROUGE スコアは向上した．しかし，要約文が長くなったことによるスコアへの影響を検証できていない．

表 4 生成された「背景」の要約文における ROUGE スコア

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| 重要度データセット未使用 | 0.387 | 0.131 | 0.194 |
| 重要度データセット使用 | 0.412 | 0.195 | 0.218 |

## 5. おわりに

研究者や学生が分野理解やサーベイ等の研究活動を行う際，観点を考慮した要約を生成することは効率及び効果の面からしても重要である．本研究では，指定した観点の内容を含んだ要約を生成するため，論文の本文を観点ごとに分類した．そして，分類された文章を入力として要約を生成することで観点を含んだ要約生成を検証した．観点分類においては，BERT を用いることで先行研究の実験設定においては先行研究と同様の精度が得られたことに加え，学習データとテストデータを論文単位で分離するより汎用的な評価を実施することで先行研究手法の精度を上回ることを確認した．また，本文に加えてセクション情報を付与することで精度向上を図り，セクション番号とセクション名を合わせたセクション情報全体を付与することが最も観点分類精度を高めるという知見を得た．

論文要約においては，抽象要約手法を用いたことで各観点においてどのような要約が得られるのか知見を得た．観点ごとの特性を考慮して要約生成手法を改良していくことで，研究者や学生の分野理解やサーベイの効率化につながることが期待される．

## 参 考 文 献


(1) “arxiv,” https://arxiv.org/.

(2) Danish Contractor, Yufan Guo, and Anna Korho-nen. Using argumentative zones for extractive sum-marization of scientific articles. COLING ’12, pp.663–678, 2012.

(3) SHIN Wonha and 白井 清昭, “セグメント構造を考慮した学術論文の包括的要約の自動生成の提案,” 言語処理学会 第 23 回年次大会 発表論文集, 2017

(4) リ ケンコン, 太田光一, and 長谷川忍, “観点を反映した深層強化学習による学術論文の自動要約生成,” in 信学技報, ser.ET2020-19, vol.120, no.167, 石川, 9 月 2020, pp.53–56, 2020 年 9 月 10 日

(5) 重松 祐匡 and 尼崎 太樹, “文章間情報を用いた抽出的要約生成器 est,” 人工知能学会全国大会論文集, vol. JSAI2022, pp.1P5GS605–1P5GS605, 2022

(6) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, et al., “Attention is all you need,” CoRR, vol. Abs/1706.03762, 2017

(7) J. Devlin, M. Chang, K. Lee, et al., “BERT: pre-training of deep bidirectional transformers for language understanding,” CoRR, vol. Abs/1810.04805, 2018

(8) N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in EMNLP/IJCNLP (1), K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019

(9) T. KUDO, “Mecab: Yet another part-of-speech and morphological analyzer,” http://mecab.sourceforge.net/, 2005

(10) Google. (2018) The wordpiece algorithm in open source bert. [Online]. Available: https://github.com/google-research/bert/blob/master/tokenization.py

(11) A. Radford, K. Narasimhan, T. Salimans, et al., “Improving language understanding by generative pre-training,”, 2018

(12) M. Lewis, Y. Liu, N. Goyal, et al., “BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” CoRR, vol. Abs/1910.13461, 2019

(13) A. Tolmachev, D. Kawahara, and S. Kurohashi, “Juman++: A morphological analysis toolkit for scriptio continua,” in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018

(14) R. Sennrich, B. Haddow, and A. Birch, “Neural machine translation of rare words with subword units,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, Aug. 2016

(15) M. Suzuki, “Pretrained japanese bert models,” https://github.com/cl-tohoku/bert-japanese.

(16) I. Sonobe, “Japanese sentence-bert model,” 2021, https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2.

(17) 小平 知範 and 小町 守, “Tl;dr 3 行要約に着目したニューラル文書要約,” 電子情報通信学会 技術研究報告 = IEICE technical report : 信学技報, vol. 117, no. 212, pp. 193–198, 09, 2017

(18) M. Ott, S. Edunov, A. Baevski, et al.,“fairseq: A fast, extensible toolkit for sequence modeling,” in Proceedings of NAACL-HLT2019: Demonstrations, 2019