1C-01

セクション情報を用いた日本語長文文書の抽象型自動要約

沢畑英之 *
電気通信大学 *
情報学専攻

大久保誠也 †
静岡県立大学 †
経営情報学部

若月光夫 ‡
電気通信大学 ‡
情報学専攻

西野哲朗 §
電気通信大学 §
情報学専攻

1 はじめに

近年では自動要約に対する需要は高まってきており, ニュース記事などの短文の文書に対する自動要約の研究がされてきており, 日本語でも同様の研究が行われている. また, 最近では長文の文書に対する自動要約研究もされてきている.

長文の文書に対する自動要約研究では, 主に論文要約が対象にされており, 様々な成果が挙げられている. しかし, これらの研究は主に英語で書かれた論文を対象に行われており, 日本語で書かれた論文を対象として行われた自動要約研究の数は少ない. その主な理由として 2 点挙げられる. まず, 入出力に必要な文字数が増えているため, 計算量が増加し, 学習が上手くできないという点. もう一つは, そもそも学習に使用できる十分な数を持つコーパスがないという点である.

2 研究目的

そこで本研究では, 上記 2 つの問題点を解決するために, DANCER[1] と呼ばれる論文自動要約の学習及び, 生成手法を基に, 日本語で書かれた論文を対象とする自動要約に用いるセクションを自動で選択し, 抽象型自動要約を行う手法を提案する. そして, 提案手法が, 日本語論文の抽象型自動要約手法において, 全文入力から要約を生成する場合よりも有効であることを示す.

Automatic abstractive summarization for long Japanese documents using section information

*Hideyuki Sawahata, Department of Informatics, The University of Electro-Communications
†Seiya Okubo, School of Management & Information, University of Shizuoka
‡Mitsuo Wakatsuki, Department of Informatics, The University of Electro-Communications
§Tetsuro Nishino, Department of Informatics, The University of Electro-Communications

3 関連研究

3.1 論文に対する自動要約

論文を対象に行った自動要約手法には, 抽出型要約手法では, 論文の要約とセクションの関係に着目し, 序論と結論から要約を生成する手法 [2] がある. この手法ではトピックモデルを用いてそのまま抽出型要約を行うよりも, 序論と結論のセクションから抽出型要約を行うことによって改善できることを示した. 抽象型要約においては論文の構造をニューラルネットワークに当てはめて, 学習及び生成を行うことが多い. 例えば Discores-aware モデル [3] は, 論文内の各セクションの特徴量から論文全体の特徴量を生成し, 要約を生成するという構造になっている. 次に述べる DANCER は自動要約モデルに対する, 論文内のセクションを利用した学習方法, 及び要約生成手法である.

3.2 DANCER

DANCER は, 論文の構造を利用した自動要約モデルの学習・生成手法である. DANCER では, モデルの学習の前に, 要約内の各文を各セクションに割り当てる. 学習の際には, セクションごとに入力し, 割り当てられた要約を生成するように学習させる. 要約生成の際にはセクションごとに要約を生成し, 最終的にそれらすべてを結合して, 論文本体の要約とする. DANCER では要約の生成に用いるセクションを事前に限定しており, 序論と結論, 実験, 提案手法の 4 種類のセクションに限定している. 本研究では事前に限定はせず, 要約生成の際に使用するセクションを自動で選択し, 選択されたものから要約生成を行う手法を提案する.

4 提案手法

本研究では, 先述したように要約に使用するセクションを自動選択し, それらを用いて要約を生成する手法を

提案する. この手法では, 論文内のセクション名を入力として, 要約に使用するセクション番号を出力する, セクション選択モデルを構築し, 選択されたセクションから要約を生成する. セクション選択モデル構築のための学習データは, DANCER における要約文のセクションへの割り当てにおいて, 割り当てられたセクションの番号を正解とするデータを作成・使用する. しかし, 日本語論文コーパスのみでは, コーパスに含まれる論文の数が少ない. そのため, 日本語論文のみで学習を行っても十分な性能を発揮しない可能性がある. そこで, 英語コーパスを用いた few-shot 学習によるセクション選択モデルを構築する. 要約生成モデルの学習方法と要約生成の手法は DANCER と同様の手法を取る.

5 実験

5.1 実験内容

提案手法を用いることで全文から要約を生成する場合よりも改善が可能かどうかを検証した. 本実験にて要約の対象として自然言語処理学会論文誌コーパス [4] を使用した. 本コーパスより, 342 件を学習データとし, 86 件を評価データとして使用した. 正解となる要約は論文内に記載されている要旨を使用した. 評価手法として bert-score[5] を使用した.

5.2 実験結果

表 1 に全文を入力した場合と自動選択したセクションから生成した場合の結果を示す. 全文入力で生成した場合と比較して, 提案手法の方が有効であることを確認できた.

表 1: 全文入力と提案手法の結果

	全文入力	提案手法
Bert-score	0.623	0.729

5.3 考察

実験の結果より, 全文入力よりも, 提案手法に則って要約生成する手法が有効であることが確認できた. これは, セクション選択を行うことで, 要約に必要な文章のみに限定することができたためだと思われる. しかし, セクション選択において 1 つのセクションしか選択されない例も確認できたため, 選択すべきセクションの最低数を保証できるような学習方法について模索していく必要がある.

6 結論

本研究では日本語の長文文書の抽象型自動要約として, 論文を対象とした, 要約対象のセクションを自動で限定する抽象型自動要約の手法を提案した. この提案手法は全文を入力として要約を生成するよりも, bert-score 上で上回る結果となった. このことから, 提案手法は論文要約において有効な手法であるといえる. また, 提案手法は論文と類似した構造を持つ文書において有効である可能性も考えられるため, 検証を行う必要がある.

参考文献

[1] Alexios Gidiotis and Grigorios Tsoumakas. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 3029–3040, 2020.

[2] Hideyuki Sawahata and Tetsuro Nishino. Automatic extractive summarization for japanese documents by lda. In Tokuro Matsuo, editor, *Proceedings of 11th International Congress on Advanced Applied Informatics*, Vol. 81 of *EPiC Series in Computing*, pp. 41–52. EasyChair, 2022.

[3] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[4] 言語処理学会. 言語処理学会論文誌 latex コーパス. https://www.anlp.jp/resource/journal_latex/index.html, 2020.

[5] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.