The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

ニューラルネットワークを用いたニュース記事の抽出的三行要約

Extractive Three-line Summarization of News Articles using Neural Network

小川 恒平[*1] Kouhei Ogawa
永塚 光一[*1] Koichi Nagatsuka
渥美 雅保[*1] Masayasu Atsumi

[*1] 創価大学大学院工学研究科情報システム工学専攻
Dept. of Information Systems Science, Graduate school of Engineering, Soka University

The purpose of text summarization is to enable readers to understand the point of documents without reading overall of them. With the spread of smartphones, opportunities to view news articles within the limited screen area are increasing, so that there are needs easy to read summaries within the limited screen area. In this paper, we propose a neural network model that extracts three-line summaries from news articles. As a dataset, we use articles and three-line summaries of Livedoor NEWS on the Internet. In an experiment, we show that the proposed model performs better than a conventional method.

1. はじめに

要約の目的は,読み手が文書全体を読むことなく,その文書の内容をある程度把握できるようにすることである.今までのニューラルネットワークを用いた要約の研究では,機械翻訳で性能を向上させた,アテンションを取り入れたエンコーダデコーダモデルによる研究が盛んに行われてきた. Rushらによる単文要約[Rush 15]や,Paulusによる複文要約[Paulus 17]の研究などにおいて,大きな性能の向上が報告されている.また,スマートフォン等の普及により,限られた画面の範囲でニュース記事などのテキストを閲覧する機会が増えてきている.そのため,スマートフォンでニュース記事を読む場合は,小さな画面でも見やすい長さの要約文が理想的である.本研究では,ネット上のニュースサイトであるLivedoor NEWSの記事とそれに対応した三行要約のペアをデータセットとして使い,ニュース記事から,スマートフォンでも見やすい三行要約を自動で出力するニューラルネットワークモデルを提案し,実験によりその性能を評価する.

2. 三行要約

図1は,本研究で使用するLivedoor NEWSというサイトの人手により作られた三行要約の例である.タイトルに加えて三行要約が付いていることで,タイトルだけでは元記事の情報を把握しきれないというデメリットをカバーしている.三行要約は,スマートデバイスでも見やすい要約で,私たちが小さな画面で読みたい記事を探す際のサポートになる.

図1. Livedoor NEWS

連絡先:小川恒平,創価大学情報システム工学科,〒192-8577 東京都八王子市丹木町1-236, Tel:042-691-2211

3. 関連研究

3.1 LEAD法

抽出型要約におけるベースラインとして用いられる手法としてLEAD法がある.LEAD法は,入力文書中の文を上から任意の数取ってくるものである. これは, 文書中の重要な内容が,文書の先頭にくるという仮定のもと用いられている. 単純でありながら精度の高い手法で,特にニュース記事に有効である.本研究では,このLEAD法と提案手法の精度を比較している.

3.2 要約構造分類モデル

三行要約の先行研究として,小平が提案した要約構造分類モデル[小平 18]がある.この要約モデルでは,まず与えられたテキストに対して“並列”または“直列”タイプの三行要約構造が存在すると仮定する.“並列”タイプの三行要約構造では, 1文目の要約をトピックの中心として,2,3文目を1文目から派生するサブトピックとみなして三行要約を生成する.これに対し,“直列”タイプの三行要約構造では, 1文目のトピックを起点として, 2,3文目のサブトピックを時系列的に並べ,三行要約生成する. 各タイプに対してモデルをそれぞれ個別に訓練することで,各要約構造タイプに特化した要約生成モデルを学習する.しかし小平の研究では,“並列”と“直列”タイプに分類分けを行ったことで,データセットの量に偏りが起きてしまうという問題がある.実際,小平の研究では直列タイプが少ないと述べている.本研究では,“並列”,“直列”タイプのように分類分けは行わずに,データセットは一つにまとめている.

4. 提案モデル

図2に提案モデルを示す.モデルはHsuらのモデル[Hsu 17]をもとにしている.まず,元記事を一つ一つの文に分ける.そして各文の単語列を正方向と逆方向にBiGRUでエンコードしたのち,最初と最後の隠れ状態を足し合わせる.次にそれら隠れ状態の系列をBiGRUによりエンコードしてsoftmax関数により,各文の重要度分布を出す.最後に,重要度で各文を並べたのち,重要度が高い3文を三行要約として抽出する.訓練では,重要文の教師ラベルを,人手による

各要約文に対してそれと記事中の各文間での ROUGE-2(recall)スコアが最も高かった文に付与する.

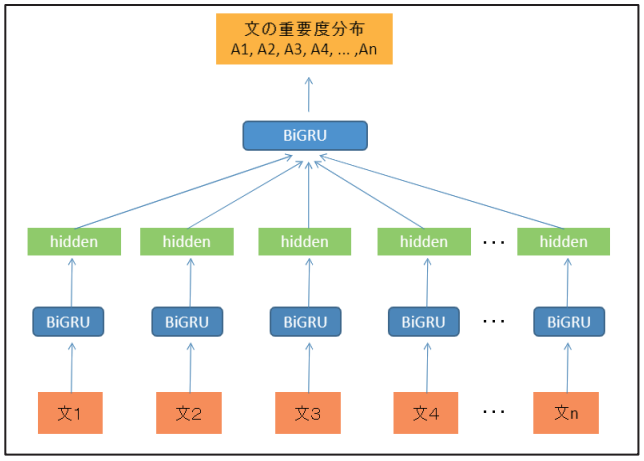

図 2.提案モデル

5. データセット

データセットは, 小平の研究[小平 18]において公開されていた Livedoor NEWS の記事と要約の URL を使い, スクレイピングをして収集したものを用いる. また, 三行要約の 1 文あたりの文長を約 30 字程度と考えたとき, 要約の合計文字数は 100 字程度になるため, 要約する元記事の文字数は要約文の倍である 200 字以上とし, 最大長は 2000 字までの範囲に限定した. データは文字化けしてしまっているものや, 記事の一部しかスクレイピングできていないものなど, データとして使えないものは除去を行い MeCab の形態素解析で分かち書きをしている. 記事とそれに対応する三行要約を 1 ペアと数え, 実験ではトレーニングデータ用に 8000 ペア, テストデータ用に 1000 ペア, バリデーションデータ用に 1000 ペアを使用した.

6. 実験

6.1 実験概要

表 1 にデータのペア数, 合計文数, 平均文数, Vocabulary 数, 平均文長を示す. 実験では評価指標として, 要約タスクに用いられる ROUGE を採用し, スコアを求めた.

表 1. データセット

	Train	Test	Validation
ペア数	8000	1000	1000
合計文数	165762	21160	19680
平均文数	20.72	21.16	19.68
Vocabulary 数	74782	29255	24592
平均文長	27.33	27.23	27.61

6.2 実験結果

今回行った実験に対する ROUGE スコアを下の表 2 に示す. LEAD 法と提案手法の二つの手法とも実験を数回繰り返し, ROUGE の平均値を取っている. 表 3 は今回の提案手法で出力された三行要の一例である.

表 2. ROUGE スコア

	ROUGE-1	ROUGE-2	ROUGE-L
提案手法	0.48	0.17	0.06
LEAD 法	0.45	0.105	0.07

実験の結果, ROUGE-1 と ROUGE-2 では, LEAD 法よりも提案手法の方が高いスコアを得ることができた. ニュース記事の先頭部分に重要な内容は固まりやすいが, 提案手法のように先頭以外からも重要だと思われる文を抽出し, 三行要約を構成した方が, 記事の先頭を抽出してくる LEAD 法よりも提案手法の方が高いスコアを得ることができた.

7. むすび

本研究では, ニュース記事から三行要約を抽出するニューラルネットワークモデルを提案した. そして Livedoor NEWS のデータセットを使い, 記事から三行要約を抽出した. LEAD 法の抽出的要約よりも ROUGE-1, ROUGE-2 に関して良いスコアを獲得することができた. しかし, データセットには少量であるが, まだノイズが残ってしまっていた. 今後もデータセットの整形やモデルの改良などを加えていき, さらなる性能の向上を試みる.

表 3. 提案手法の出力結果の一例

正解の三行要約
東日本 大震災 から 5 年 を 迎え た 11 日 , Twitter で は 「 黙祷 なう 」 が 散見 され た これ に 「 心 の 中 で やっ とけ よ 」 と いった 否定 的 の 声 が 多く 上がっ て いる 一方 で , どういう 形 で あれ 「 思い出す こと 」 が 大切 と 理解 を 示す 声 も ある
[元記事の文番号] 重要度が高いとされた 3 文（三行要約）
[0] 死者 ・ 行方 不明 者 あわせ て 1 万 8455 人 という 甚大 な 被害 を もたらし た 東日本 大震災 から 3 月 11 日 で 5 年 目 を 迎え た . [2] そんな 中 , 今年 も 物議 を 醸し て いる の が , Twitter に 書き込ま れる 「 追悼 なう 」 という 言葉 だ . [5] 」 「 心 の 中 で やっ とけ よ 」 と いった 否定 的 の 声 が 多く 上がっ て いる .

参考文献

[Rush 15]Alexander M. Rush, Sumit Chopra, Jason Weston: A Neural Attention Model for Abstractive

Sentence Summarization, *arXiv preprint arXiv:1509.00685*, 2015.

[Paulus 17]Romain Paulus, Caiming Xiong, Richard Socher: A Deep Reinforced Model for Abstractive Summarization, *arXiv preprint arXiv:1705.04304*, 2017.

[Hsu 17] Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, jing Tang, Min Sun: A Unified Model for Extractive and Abstractive Summarization using Inconsistency Loss, *arXiv preprint arXiv:1708.00391,* 2017.

[小平 18]. 文書構造に着目したニューラル文書要約. 首都大学東京大学院システムデザイン研究科修士論文.