Attention モデルを用いた生成型要約

1 はじめに

近年，インターネット上にはニュース記事やブログ記事などの文書データが溢れ，多くの人々が情報として利用している．しかしそれらの中には冗長な文章が存在し，簡潔に要約された文章が求められている．また簡潔にまとめられた要約は元文書を読む際の手がかりともなる．

これらの要求を満たすために様々な文書要約手法が提案されてきた．自然言語処理を用いた文書要約には生成型要約と抽出型要約の 2 種類の方法がある．抽出型要約は入力文書に存在する入力文書の代表文や重要文を抽出することで要約し，生成型要約は要約元となる文から文章を再構築し，新しい文を生成することで要約を行う．生成型要約は助詞の接続の問題や未知語の問題などがありコンピュータで行うには難しく，既存研究の多くは抽出型要約に属している．特に日本語は表記体系がひらがな，カタカナ，漢字, アルファベットなど多く存在し，助詞の接続が複雑なため，生成型要約がより困難とされている．しかし人間の作る要約文のような自然な要約文には生成型要約が必要不可欠であり，そのため生成型要約に関するニューラルネットワークを用いた研究が盛んに行われている．実際に機械翻訳手法として考案された Encoder-Decoder モデルや Attention モデルを生成型要約に用いる手法が提案されており，短文要約やタイトル生成といったタスクにおいて従来の手法を上回る性能が報告されている．

そこで本研究ではこの Attention モデルと LSTM を用いた言語モデルを組み合わせて用いることでニュース記事を 3 文から 1 文に要約することを目標とする．

2 要素技術

2.1 RNNLM

RNNLM (Recurrent Neural Network Language Model) とは，RNN を利用した言語モデルのことである．RNN は系列を扱うことができるニューラルネットワークで，隠れ層に帰還路を持つことにより，任意の個数の入力が行える．文を単語の系列として扱うことができ，単語を順に入力することで語順を考慮して文を扱うことができる．RNN は原理的には隠れ層は全ての入力を考慮することが可能であるが，実際には長期的な記憶は困難であるため，長期的な記憶が可能な LSTM(Long Short Term Memory) を用いることで性能が向上する．

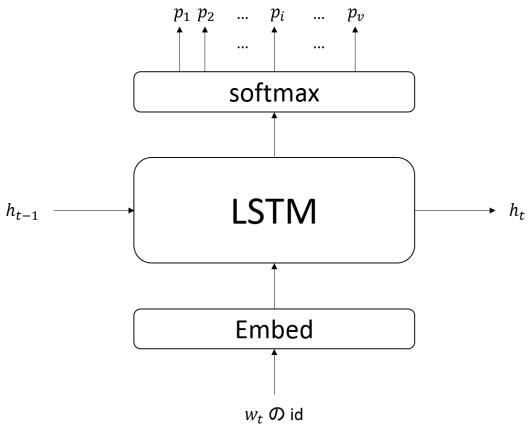

図 1: 時刻 t における RNNLM

また言語モデルとは文 s が現れる確率 $P(s)$ を与える確率モデルのことである．文 s が $w_1 w_2 w_3 \cdots w_N$ という N 個の単語の列であるとき，$P(s)$ は以下のように分解できる．

$$\begin{aligned} P(s) &= P(w_1)P(w_2|w_1)\cdots P(w_N|w_1 w_2 w_3 \cdots w_{N-1}) \\ &= P(w_1)\prod_{t=2}^{N} P(w_t|w_1 w_2 w_3 \cdots w_{t-1}) \end{aligned} \quad (1)$$

よって式 (1) より，単語列が与えられたときに次に現れる単語の確率をそれぞれ算出できるので，次の単語の予測ができる．次に現れる単語の確率の算出は RNN や LSTM を用いたニューラルネットワークモデルで次に出る単語の確率をあげるように学習させていくことで実現される．LSTM を用いた時刻 t における RNNLM は図 1 のようになっている．

2.2 Encoder-Decoder モデル

Cho らは 2014 年，2 つの RNN を用いて言語 A の文から言語 B の文へ機械翻訳を行う Encoder-Decoder モデルを提案した [1]. Encoder では RNN に言語 A

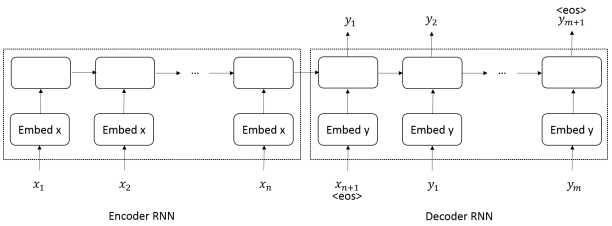

図 2: Encoder-Decoder モデル

の文に含まれる単語 x_i を順に入力していき，n 語まで入力したときの隠れ層の出力を言語 A の文をエンコードしたベクトルとして出力する．Decoder は RNN を用いた言語モデル RNNLM に潜在状態として Encoder の出力を入力したモデルで，言語 B の単語 y_i の確率を出力する．図 2 に Encoder-Decoder モデルを示す．

2.3 Attention モデル

Encoder-Decoder モデルでは可変長の文を固定長のベクトルにエンコードするため，長い入力文になるほど隠れ層のノード数が不足し，学習が難しくなる問題がある．そこで Bahdanau らは 2015 年，Eecoder 側が入力文の各単語の荷重を決定してエンコードするべき場所を制御する Attention モデルを提案した [2]. 図 3 に Attention を導入した Encoder-Decoder モデルを示す．図 3 中の α_t は入力 y_t における入力文の各単語 x_i の荷重を表し，c_t は y_t における Eecoder の出力結果になっている．α_t は x_i からの出力 $\overline{h_i}$ と y_t からの出力 h_t の類似度を正規化することによって得られる．それぞれのモデル式は以下の通りである．

$$\alpha_t(i) = \frac{\exp((\overline{h_i}, h_t))}{\sum_{j=1}^{n} \exp((\overline{h_j}, h_t))} \quad (2)$$

$$c_t = \sum_{i=1}^{n} \alpha_t(i)\overline{h_i} \quad (3)$$

また式 (3) からの c_t と y_t から出力された h_t を連結させたベクトル $[c_t; h_t]$ に重み W_c をつけて活性化関数 tanh を被せて出力された $\tilde{h_t}$ を Decoder の出力としている．

$$\tilde{h_t} = tanh(W_c[c_t; h_t]) \quad (4)$$

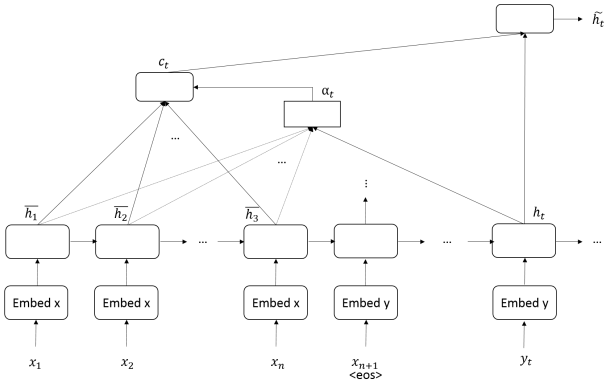

図 3: Attention モデル

2015 年に Rush らによって，Attention モデルを用いた生成型要約手法が提案されており，従来の手法を上回る性能が報告されている [3].

3 実験

3.1 実験 1

3.1.1 実験方法

Attention モデルと LSTM を用いたニュース記事の生成型要約をした．

初めにデータセットについての説明をする．livedoor は国内や海外のニュース記事やスポーツに関する記事を掲載しているサイトであり，その中には記事に対する 3 行の要約文を掲載している場合もある．そこで livedoor に 10 月 1 日から 12 月 10 日に掲載された国内と海外のニュース記事のうち 3 文の要約文を掲載しているものを収集し，学習用データとして用いた．収集した 1442 件のニュース記事からランダムに選んだ 1000 件を訓練データ，100 件をテストデータとした．その中から入力文 X としてニュース記事の最初の 3 文を用い，それに対する出力 Y として 3 文の要約文の最初の 1 文を用いた．総出現語彙数は 11106 個であった．これらを Attention モデルで学習させた．なお形態素解析には Mecab を用い，数字に関してはすべて Num トークンに置換した．

次にモデルの構成と学習方法について説明する．Encoder には図 3 における RNN を利用しない Attention モデルを用い，Decoder には図 1 の RNNLM を用いた．モデルを図 4 に示す．モデルは chainer フレームワークで実装した．モデルのパラメータは以下の通りである．

表 1: モデルのパラメータ

エンベディングサイズ	200
LSTM の次元数	200
optimizer	Adam
epoch	200

本研究では入力単語を 1-hot ベクトルで与え，これを実数値ベクトルにエンベディングしたうえでエンコードし，このベクトルとそれまでの要約文 Y_c を Decoder に入力として与えることで次に出現する単語の確率 $P(y_{t+1}|X, Y_c)$ を出力した．学習では出力される正解となる単語の確率値の負の対数尤度を最小化するように学習を行った．K 個の参照要約の対 $D = \{(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \cdots, (X^{(K)}, Y^{(K)})\}$ があるとき負の対数尤度 (NLL) は以下のようになる．

$$\begin{aligned} NLL &= -\sum_{k=1}^{K} \log P(Y^{(k)}|X^{(k)}) \\ &= -\sum_{k=1}^{K} \sum_{l=1}^{|Y^{(k)}|-1} \log P({y^{(k)}}_l|X^{(k)}, {Y^{(k)}}_c) \end{aligned} \quad (5)$$

最後に文章生成方法と評価方法について説明する．出力される要約文に対する負の対数尤度を最小化するように学習をしたため，要約文として最も生成される確率の高い文章を選ぶ必要がある．これは語彙数 V の状態をもった最適経路問題を解くという問題に帰着され，これは NP 問題であり計算時間内に解くことが不可能である．そのため，枝刈りを行いながら探索を行う必要がある. そこで本研究では探索アルゴリズムとしてビームサーチを用いた．ビームサイズは 3 に設定した．また評価方法としてはシステム要約と参照要約で n-gram 単位でどれほど一致したかを表す RUOGE-N という指標を用い，ROUGE-1，ROUGE-2 を用いた．

3.1.2 実験結果

初めにそれぞれの 10 epoch ごとに対する 100 件の訓練データ，テストデータの ROUGE-1,2 値の平均を図 5，図 6 に示す．また実際に出力された 200 epoch 目の要約文の 1 例を以下に示す.

入力文 1 自民党の応援演説の顔として全国を飛び回っている小泉進次郎が、Num 日 (Num 年 Num 月) に地元 (神奈川 Num 区) 入りし、「きょうは全国から帰ってきた。Num 日だけ地元・横須賀で活動ができる日になりました」と演説した 。前回 Num 年の衆院選でも地元にほとんど帰れなかったが 、約 Num 票を獲得し 、全国最多得票で当選だった。

出力文 1 小泉進次郎氏が Num 日から Num 日連続で 、暴行の容疑者についてゲンダイが報じた。

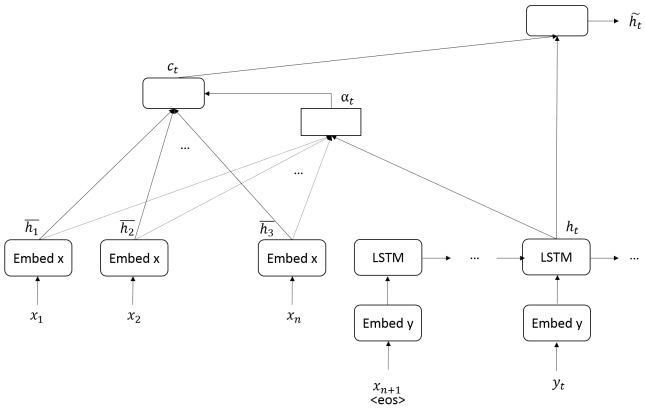

図 4: 生成型要約モデル

図 5，図 6 の ROUGE 値から学習が進んでいることは確認できるが，テストデータに対しては精度が低いことが見てとれる．これは訓練データのデータ数が少なかったために過学習が起きたことが原因と考えられる．また出力文において主語はある程度とれているがそれ以降の文で内容と異なる文を生成していることからも，過学習が起きていると推測できる．これは言語モデルに LSTM を用いたことが原因と考えられる．そのほかの原因としては未知語に対して対応できなかったことが考えられる．また今回，入力分として 3 行しかいれていなかったので，参照要約文との対応がとれていなかった可能性もある．

3.2 追加実験

先の実験で訓練データが少なくとも，主語に関してはある程度とれることが分かった．そこで主語に関する部分だけ Attention モデルから取り，残りの部分は入力文からそのまま抜き取ってくるように生成した．200 epoch 目に対する要約文の例を以下に示す．また 100 件の訓練データ，テストデータに対する ROUGE-1，2 値を表 2 に示す.

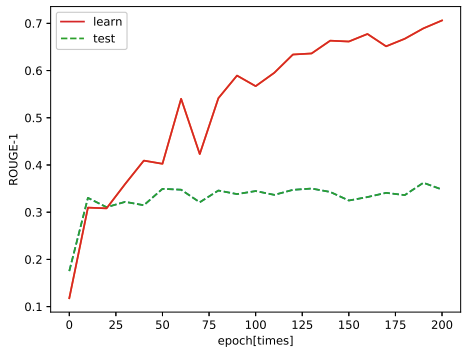

図 5: ROUGE-1 score

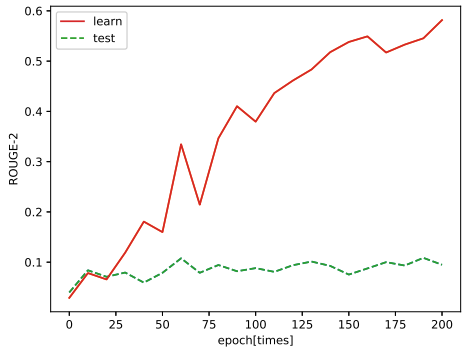

図 6: ROUGE-2 score

入力文 1 同上

出力文 1 小泉進次郎が Num 日 (Num 年 Num 月) に地元 (神奈川 Num 区) 入りし、「きょうは全国から帰ってきた。Num 日だけ地元・横須賀で活動ができる日になりました」と演説した。

入力文 2 東京都江戸川区で Num 年 11 月、都立高校 Num 年岩瀬加奈さん (当時 Num 歳) を殺害して現金を奪うなどしたとして強盗殺人罪などに問われ、Num 審・東京地裁の裁判員裁判で求刑通り無期懲役の判決を言い渡された元コンビニ店従業員・青木正裕被告 (Num) の控訴審第 Num 回公判が Num 日、東京高裁 (栃木力裁判長) であった。弁護側は刑が重いとして有期刑の適用を主張。検察側は控訴棄却を求め、即日結審した。

出力文 2 東京の裁判員裁判で求刑通り無期懲役の判決を言い渡された元コンビニ店従業員・青木正裕被告 (Num) の控訴審第 Num 回公判が Num 日、東京高裁 (栃木力裁判長) であった。

表 2: ROUGE score

	ROUGE-1	ROUGE-2
訓練データ	0.462	0.251
テストデータ	0.437	0.224

表 2 から見て取れるようにシステム要約の ROUGE 値は向上していたが，入力文の一部をそのまま抽出しているために冗長な文になってしまう場合もあった．

4 まとめと今後の課題

Attention モデルを用いて生成型要約をした結果，過学習のためにテストデータに対してよい精度は得られなかった．しかし，ある程度入力文をそのまま抽出することで ROUGE 値が上昇することが確認できた．今後の課題として過学習を避けるために訓練データを増やすことやパラメータサーチ，LSTM を用いた言語モデルが適切かどうかの検証が挙げられる．また入力文数を増やして複数文から複数文の生成をすることなどが挙げられる．

参考文献

[1] Cho,K.,Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.:“Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation ”, In Proceedings of Empirical Methods in Natural Language Processing 2014, pp.1724-1734, (2014).

[2] Bahdanau, D., Cho, K. and Bengio, Y.: “Neural Machine Translation by Jointly Learning to Align and Translate”, In Proceedings of the International Conference on Learning Representation 2015, (2015).

[3] M. Rush, A., Chopra, S. and Weston, J.: “A Neural Attention Model for Abstractive Sentence Summarization”, In Proceedings of Empirical Methods in Natural Language Processing, pp. 379-389, (2015).