

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、多くの人々が情報源として利用している。しかしそれらの中には冗長な文章も存在するため、簡潔に要約された文章への要求が高まっている。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り、情報検索の負荷を軽減できる点で有用である。

上記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する。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は難しい処理であり、そのため既存研究の多くは抽出型要約に属しているのが現状である。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を得るため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近年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 Attention モデルを生成型要約に用いる Attention based Summarization (ABS) という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。また要約は複数の文を参照して行われるものであるという考えのもと、ABSを拡張し、複数文をエンコードし、統合することで複数文のエンコードをするマルチ Attention モデルも提案されている。

本研究では ABS モデルと マルチ Attention モデルを用いてニュース記事を要約し、それぞれのモデルの性能の比較をすることを目的とする。

2 要素技術

2.1 Feed Forward Neural Network による言語モデル

FFNN による言語モデルは、Bengio らによって考案された言語モデルである。このモデルはある時刻 t 以前の単語を n -gram 単位で単語列として入力し、次の単語の予測をするニューラルネットワークモデルである [1]。FFNN による言語モデルの式を以下に示す。

$$\tilde{y}_c = \text{cat}(Ey_{t-c+1}, \dots, Ey_t) \quad (1)$$

$$h = \tanh(U\tilde{y}_c) \quad (2)$$

y_{t-c+1}, \dots, y_t は生成された時刻 t 以前の c 個の単語を表す。 E は生成された単語 y_{t-c+1}, \dots, y_t を実数値ベクトルに変換するエンベディング行列で U は重みパラメータである。 \tilde{y}_c は y_{t-c+1}, \dots, y_t をエンベディングしたベクトルを連結したベクトルを表し、 h は FFNN の隠れ層のベクトルを表す。 cat 関数はベクトルを連結する関数を表す。

2.2 Attention モデル

機械翻訳などのタスクで考案された Encoder-Decoder モデル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、エンコーダ側で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]。

Attention モデルでは入力文の各単語 x_i に対する荷重 α_t を計算することで、エンコーダで出力される中間表現 c_t を得る。 α_t はエンコーダで出力されるベクトル h_i とデコーダから出力されるベクトル h_t の内積を正規化することによって得られる。

$$\alpha_t(i) = \frac{\exp((\bar{h}_i, h_t))}{\sum_{j=1}^n \exp((\bar{h}_j, h_t))} \quad (3)$$

$$c_t = \sum_{i=1}^n \alpha_t(i) \bar{h}_i \quad (4)$$

2.3 Attention based Summarization モデル

Rush らは 2015 年、機械翻訳モデルである Attention モデルをもとに、文から短い文への生成型要約を行う Attention Based Summarization (ABS) モデルを提案した [3]。このモデルは Attention モデルの持つ単語に荷重をかけながらエンコードするという特徴を利用することで、要約に必要な単語に荷重をかけながら、生成的に文を要約することを可能とした。Rush らはこのモデルでニュース記事からタイトル生成をする実験により、従来の手法より高い精度を得た。図 1 に ABS モデルを示す。

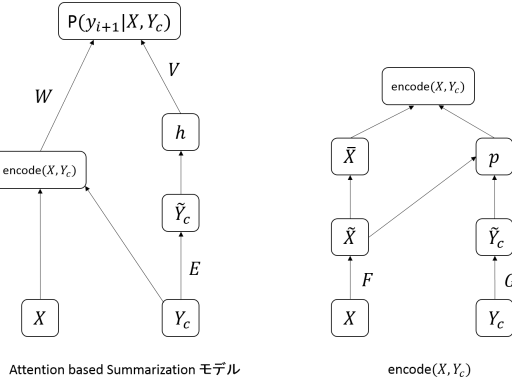


図 1: ABS モデル (文献 [3] をもとに作成)

ABS モデルはエンコーダ部分とデコーダ部分に分けられる。エンコーダ部分は 2.2 節で述べた Attention モデルが使用され、デコーダ部分では 2.1 節で述べた FFNN による言語モデルが使用されている。

図 1 において X は元文を表し、 Y_c は生成した c 個の単語を表す。 E , F , G は元文、生成文の単語を 1-hot ベクトルから実数値ベクトルに変換するエンベディング行列であり W , V は重みパラメータである。また p は単語の荷重、 \bar{X} は \tilde{X} をスムージングした行列である。

それぞれのモデル式は以下の通りである。 q はスムージングウィンドウサイズの大きさを、 cat 関数はベクトルを連結する関数を表す。

$$\text{encode}(X, Y_c) = p^T \bar{X} \quad (5)$$

$$p \propto \exp(\tilde{X} P \tilde{Y}_c) \quad (6)$$

$$\bar{x}_i = \frac{\sum_{j=i-q}^{i+q} \tilde{x}_j}{q} \in \bar{X} \quad (7)$$

$$\tilde{x}_i = F x_i \in \tilde{X} \quad (8)$$

$$\tilde{Y}_c = \text{cat}(G y_{i-c+1}, \dots, G y_i) \quad (9)$$

$$P(y_{i+1}|X, Y_c) \propto \exp(Vh + W \text{encode}(X, Y_c)) \quad (10)$$

ABS モデルは入力が一文となっており、複数文を入力とすることは不可能である。

2.4 マルチ Attention モデル

マルチ Attention モデルについて述べる。マルチ Attention モデルでは ABS モデルを用いて元文書に含まれる各文をエンコードし、そのエンコード結果を統合することで複数文のエンコードをする。マルチ Attention モデルとして、平均マルチ Attention モデルとディープマルチ Attention モデルが提案されている [4]。ここでは平均マルチ Attention モデルについて説明する。平均マルチ Attention モデルは各文の Attention モデルのエンコード結果を平均したものをエンコードの中間表現とする。平均マルチ Attention モデルでは各文が同等に反映されたものとなり、文間に荷重をかけることができない。図 2 に平均マルチ Attention モデルのエンコーダ部を示す。デコーダ部は ABS モデルと同様である。

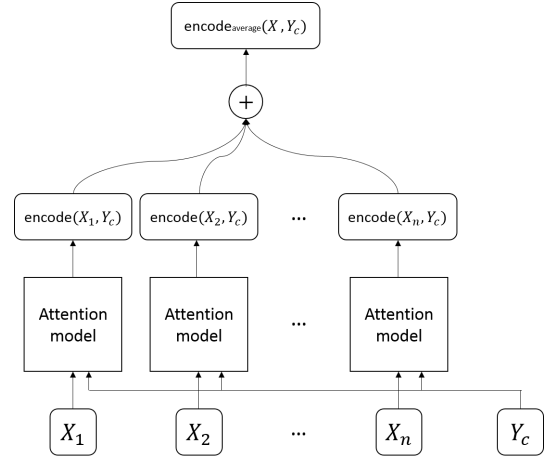


図 2: 平均マルチ Attention モデルの Encoder 部 (文献 [4] をもとに作成)

また平均マルチ Attention モデルのエンコーダの中間表現のベクトルは以下の通りである。

$$\text{encode}_{\text{average}}(X, Y_c) = \frac{\sum_{i=1}^n \text{encode}(X_i, Y_c)}{n} \quad (11)$$

3 実験

3.1 実験 1

3.1.1 実験方法

Attention モデルと 平均マルチ Attention モデルを用いてニュース記事の生成型要約をし、モデルの性能の比較をした。

初めにデータセットについて説明する。livedoor ニュース¹から国内、海外に関するニュース記事とその記事に対する要約文のセットを人手で収集し、学習用データとして用いた。livedoor ニュースでは、ニュース記事とそれに対する人手による 3 文の要約文が掲載されている。収集した 2477 件のニュース記事からランダムに選んだ 2400 件を訓練データ、77 件をテストデータとした。

ABS モデルでは入力文 X としてニュース記事の最初の 1 文(ヘッドライン)を用い、それに対する正解要約文 Y として人手による要約文の最初の 1 文を用いた。マルチ Attention モデルでは入力文 X としてニュース記事全文を用い、それに対する正解要約文 Y として ABS モデル同様に人手による要約文の最初の 1 文を用いた。

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。また出現頻度 1 以下の単語を未知語として扱い、「***」トークンに置換した。未知語を取り除いた総語彙数は 36401 語であった。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表 1 の通りである。

最後に文章生成方法と評価方法について説明する。実験では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、要約文としては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るが、これは NP 困難であり計算時間内に解くことが不可能である。そのため、枝刈りをしながら探索するために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 2 に設定した。

評価方法として、システム要約(生成文)と参照要約(正解文)間において n-gram 単位でどれほど一致したかを表す ROUGE-N [5] という指標を用いた。今回は ROUGE-1, ROUGE-2 を用いた。ROUGE-N の計算式を以下に示す。

¹livedoor ニュース <http://news.livedoor.com/> (2018 年 6 月 10 日アクセス)

表 1: ABS, 平均マルチ Attention モデルにおけるパラメータ

パラメータ	値
エンベディングサイズ	100
隠れ層サイズ	100
FFNN に入れる直前単語数	5
スムージングサイズ	2
optimizer	Adam(alpha=0.001)
epoch 数	200
バッチサイズ	50

表 2: ROUGE 値

	ROUGE-1	ROUGE-2
ヘッドライン抽出	0.5378	0.3114
ABS モデル	0.3168	0.0724
マルチ Attention モデル	0.2641	0.0509

$$\text{ROUGE-N}(C, R) = \frac{\sum_{e \in \text{n-gram}(C)} \text{Count}_{\text{clip}}(e)}{\sum_{e \in \text{n-gram}(R)} \text{Count}(e)} \quad (12)$$

n-gram(C) は、システム要約に含まれる n-gram, n-gram(R) は、参照要約に含まれる n-gram 集合を表す。Count(e) は、ある n-gram の出現頻度を数える関数であり、Count_{clip}(e) は、システム要約における n-gram の出現頻度 Count($e \in \text{n-gram}(C)$) と参照要約における n-gram の出現頻度 Count($e \in \text{n-gram}(R)$) の小さい方の値を採用する。

3.1.2 実験結果

初めに ヘッドライン抽出, ABS モデル, 平均マルチ Attention モデルに対するテストデータ 77 件の ROUGE 値を表 2 に示す。

また実際の生成文例を以下に示す。なお ABS モデルの入力文は記事の一文目, マルチ Attention モデルの入力文は記事の全文である。

記事例 平昌オリンピックの警備員が、ノロウイルスに集団感染した疑いがあることがわかった。韓国の保健福祉省によると、オリンピックの警備員が下痢などの症状を訴え、Num 日までに、約 Num 人にノロウイルスに感染した疑いがもたれている。水道水が原因とみられ、同じ宿舍の、約 Num 人が業務から外れて待機している。

正解文 平昌五輪の警備員約 Num 人がノロウイルスに感染した疑いがある。

ヘッドライン抽出 平昌オリンピックの警備員が、ノロウイルスに集団感染した疑いがあることがわかった。

出力文 ABS 平昌五輪の警備員らがノロウイルスに集団感染した問題。

出力文 平均マルチ Num 日、埼玉県の前が流出した。

ROUGE 値の結果はヘッドライン抽出が最も高かった。これは生成型要約の学習が十分でないこともあるが、ニュース記事の要約においてはヘッドラインが用いられることが多いので、ROUGE 値が高いのは妥当といえる。しかし、ヘッドラインに含まれない情報が有効となる場合もあるので、汎用性を考えると、生成型要約のほうが重要になると考える。

生成型要約について比べると、表 2 より、ABS モデルのほうがマルチ Attention モデルより ROUGE 値が高いことが分かる。また生成される要約文の内容についても、ABS モデルに比べてうまく生成できていなかった。マルチ Attention モデルでは文章は生成できているが、内容や意味的に正しい要約がとれていなかった。全文を入力する場合、冗長性があがるために文生成の精度が落ちたためと考えられる。また全文を入力した際は、訓練データ数が少ないこともあり学習が十分にされなかったことも精度低下の要因と考えられる。また、ニュース記事の要約は最初の文が用いられることが多い。平均マルチ Attention モデルが文間の荷重をかけないモデルであり、すべての文を同等に扱うこのモデルの特性が、精度低下の一因とも考えられる。

今後、入力文を全文ではなく最初の 3 文程度を入力したり、文間に荷重をかけたりすることで平均マルチ Attention モデルより ROUGE の精度が上がると思予想する。

4 まとめと今後の課題

ABS モデルと平均マルチ Attention モデルで文章要約をし、両者のモデルの性能の比較をした。全文を同等に扱うマルチ平均 Attention モデルでは単一文だけを利用する ABS モデルより精度が落ちることが分かった。しかし要約は複数の文を参照して行

われるものであるためにマルチ Attention モデルを用いた要約文生成が望ましい。平均マルチ Attention モデルの精度向上のために、ニュース記事の要約は最初の文が用いられることが多いことを考慮して入力文数の制限や、文間の荷重付与が今後必要になると考えられる。それ以外の今後の課題としては、文間の荷重をかけるディープマルチ Attention モデルの実装、ひとつの記事から複数の要約文を生成する手法の考案などが挙げられる。

現在、複数の要約文の生成法として以下の手法を考えている。

- 記事を単純に 3 分割し、それぞれの分割した複数文章を異なる 3 つのモデルを用いて学習させ、それぞれのモデルから 3 つの要約文を生成する手法
- 抽出型要約 (LexRank 法など) で代表文を 3 文取り出して、その 3 文から生成型要約を利用して要約文を生成する、抽出型要約と生成型要約を組み合わせた手法

参考文献

- [1] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, Vol. 3, No. Feb, pp. 1137–1155, 2003.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [3] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.
- [4] 吉岡重紀, 山名早人. 生成型一文要約のためのマルチアテンションモデルの提案. *DEIM Forum E8-3*, 2016.
- [5] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.