線形化された構文情報を用いた生成型ニューラル文要約

Abstractive Neural Sentence Summarization with Linearized Syntax Structures

瀧川　雅也　　三輪　誠　　佐々木　裕
Masaya Takikawa　Makoto Miwa　Yutaka Sasaki
豊田工業大学
Toyota Technological Institute

{sd13057, makoto-miwa, yutaka.sasaki}@toyota-ti.ac.jp

1. 研究背景と目的

Web などの情報の増加に伴い，情報収集の労力も増加し続けている．そこで注目されているのが文要約の技術である．文要約の手法としては Jing[1]を筆頭に，構文木を刈り込む抽出型の手法が長らく使われてきた．近年，ニューラルネットワーク（NN）を用いた生成モデルが登場し，それを文要約に応用することで，生成型要約を実現する手法が提案された．しかし，この手法では構文情報を利用していない．要約において構文情報は，多重修飾などの冗長性を排除するために有用である．そこで本研究では，既存の生成モデル[2]を基に，新たに構文情報を利用することで，より良い要約文の生成を目指す．

2. 関連研究

2.1. 構文木を用いた要約

文要約の主要な手法は，要約元文から単語や句などを取り除く方法である． Jing[1]は，要約元文の構文木から，不要な要素を取り除くことにより，文を要約する手法を提案している．この手法の利点は，元文の語彙をそのまま利用できる点や，正しい文法を保ちやすい点にある．一方で，元文から非重要箇所を削除するだけなので，人間のように生成型の要約が実現できない欠点がある．

2.2. ニューラルネットを用いた要約

Nallapati ら[3]は、機械翻訳のタスクで提案されたアテンションモデル[2]を文要約に適用し，生成型文要約のタスクで最先端の結果を出した．アテンションモデルは，エンコーダ・デコーダモデルをベースとしている．

エンコーダ･デコーダモデルとは,Cho ら[4], Sutskever ら[5]によって提案された，再帰型ニューラルネットワーク（Reccurent NN; RNN）を 2 つ用いるモデルである．エンコーダ側で入力文を RNN によりベクトル化し，デコーダ側でそのベクトルを初期状態とし，別の RNN によって出力文を生成する．このエンコーダ・デコーダモデルには，入力文が長文になると，情報の圧縮ロスが大きくなり，性能が下がるという問題がある．そこで，長文に対応するため提案されたのが，Bahdanau ら[2]によるアテンションモデルである．

アテンションモデルは，入力文中の注目するべき単語に焦点を当てて，出力文を生成するモデルである．注目するべき単語に焦点を当てるため，効率よく情報を圧縮することができ，長文でも比較的質の高い文を生成できる．入力文 x_1，x_2，…，x_Tを与えた時に，t 番目の出力y_tを生成するモデル図を図 1 に示す．その時の生成確率は次式で表される．

$$\mathrm{P}(y_t|y_1,\dots,y_{t-1},\boldsymbol{x}) = g(y_{t-1},\boldsymbol{s_t},\boldsymbol{c_t})$$

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

$$c_t = \sum_{j=1}^{T} \alpha_{t,j} h_j$$

$$\alpha_{t,j} = \frac{\exp(e_{t,j})}{\sum_{k=1}^{T} \exp(e_{t,k})}$$

$$e_{t,j} = \alpha(s_{t-1}, h_j)$$

s_tは t 番目のデコーダの隠れ状態で，文脈ベクトルc_tは j 番目の入力の隠れ状態h_jとアテンション荷重$\alpha_{t,j}$の積和である．j 番目のアテンション荷重$\alpha_{t,j}$は 1 つ前の隠れ状態s_{t-1}と j 番目の入力の隠れ状態h_jで計算される．g，f，αはそれぞれ非線形関数である．

3. 提案手法

Bahdanau ら[2]が提案したアテンションモデルを基に，新たに構文情報を利用した生成型文要約を行う．構文情報には，句構造，係り受け構造を用いる．構文情報の利用は，入力文の表現形式を変えることで実現する．TreeLSTM[6]で構文情報を捉える手法も考えられるが，学習に時間が掛かる欠点がある．そこで，今回は線形化された構文情報を利用する．

まず，構文木を得るために要約元文を構文解析する．そして得られた構文木を S 式で表現した形を，線形化された構文情報として，平文の代わりに入力にする．句構造木とその S 式表現の例を図 2 に示す．これをアテンションモデルで学習させることによって，構文情報を考慮した生成型文要約の実現が期待できる．

4. 評価実験

4.1. データセット

今回の評価実験は，英語のニュース記事を対象に行う．要約文と要約元文をモデルに学習させ，そのモデルで要約文を生成し，スコアを確認する．データとしては，CNN とデイリー・メー

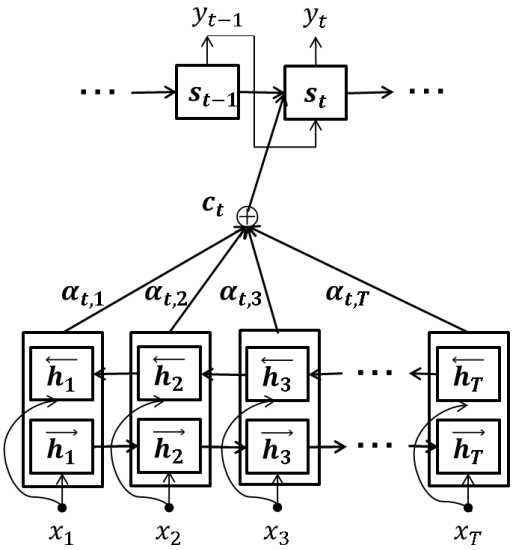

図 1. アテンションモデル

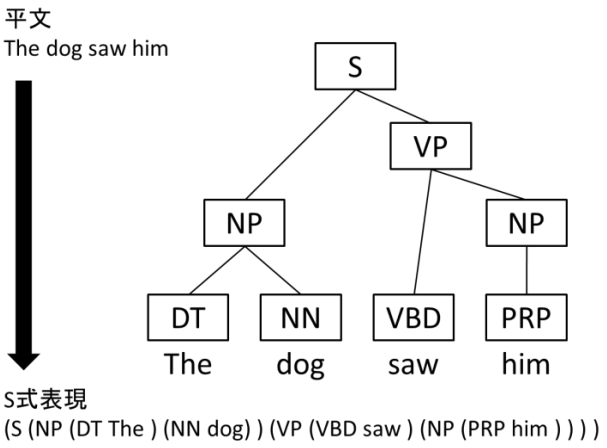

図 2. 句構造木と S 式表現

ルのニュース記事 29 万件を用いる．これらには，一つの記事に対して，三，四文の人手で書かれた要約文がついている．その要約文と，それに対応する本文中の一文をセットにして，要約文と要約元文として扱う．そのための要約元文の決定については，単語の一致度を評価する ROUGE-L [7]で決定する．ROUGE-L は，m 単語からなる文 X と，n 単語からなる文 Y との最長共通部分列（Longest Common Subsequence; LCS）を用いて，次のように算出される．

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad \beta = \frac{P_{lcs}}{R_{lcs}}$$

要約文とニュース記事本文の各文に対する ROUGE-L の R 値を算出し，値が最大の文を要約元文として採用する．また，質の悪いデータを省くために,ROUGE-L の R 値に閾値を設定した．閾値の設定に際しては，ROUGE-L の R 値に対するデータとしての有用性を，人手で五段階評価して検討した．評価基準を表 1 に，評価結果を表 2 に示す．ROUGE-L の R 値 0.1 毎に 100 件の文対を評価した所，0.4 未満は信頼度が極端に低く，0.6 以上は信頼度が高かった．また，0.6 以上は単語を抽出するだけで要約できるような文が多かった．そこで，閾値は 0.4 か 0.5 の二択に絞り，小規模データで良い結果を出した 0.5 を閾値として採用した．

4.2. 実験設定

今回の実験では，構文情報を用いた場合と用いなかった場合で要約を行い，ROUGE スコアの F 値による比較を行った．学習データには，一文あたり 100 単語以下，かつ ROUGE-L の R 値が 0.5 以上の組 38 万文を用意し,開発･テストデータ用にそれぞれ 1,000 文用意した．また，構文解析には Stanford Parser（version 3.5.2）[1]を用いた．今回使用したモデルでは，エンコーダに 3 層の双方向 RNN を用いている.エンコーダ，デコーダともに隠れ層に Gated Recurrent Unit を使い，次元数は 128 次元に設定した．単語ベクトルも 128 次元で,事前学習は行っていない．語彙は学習データ中に二回以上現れた単語を採用し，入力語彙数は 9 万，出力語彙数は 6.5 万となった．学習には Adam を使い，バッチサイズは 64 に設定した．ドロップアウトや正則化は行っていない．

表 1.　評価基準

評価	評価基準
5	9 割方要約文の内容が元文にある． ほぼすべての単語が一致．
4	7 割方要約文の内容が元文にある． 一部（日付や地域）の情報が欠落．
3	5 割方要約文の内容が元文にある． 複数の情報が欠落，残りは別の文．
2	3 割方要約文の内容が元文にある． 一部（固有名詞や日付，地域）の 情報だけ一致．
1	要約文の内容が元文にない． 一般的な単語だけ一致．

表 2.　ROUGE スコア毎の評価結果

評価＼ROUGE	0.3～	0.4～	0.5～	0.6～	0.7～
5	29	50	69	67	81
4	24	27	22	29	19
3	31	19	7	4	0
2	13	4	2	0	0
1	3	0	0	0	0

5.　結果と考察

開発データの評価結果を表 3 に示す．評価結果より，句構造を用いた場合，平文のままよりもスコアが大きく上昇していることが確認できる．この結果より，要約において句構造が有用であると言える．しかし，係り受け構造を利用した場合は，句構造のようにスコアが伸びなかった．係り受け構造を用いても，スコアが伸びなかった原因として，語順の欠落が考えられる．平文を係り受け構造木で表現すると，文の語順がバラバラになってしまう．その対策として，語順の情報を特徴として入力する方法が考えられる．

句構造を利用した場合と平文の場合のアテンションヒートマップを図 3 に示す．横軸が入力で，縦軸が出力である．平文では，注目した入力単語がそのまま出力単語と一致することが多い．一方，句構造の場合，出力と同じ単語

表 3.　各入力形式における ROUGE スコア

入力形式	ROUGE-1	ROUGE-2	ROUGE-L
平文	0.3554	0.1869	0.3370
句構造	**0.3718**	**0.2060**	**0.3526**
係り受け	0.3497	0.1668	0.3296

[1] http://nlp.stanford.edu/software/lex-parser.shtml

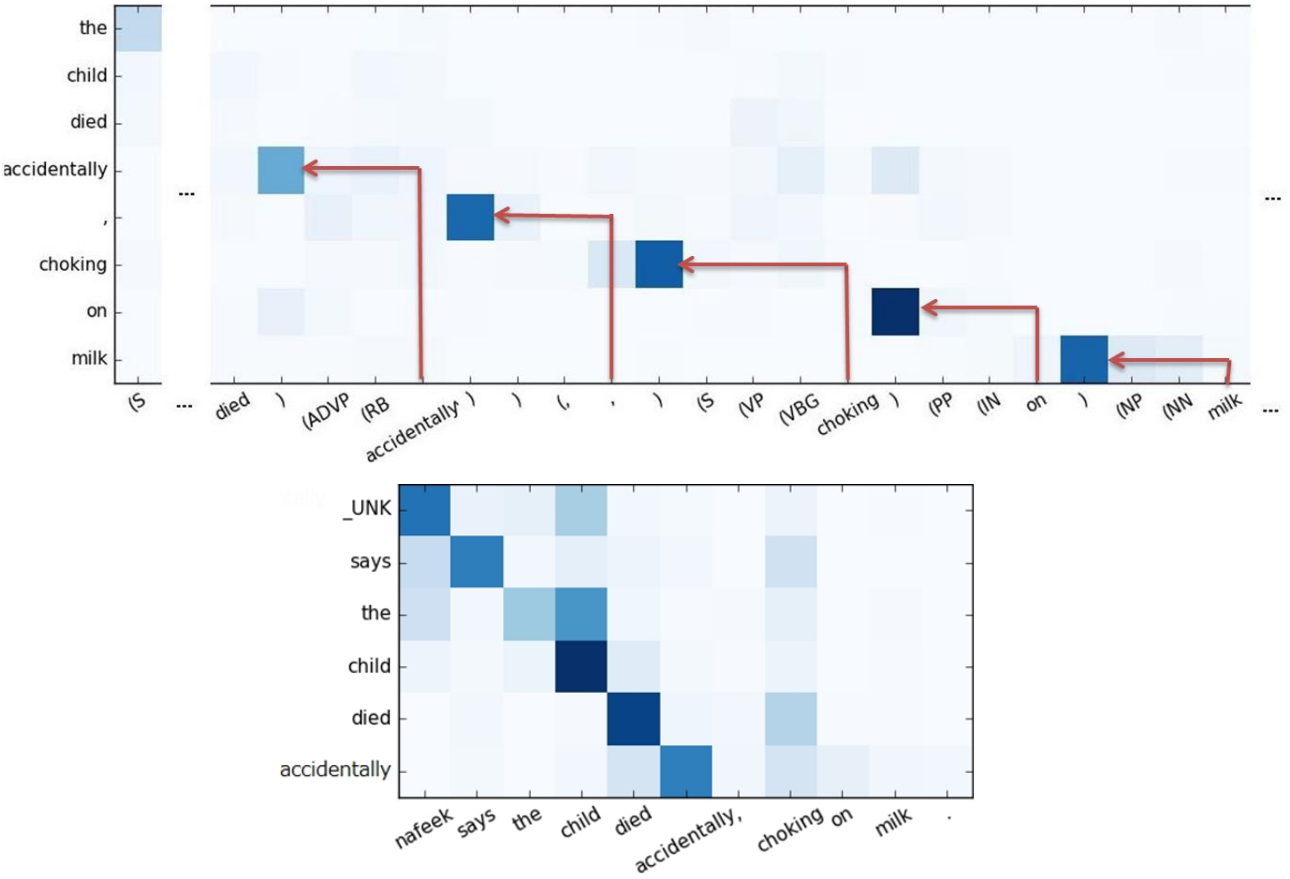

図 3. アテンションヒートマップ（上：句構造，下：平文）

が入力にある場合，そこから文頭方向にある「)」を注目しているように見える．これは単語だけでなく，句構造に注目していると考えられる．

6. まとめ

本研究では，構文情報の利用による要約性能を調べるため，線形化された構文情報を用いた生成型ニューラル文要約を提案した．この手法で実験した結果，句構造を利用した場合，平文の時よりも ROUGE スコアが上昇した．これより，要約における句構造の有用性が確認できた．今回用いた手法は，入力形式を変えただけであるため，直接的に構文を理解できる形にはなっていない．今後の課題は，直接的に構文情報を扱えるモデルの考案である．

参考文献

[1] Hongyan Jing, Sentence Reduction for Automatic Text Summarization, In Proc. of NAACL 2000.

[2] Dzmitry Bahdanau et al., Neural machine translation by jointly learning to align and translate. In Proc. of ICLR, 2015

[3] Ramesh Nallapati et al., Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond, In Proc. of CoNLL 2016.

[4] Kyunghyun Cho et al., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, In Proc. of Conference on EMNLP 2014.

[5] Ilya Sutskever et al., Sequence to sequence learning with neural networks, In Proc. of NIPS 2014.

[6] Kai Sheng Tai et al., Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, In Proc. of ACL2015.

[7] Chin-Yew Lin, ROUGE: A Package for Automatic Evaluation of Summaries, In Proc. of ACL2004.