The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

自己学習を用いたニューラル見出し生成

Neural Headline Generation with Self-Training

竹前 慎太郎 *1 Shintaro Takemae　村尾 一真 *2 Kazuma Murao　谷塚 太一 *2 Taichi Yatsuka　小林 隼人 *2*3 Hayato Kobayashi　野口 正樹 *2 Masaki Noguchi　西川 仁 *1 Hitoshi Nishikawa
徳永 健伸 *1 Takenobu Tokunaga

*1東京工業大学 Tokyo Institute of Technology　*2ヤフー株式会社 Yahoo Japan Corporation　*3理化学研究所 AIP センター RIKEN AIP

In this paper we propose a novel method which incorporates self-training into a sequence-to-sequence model in order to improve the accuracy of the headline generation task. Our model is based on neural network-based sequence-to-sequence learning with an attention mechanism and trained with approximately 100,000 labeled examples and 2,000,000 unlabeled examples. Through experiments, we show our proposal significantly improves the accuracy and works effectively.

1. はじめに

自動見出し生成は, 自動要約研究における重要な課題のひとつである [Nenkova and McKeown 2011]. 見出し生成課題は一般的に, 新聞記事などの文章に対して, その文章の内容をより少ない文字数で表す文を生成する課題である. 自動要約によって作成された見出しを提供することで, 読み手は少ない文字数の見出しから記事の内容を把握し, 読むべきかどうかをより迅速に判断できるようになるため, この課題は実用上非常に有用である. 見出し生成における一般的な設定では記事に付属しているタイトル（記事タイトル）が生成されるが, 本論文では, あるニューストピックを指し示すための, 記事タイトルよりもさらに短い文字数の見出し（トピックス見出し）を生成する方法を提案する. Yahoo!ニューストピックスのような総合ニュースサイトにおいては, 記事タイトルの全文字を描画領域に表示できないことが頻繁に起こるため, 記事タイトルよりも短く, そのトピックにおける重要な情報を含む見出しを生成することができれば, 読み手の負担を減らし効率よく記事を提供できるようになる.

一方で, 近年, 自動要約システムは機械学習を用いた手法に大きく依存している. 機械学習の枠組みにおいては, 訓練データ, すなわち入力とそれに対応する出力の組を準備することが必要である. しかしながら, 出力に相当するトピックス見出しは記事タイトルや内容から編集者の手によって作成されており, その作業コストが高いことからこのようなデータセットを十分に用意することは困難である. 学習のためのデータが不足しがちという問題点を克服するため, 本論文では自己学習によって学習データを拡張する手法を提案する. 提案手法では, まず最初のモデルを人間の手によって作成された学習データセットから作成し, 続いて, 記事タイトルのみを持つデータに対して, このモデルを用いて疑似トピックス見出しを付与する. その後, 疑似トピックス見出しが付与されたデータと, 人間がトピックス見出しを作成したデータの両方のデータセットを学習データとしてモデルの再学習を行う. 比較実験により, 提案手法がベースライン手法よりも統計学的に有意に高い精度を達成することを確認した.

連絡先: *1{takemae.s.aa@m, hitoshi@c, take@c}.titech.ac.jp
*2{kmurao, tyatsuka, hakobaya, manoguch}@yahoo-corp.jp

2. 関連研究

自動要約についての手法は, 抽出的要約と生成的要約の大きく 2 つに分類される. 抽出的要約では, システムは入力文章の一部分を抜き出し, それを組み合わせることで出力文を構成する. 生成的要約では, 入力文書に含まれない全く新しい表現を含む出力文を生成する. 後述するように, 編集者がトピックス見出しを作成するときにはしばしば入力文にない表現を用いることがあるため, 本研究課題では生成的要約のアルゴリズムを適用することが望ましいと考えられる.

見出し生成については多くの研究がなされており [Banko et al. 2000, Knight and Marcu 2002, Alfonseca et al. 2013], 近年ニューラルネットワークを用いる手法が活発に研究されている [Rush et al. 2015, Filippova et al. 2015, Takase et al. 2016, Ayana et al. 2016, Xu et al. 2016, Chopra et al. 2016, Zhou et al. 2017, Tan et al. 2017, Tilk and Alumäe 2017]. ニューラルネットワークを用いたアルゴリズムでは主に, 文章の単語分散表現の系列を入力とし, encoder-decoder 型のネットワークにアテンション機構を取り入れた構成によって出力文を生成する.

自己学習は, 自然言語処理において広く用いられており, 構文解析 [McClosky et al. 2006] や語義曖昧性解消 [Mihalcea 2004] などに利用されている. sequence-to-sequence の研究においても, Dai と Le によって半教師あり学習の手法が提案されている [Dai and Le 2015]. 彼らは, 前段の教師なし学習によって後段の教師あり学習におけるより良い初期パラメータを計算する事前学習手法を提案した. 彼らの手法と異なり, 本論文の提案手法では, 正解付きのデータセットを用いて前段のモデル学習を行うが, その後, 正解を持たないデータセットに対してこのモデルから疑似正解を付与し, これらの正解付き・疑似正解付きの両方のデータセットを利用して学習することで最終的なモデルを獲得する. 前段の教師あり学習によって作成されるモデルと, 後段の自己学習を経たモデルを比較した結果, 後者のモデルが前者のモデルより高い精度

記事タイトル	トピックス見出し
＜天皇陛下＞退位後に「赤坂御用地」に転居　宮内庁検討	陛下 退位後は赤坂に転居案

表 1: 記事タイトルと対応するトピックス見出しの例 *4

を達成することを確認した.

3. 提案手法

本論文の提案手法は以下のような手順による.

1. 教師あり学習 (ベースライン): ベースラインのモデルは, 人間の作成したトピックス見出しと記事タイトルの組から学習される.
2. 疑似見出しの生成: 次に, 記事タイトルのみからなるデータセットに対して, 手順 1 で作成したモデルを用いて疑似トピックス見出しを作成する.
3. 提案モデルの学習: 最後に, 手順 1 のオリジナルデータと, 手順 2 で疑似トピックス見出しが付与されたデータの両方を用いて, 提案モデルの学習を行う.

言い換えると, 提案手法においては, ベースラインのモデルから新しいデータセットを作成することで自己学習を行っている.

4. 実験

4.1 実装と設定

ベースラインのモデルを作成するため, OpenNMT ツールキット [Klein et al. 2017] を利用した. encoder-decoder モデルのネットワークとして, Luong らによって提案された, stacking recurrent neural network に global attention を加えた構成を用いている [Luong et al. 2015].

encoder の入力には 3000 次元の one-hot ベクトルを埋め込んだ分散表現を用い, 250 次元の内部ベクトルを持った 3 層の隠れ層による双方向 LSTM を採用した. decoder においても, 出力した語の再入力に同様の分散表現を用いたが, 3 層の隠れ層の内部ベクトルの次元数は 500 とした. 学習時の最適化手法として Adam [Kingma and Ba 2014] を採用し, そのハイパーパラメータはそれぞれ, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\varepsilon = 10^{-9}$, $\alpha = 10^{-3}$ とした. 前段の学習と後段の学習はそれぞれ 30 エポックずつ行い, そのミニバッチサイズは 64, dropout 率は 0.3 とした.

本研究では対象とする見出しが非常に短いことから, 分散表現として単語単位ではなく文字単位の入力を用いた. すなわち, システムは記事タイトルを 1 文字ずつ読み込み, その出力として見出しを 1 文字ずつ生成する.

また, 出力時には探索幅 5 によるビームサーチを用いて見出しを生成した.

4.2 比較手法

以下の 3 手法を比較した.

- ベースライン手法: 前述の通り, アテンション機構付きの一般的な encoder-decoder 型モデルをベースラインとして採用した.
- 事前学習手法: 他の半教師あり学習との比較のため, Dai と Le が提案した sequence autoencoder に基づく事前学習手法 [Dai and Le 2015] を比較対象とした. この手法では, 正解見出しなしデータを用いた sequence autoencoder モデル (記事タイトルから記事タイトルへの sequence-to-sequence モデル) の学習を行った後, そのパラメータで初期化したモデルの学習を行う. このとき初期化するパラメータの対象を全てのネットワークパラメータとする場合と, アテンション機構を除く encoder-decoder 部分のみのパラメータを初期化する場合について実験を行った. なお, Dai と Le はこの事前学習手法を感情分析課題で評価している. そのため, 事前学習手法が想定していた, 事例を正負に分類する単純な分類問題と比較して, 本研究課題のほうが複雑であることを考慮に入れる必要がある.
- 提案法: 正解見出しを持つデータと, 疑似正解見出しを付与したデータの双方を学習データとして学習を行った.

4.3 データ

実験用のデータセットとして, 日本語ニュース記事タイトルと, それに対応するトピックス見出しの組を用いた. 表 1 にその例を示す. トピックス見出しにおいては全体の長さをコンパクトにまとめるため, 記事のタイトルに含まれているいくつかの語を省いたり, 言い換えなどが行われている.

クローリングによって, 記事タイトルのみからなるデータセットを準備した. このうち, トピックス見出しと紐付けられた約 10 万組を教師ありデータとし, 紐付けをしなかったものを自己学習用のデータとした. 教師ありデータは, 約 60%を学習データ, 約 20%を開発データ, 約 20%をテストデータとして分割した. 自己学習を経たモデルの精度にデータセットのサイズが与える影響を確認するため, 自己学習用データからそれぞれ 20 万件, 40 万件, 80 万件, 100 万件, 150 万件, 200 万件をサンプリングしたデータセットを作成した.

トピックス見出しは全て 13.5 文字以内で記述されている. 日本語のひらがな, カタカナ, 漢字はそれぞれ 1 文字として数えられ, 英語のアルファベット, 数字および空白文字については 0.5 文字換算として数えられる. すなわち, 編集者は 13 文字の日本語の文字と空白文字 1 つによってトピックス見出しを作成することができ, またアルファベットや数字を用いる場合はそれらの 2 文字を日本語の 1 文字分と換算することができる.

4.4 評価尺度

評価尺度として, 文字ベースの ROUGE-1 指標を用いた [Lin 2004]. これは比較対象のシステムが単語ではなく文字を単位として出力するためである. データセット中の記事タイトルとトピックス見出しは非常に短く, トピックス見出しはしばしば言い換えが用いられている. 単語ベースの ROUGE は言い換え表現を過剰に低く評価するが, 文字ベースで評価すると, 元の単語を構成する文字を抽出したような略語についてはその語を正解に含まれるものとして評価できる. このように, 文字ベースの評価を行うことで, 単語ベースの評価よりもより正確な生成文字列の評価が可能であると考えられる.

また評価において初期値による影響を軽減するため, それぞれの手法に対し, 異なる初期値による 10 回の学習とトピック

*4 https://news.yahoo.co.jp

	ROUGE-1
ベースライン	0.568
事前学習（アテンション機構除く）	0.520
事前学習（全て）	0.503
提案手法（60 万件）	**0.574**

表 2: ROUGE-1 の平均

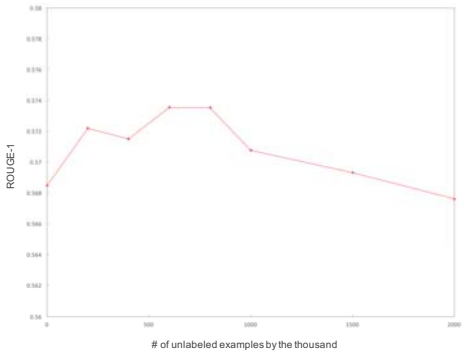

図 1: 正解なしデータのサイズに対する提案手法の ROUGE-1

ス見出しの生成を行い, その評価値の平均をその手法のスコアとして採用した.

5. 結果と考察

表 2 に実験結果を示す. 提案手法のスコアは, 60 万件の自己学習用データセットを用いた場合のモデルを採用した. 提案手法によるモデルは, 自己学習を行わないベースラインモデル, および事前学習を行ったモデルより高い精度を達成した. t 検定により, 提案法と比較手法の間の ROUGE-1 スコアは統計学的に有意であることが示された ($p < 0.05$).

事前学習手法はベースラインよりも低い精度となった. 全てのパラメータを事前学習する手法と比較して, アテンション機構のパラメータを事前学習による初期化の対象から除く手法が良い精度を達成した. このことは, アテンション機構が, 入力された記事タイトルに含まれる情報のうち, どのような情報をトピックス見出しに残すべきかといった, 教師なし学習からは本質的に学習できないパラメータを学習しており, そのため正解のないデータからは望ましい初期値を得られなかったことを示唆している.

図 1 に, 自己学習用のデータセットのサイズに対する提案法の精度の変化を示した. データセットのサイズが 60 万件, および 80 万件のときスコアは 0.574 と最も高い値を記録した. データセットのサイズが小さいとき, もしくは大きすぎるときに精度が低下する現象は, 正解なしデータが多くなりすぎることで, 本来学習すべき対象である正解ありデータよりも大きな影響を持ってしまうことで生じると考えられ, 自己学習ではしばしば確認される. そのため, 自己学習用に疑似正解を付与する際に望ましい出力例のみを採用することができれば, 自己学習による精度向上はより大きくなると考えられる.

表 3 に, 提案法と比較手法による出力文字列の長さ (文字ベース, 半角英数字も 1 文字として数える) を掲載した. 要約率は約 50%で, 機械によって生成された見出しの長さはどれも, 編集者が作成したものと比較して若干短い傾向にあった.

図 4 に, 評価データに対するモデルの出力例をいくつか掲載した. 1 番目の例では, 記事タイトル中の「割引切符」という名詞を, モデルによる出力で「割符」と略語化している. 提案手法においては単語単位ではなく文字単位で処理が行われるため, 複合語が再構成されて出力されることがある. この出力例に対し ROUGE 値は高い値を示すが, 「割引切符」と「割符」は異なるものを指し示すため, この出力例から正しい意味を読み取ることは難しい.

	文字数
本文	595.6
記事タイトル	26.4
トピックス見出し	14.0
ベースライン	13.6
事前学習（アテンション機構除く）	13.3
事前学習（全て）	12.8
提案手法（60 万件）	13.5

表 3: 平均文字数

2 番目の例では, ベースラインモデルは「ロボット」という語を反復して出力し, この見出しで正しい記事内容を説明することは困難である. 一方で, 提案手法の出力では「ソフトバンク」という文字列を「ソフト B」のようにうまく略語化しているように見えるが, 正解トピックス見出し中に「ソフトバンク」に相当する語がなかったため ROUGE 値はベースラインモデルを下回った.

3 番目の例においては, 正解の見出しは記事タイトルの削除操作で生成することができ, 学習モデルはこれを模倣している. Filippova らの先行研究によって, ニューラルネットワークによる文圧縮モデルは, データセット中に存在する正解文字列が全て元文字列の削除操作のみから生成することができる場合に比較的良い結果を見せることが報告されている [Filippova et al. 2015].

4 番目の例のように正解見出しが記事タイトルと全く異なっていると, モデルが生成した見出しがある程度意味をとれるものであっても, 正解見出しと共通する文字を持たないことから ROUGE-1 が低い値をとる場合がある.

最後の例では, 提案手法による出力はベースラインモデルのものよりも望ましい出力を得た. ベースラインモデルは記事の重要な要素とみられる「試験」という語を生成見出しから欠落させたが, 提案法ではこれを残したままトピックス見出しを生成できた.

全般的に, 人間の編集者によって作成されたトピックス見出しは洗練された言い換え表現を含んでおり, これらはしばしば文字単位で置き換えられている. 正解のトピックス見出しが入力文の削除操作のみから生成することができる場合には, ベースラインモデルと提案モデルの双方から比較的良い出力結果を得ることができた.

6. まとめ

本論文では, アテンション機構付き sequence-to-sequence のモデル学習に, 自己学習を追加する学習手法を提案した. 実験により, 自己学習を追加して行うことで学習モデルの精度が有意に向上したことが示された.

5 節で議論したように, 自己学習のために生成された疑似正解見出しの中から品質の高い組のみを選ぶことができれば, より自己学習の効果が高まると考えている. また, 人間の編集者はトピックス見出しを作成する際に記事タイトルだけでなく記事の内容などを考慮に入れていることから, モデルにおいても記事リード文などの情報を考慮することで生成見出しの品質を高めていくことを検討している.

記事タイトル	トピックス見出し	ベースライン	提案手法
JR 東日本の割引切符に若者からため息，高齢者優遇の背景は？	JR のシニア優遇に若者ため息	JR 東の割引切符高齢者の背景 (0.462)	JR 東割符に若者からため息 (0.615)
ソフトバンクがロボット事業に参入，日本のロボット産業どうなる？	日本のロボット産業 未来は?	日本ロボットロボットに参入 (0.462)	ソフト B ロボット事業に参入 (0.385)
＜ボクシング＞亀田ジムが新会長での再出発を検討	亀田ジム 新会長で再出発検討	亀田ジム新会長で再出発検討 (1.000)	亀田ジム新会長で再出発検討 (1.000)
カーボヴェルデ代表が世界ランクで日本を追い抜いた理由	FIFA ランク 無名国がなぜ躍進	カーボヴェルデ代表日本追い抜く (0.000)	カーボヴェルデ代表日本追い抜く (0.000)
ドワンゴの入社試験有料化，その是非をどう考えるべきか	入社試験「有料化」の是非は	ドワンゴ有料化その是非どう (0.462)	ドワンゴの入社試験有料化へ (0.615)

表 4: 出力例 と ROUGE-1 *5

参考文献

[Alfonseca et al. 2013] Alfonseca, E., Pighin, D., and Garrido, G. (2013). Heady: News headline abstraction through event pattern clustering. In *Proceedings of the 51st Annual Meeting of ACL*, pages 1243–1253.

[Ayana et al. 2016] Ayana, Shen, S., Liu, Z., and Sun, M. (2016). Neural headline generation with minimum risk training. *CoRR*, abs/1604.01904.

[Banko et al. 2000] Banko, M., Mittal, V. O., and Witbrock, M. J. (2000). Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on ACL*, pages 318–325.

[Chopra et al. 2016] Chopra, S., Auli, M., and Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of NAACL-HLT*, pages 93–98.

[Dai and Le 2015] Dai, A. M. and Le, Q. V. (2015). Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28*, pages 3079–3087.

[Filippova et al. 2015] Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on EMNLP*, pages 360–368.

[Kingma and Ba 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

[Klein et al. 2017] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *CoRR*, abs/1701.02810.

[Knight and Marcu 2002] Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 1(139):91–107.

[Lin 2004] Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of ACL Workshop Text Summarization Branches Out*, pages 74–81.

[Luong et al. 2015] Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on EMNLP*, pages 1412–1421.

[McClosky et al. 2006] McClosky, D., Charniak, E., and Johnson, M. (2006). Effective self-training for parsing. In *Proceedings of the Main Conference on NAACL-HLT*, pages 152–159.

[Mihalcea 2004] Mihalcea, R. (2004). Co-training and self-training for word sense disambiguation. In *Proceedings of the Eighth Conference on CoNLL*, pages 33–40.

[Nenkova and McKeown 2011] Nenkova, A. and McKeown, K. (2011). *Automatic Summarization*. Now Publishers.

[Rush et al. 2015] Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on EMNLP*, pages 379–389.

[Takase et al. 2016] Takase, S., Suzuki, J., Okazaki, N., Hirao, T., and Masaaki, N. (2016). Neural headline generation on abstract meaning representation. In *Proceedings of the 2016 Conference on EMNLP*, pages 1054–1059.

[Tan et al. 2017] Tan, J., Wan, X., and Xiao, J. (2017). From neural sentence summarization to headline generation: A coarse-to-fine approach. In *Proceedings of IJCAI-17*, pages 4109–4115.

[Tilk and Alumäe 2017] Tilk, O. and Alumäe, T. (2017). Low-resource neural headline generation. *CoRR*, abs/1707.09769.

[Xu et al. 2016] Xu, L., Wang, Z., Ayana, Liu, Z., and Sun, M. (2016). Topic sensitive neural headline generation. *CoRR*, abs/1608.05777.

[Zhou et al. 2017] Zhou, Q., Yang, N., Wei, F., and Zhou, M. (2017). Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104. Association for Computational Linguistics.

*5 JR 東日本は東日本旅客鉄道株式会社の商標または登録商標です. ソフトバンクの名称は, 日本国およびその他の国におけるソフトバンクグループ株式会社の登録商標または商標です. ドワンゴは株式会社ドワンゴの商標または登録商標です. その他、本論文中に記載されている会社名および商品・サービス名は、各社の商標または登録商標です.