見出し生成の忠実性の改善

松丸 和樹　高瀬 翔　岡崎 直観
東京工業大学
{kazuki.matsumaru, sho.takase, okazaki}@nlp.c.titech.ac.jp

1 はじめに

近年，新聞記事の本文から見出しを自動的に生成する，すなわち**見出し生成**の研究が盛んである．機械翻訳研究でエンコーダ・デコーダモデルが発展を遂げたことを受け，新聞記事の本文を入力，見出しを出力と見なし，記事・見出しの組を訓練データとして大量に与え，見出し生成モデルを学習する研究が主流である [9]．深層学習ベースの手法は大量の新聞記事コーパスと相性がよく，流暢な見出しを生成するモデルを容易に構築できる．

ところが，自動生成された見出しには**忠実性** — 生成された見出しが伝えるすべての事柄が元記事に基づいているか — に関する懸念がある [1]．これは，報道のような応用を考えたとき，深刻な問題を引き起こす．例えば，表1の自動生成見出しは「あす投開票」という記述を含むが，正しくは「きょう投開票」であるので，国民主権の根幹を脅かすフェイクニュースである．

しかし，人間が表1の記事本文を読んでも，投開票が今日なのか明日なのか判断できない．つまり，「きょう」や「あす」のような表現を含む見出しを生成することに無理がある．本研究では，記事内容に忠実ではない（逸脱した）見出しが生成される原因は，見出し生成のタスク設定の不備や訓練データにあるのではないか，という仮説を検証する．まず，English Gigaword と JApanese MUlti-Length Headline Corpus (JAMUL) のデータセットを分析し，見出し生成に使われるデータセット中に，記事内容に忠実でない見出しがどの程度存在するのか調査する．その結果，訓練データの約 30–40% に，記事に忠実でない見出しが含まれることを報告する．

この調査結果に基づき，訓練データから記事に忠実でない見出しを排除することで，見出し生成モデルの忠実性が改善されるのではないか，という仮説を立てる．この仮説の検証のため，記事に対する見出しの忠実性を推定する含意関係認識器を構築し，訓練データから見出しの忠実性が低い事例を除去（フィルタリング）する．

実験の結果，元々の訓練データとフィルタリングされた訓練データで学習した見出し生成モデルを比較したとき，ROUGE スコアは明確な違いを示さないが，自動評価および人手評価で忠実性を測定すると，フィルタリングされた訓練データで学習したモデルの方が顕著に高い忠実性を示した．

表1: 不適切な見出し生成の例．

記事本文	衆院選は 14 日に投開票される。前回 2012 年は 19 人だった県内五つの小選挙区の候補者は、今回 14 人に減少。少数激戦になった。
自動生成	14 候補、最後の訴え　あす投開票　衆院選
実際の見出し	14 候補、最後の訴え　きょう投開票　深夜に大勢判明

2 タスク設定の分析

2.1 本研究で用いるデータセット

Annotated English Gigaword コーパス[*1]は抽象型要約の研究で最も人気のあるコーパスの 1 つである．Rush ら [9] はこのコーパスの記事の先頭 1 文を入力とし，対応する見出しを出力対象とした．本研究では Rush らの研究に基づき，Annotated English Gigaword の約 380 万事例を訓練，約 39 万事例を開発，約 38 万事例を評価データに用いる．

また，日本語の見出し生成モデルのための大規模コーパスとして，Japanese News Corpus (JNC)[*2] を用いる．JNC は朝日新聞の記事と紙面見出しの組 1,831,812 件を収録しており，記事は冒頭 3 文のみが収録されている．また，評価用データセットとして，JApanese MUlti-Length Headline Corpus (JAMUL) も公開されている．JAMUL は朝日新聞デジタルで配信された 1,524 件の記事全文と紙面見出し，10，13，26 文字以内の各種デバイス向け見出しが付与されたデータセットである．本研究

[*1] https://catalog.ldc.upenn.edu/LDC2012T21
[*2] https://cl.asahi.com/api_data/jnc-jamul.html

表2: 記事が見出しを含意する組の割合.

データセット	先頭 1 文	先頭 3 文	全文
Gigaword	70.3%	N/A	92.8%
JAMUL	N/A	61.4%	94.2%

では，紙面見出しを使って実験を行う．

2.2 見出しが記事本文から逸脱する割合

2.1節で説明した Gigaword と JAMUL について，見出しが記事本文から逸脱している事例がどのくらいあるか調査する．本研究では，見出しが記事内容に忠実であるか否か，という判定を含意関係認識の問題として捉える．含意関係認識とは，前提と仮説が与えられた時に，前提が正しければ仮説も正しいと言える（含意する）か否かを判定するタスクである．これを記事と見出しに適用し，訓練データや評価データの記事と見出しの組に対して，記事（前提）が見出し（仮説）を含意しているものを，忠実な見出しと判定する．

本研究では，English Gigaword の評価セットから 1,000 組，JAMUL から 1,000 組をランダムに選択し，3 人の被験者に記事と見出しの含意関係を判定してもらった．表2に、2 人以上の被験者が「含意する」と判定した記事と見出しの組の割合を示す．Gigaword データでは，記事の先頭 1 文から見出しを生成する実験設定がよく用いられる．ところが，この実験設定では先頭 1 文と見出しが含意する割合は 70.3% に留まり，残りの事例では記事の先頭 1 文だけでは見出しを生成するのに必要な情報が不足する．参考のため，記事全文を見出しの間の含意関係を判定したところ，記事全文と見出しが含意する割合は 92.8% に上昇した．

JAMUL に対応する学習データである JNC は，記事の先頭 3 文と見出しが収録されている．そこで，JAMUL の記事先頭 3 文と見出しの含意関係を判定したところ，含意関係を保持する割合は 61.4% であった．これに対して，記事全文と見出しの間の含意関係を判定すると，この数字は 94.2% に上昇する．

以上の分析結果から，Gigaword も JAMUL も見出しの情報は記事全文で概ねカバーされていることが分かる．ところが，タスク設定やコーパスの仕様により，記事の先頭 1 文や 3 文から見出しを生成することを考えると，30–40% の事例に記事から逸脱した見出しが存在する．したがって，記事の先頭 1 文や 3 文から見出しを生成するというタスク設定が，見出し生成器の忠実性に悪影響を与えている可能性がある．

3 訓練データの忠実性の改善

2節で，見出し生成の訓練データ中に記事内容から逸脱した見出しが多く存在することを明らかにした．この状況を改善するためには，2 つの戦略が考えられる．1 つは記事の先頭だけでなく，記事全文を与えること，もう 1 つは訓練データから含意していない事例を削除することである．前者の戦略を取ることが理想的ではあるが，見出し生成器に与える入力系列が長くなるため訓練コストの増加や生成される見出しの質の低下が懸念される．また，JNC は記事全文を提供しないため，この戦略を採用できない．ゆえに，本研究では後者の戦略を採用し，見出し生成データセットから非含意事例を削除する．

訓練データ中の非含意事例を見つけるため，入力文書と見出しの含意関係の認識器を構築する.近年,BERT [2] などの事前学習言語モデルが，含意関係認識タスクでも顕著な進歩を示している．そこで，事前学習済みモデルを出発点として，含意関係認識器を構築する．

英語の含意関係認識として，MultiNLI [12] でファインチューニングした事前学習済みの RoBERTa large [7] をベースとする．さらに，2節で構築した記事先頭 1 文と見出しの含意関係ラベルでモデルをファインチューニングした．具体的には，2 人以上の被験者が含意または非含意とラベル付けした記事・見出しの事例のみを用い，RoBERTa モデルをファインチューニングした（10 エポック）．記事・見出しのデータによるホールドアウト検証（訓練に 761，テストに 179 事例）を行ったところ，含意関係認識の精度は 91.7% であった．

日本語の含意関係認識器は，日本語テキストの事前学習済み BERT モデル[*3]を使用する．ただし，MultiNLI に匹敵するような意味推論用の大規模な日本語コーパスはない．そこで，JNC の記事先頭 3 文と見出しの事例 12,000 件を抽出し，クラウドソーシングで含意関係ラベルを付与した．各事例に対して，5 人の作業者によって含意関係ラベルを付与し，4 人以上が含意もしくは非含意とラベル付けした事例のみを使って，含意関係認識器を学習した．5,033 件の訓練データと 1,678 件のテストデータによるホールドアウト検証を実施したところ，含意関係認識の精度は 83.9% であった．

4 見出し生成器の忠実性の改善

本節では，訓練データから含意しない見出しを排除することで，見出し生成モデルの忠実性が改善されるので

[*3] https://github.com/yoheikikuta/bert-japanese

はないか，という仮説を検証する．

4.1 データセットの準備

Gigaword データにおける実験では，Rush ら [9] と同じデータ分割の訓練（約 380 万事例），開発（約 39 万事例），およびテスト（約 38 万事例）セットを使用する．本研究では，テストセットのうち2節で使用されなかった事例から 10,000 事例をサンプリングし，評価に用いた．数字のマスキング，稀な単語の UNK への変換，および小文字への変換等の操作は行わない．データセットは，UniLM[*4]で使用されているのと同じ語彙を使用し，WordPiece によってトークン化した．

JNC は約 170 万件の訓練事例と約 3 千件の開発事例に分割し，JAMUL データセットでモデルを評価する．トークン化には SentencePiece[*5]を使用した．

4.2 見出し生成モデル

見出し生成モデルとして Transformer アーキテクチャ [11] を採用し，その実装として fairseq[*6]を用いた．非含意事例を含むデータと含まないデータで Transformer モデルを訓練する．3節で構築した含意関係認識器を用いて，記事が見出しを含意する事例のみを訓練データとして，見出し生成モデルを学習した（フィルタ有）．ただし，訓練データから非含意事例を削除すると事例数が減ってしまうため，見出し生成の性能が低下する恐れがある．そこで，自己学習戦略を適用し，元々の訓練事例と同量の訓練事例を用意した（フィルタ有＋疑似）．具体的には，フィルタ有の設定で学習した見出し生成モデルを用い，3節で削除された記事から見出しを生成し，議事訓練データとした．これらの実験設定と，すべての訓練データを用いた場合（フィルタ無）を比較する．

4.3 評価方法

多くの先行研究に従い，full-length F1 ROUGE スコアで見出しの品質を評価する．しかし，Kryscinski ら [6] はシステム要約と参照要約の間の ROUGE スコアが人間の評価と弱い相関関係しかないことを報告している．さらに，本研究では見出しの原文に対する忠実性に焦点を当てて評価をしたい．そこで，3節で説明した含意関係認識器が原文が見出しを含意すると判定する割合（含意率）と人手による評価も報告する．

4.4 結果

表3に実験結果を示す．元々の訓練データ（フィルタ無）で訓練したベースラインモデルは，English Gigaword データセットで 35.80 ROUGE-1 スコア，JAMUL で 48.08 ROUGE-1 スコアを獲得した．含意関係認識器でフィルタリングを行うと，訓練事例数が減少したためか，両方のデータセットにおいて ROUGE スコアが低下したが，自己学習戦略（フィルタ有 + 疑似）により Gigaword データセットの ROUGE スコアは上昇し，ベースラインモデルを上回った．

対照的に，自己学習戦略は JAMUL における ROUGE スコアを改善できなかった．この原因を正確に特定することは難しいが，フィルタリングにより訓練事例が減少しすぎたため（約 80 万事例），自己学習戦略がうまく働かなかった可能性がある．別の可能性としては，JNC/JAMUL で記事と見出しが非含意となる記事の執筆スタイルが，含意する記事のものと大きく異なるため，自己学習手法が参照見出しとかけ離れた見出しを生成したことも考えられる．

表3の「含意率」列は，各モデルが生成した見出しと元記事の間の含意関係を3節で構築した含意関係認識器を用いて推定し，全評価事例の中で含意すると判定された事例の割合を表す．この評価方法では，フィルタ有 + 疑似の設定が両方のデータセットにおいて最も高い含意率を示した．訓練データの記事と見出しの間で含意関係が成立するように誘導したため，この結果は自然なものであるが，見出し生成器のモデルを一切変更することなく，この実験結果が得られたことは興味深い．

含意率による自動評価は確立された評価手法ではないため，被験者を用いた評価実験も実施した．具体的には，Gigaword と JAMUL のそれぞれの評価データからランダムにサンプルした 109 件の事例に対して，各モデルで見出しを生成し，入力文書との忠実性を「忠実」「忠実でない」「判定不能」のいずれかで判定した．表3の「人手評価」列に，忠実と判定された見出しの割合を示した．人手による忠実性の評価結果は，含意率の評価結果と整合しており，フィルタ有 + 疑似で獲得した訓練データが，入力に忠実な見出しを生成することに貢献していることが確認できた．以上の実験結果は，モデルが生成した見出しを ROUGE スコアだけで評価しても，忠実性を検証することができず，含意関係認識器や人手による評価が欠かせないことを示唆している．

5 関連研究

Rush ら [9] の研究以降，ニューラル系列変換アーキテクチャを用いた抽象型要約が盛んに研究されている．

しかし一方で，系列変換モデルに基づく抽象型要約

[*4] https://github.com/microsoft/unilm
[*5] https://github.com/google/sentencepiece
[*6] https://github.com/pytorch/fairseq

表3: テストセットでの結果．各 R は F1 full-length ROUGE スコアを表す．「含意率」は含意関係認識器（3節）が含意と予測した事例の割合，「人手評価」は人間が忠実な見出しであると判定した事例の割合を表す．

データセット	訓練データ（量）		R-1	R-2	R-L	含意率 (%)	人手評価 (%)
Gigaword	フィルタ無	(3.8 M)	35.80	17.63	33.69	85.78	77.06
	フィルタ有	(2.7 M)	35.24	17.29	33.14	91.50	—
	フィルタ有 + 疑似	(3.8 M)	**35.85**	**17.94**	**33.72**	**93.56**	**85.32**
JAMUL	フィルタ無	(1.7 M)	**48.08**	**22.21**	**40.02**	90.29	89.91
	フィルタ有	(0.8 M)	46.08	20.81	38.07	95.67	—
	フィルタ有 + 疑似	(1.7 M)	45.62	20.55	38.10	**96.26**	**92.66**

が不正確な事実を含む要約を生成するとの報告がある．Cao ら [1] は，系列変換モデルによって生成された要約の 30%には入力記事と異なる事実が含まれていると報告した．Kryscinski ら [6] は，抽象型要約において ROUGE スコアは人間の評価と弱い相関しかないため，事実に関する評価尺度として不適切であると指摘した．

ニューラル系列変換アーキテクチャや学習方法を改善することで，忠実性を改善する研究もいくつか発表されている．Cao ら [1] は情報抽出器を適用し，主語・述語・目的語のタプルを抽出し，それらをモデルの入力に追加した．Pasunuru ら [8] は含意関係認識器を抽出型要約の強化学習の報酬として用いた．Guo ら [4] は含意生成（前提から含意している仮説を生成するよう訓練する）と要約のマルチタスク学習を提案した．清野ら [5] は入力と出力のトークン間の対応をモデル化することで，間違った生成を抑制した．Falke ら [3] は含意関係認識器の予測に基づき，ビームサーチのリランキングを行った．

また，Tan ら [10] は記事の先頭の文は見出し生成の入力として適切ではないと主張し，記事に加えてその要約を入力として利用することで，見出し生成のパフォーマンスが向上することを報告した．

これらの先行研究に対して本研究は，Gigaword と JNC/JAMUL の 2 種類のデータセットに関して，タスク設定や訓練データの問題点を検証し，その改善策を講じることで，見出し生成の忠実性を向上させることを実証した．

6 おわりに

本研究は，記事内容に忠実ではない見出しが生成される原因は，見出し生成のタスク設定の不備や訓練データにあるのではないか，という仮説を検証するため，Gigaword と JNC/JAMUL のデータセットを分析し，訓練データの入力文書が見出しを含意しない事例が多く含まれることを明らかにした．見出し生成器の忠実性を向上させる方法として，入力文書が見出しを含意しない事例を訓練データから除去するアプローチを提案した．このアプローチの効果は ROUGE スコアで確認することはできなかったが，自動評価および人手評価では忠実性の顕著な改善が確認できた．今後は，忠実性の評価尺度を確立するための大規模な実験や，非含意の訓練事例を効果的に活用する手法を検討していく予定である．

謝辞 本研究成果は独立行政法人情報通信研究機構（NICT）の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」により得られたものです．

参考文献

[1] Ziqiang Cao et al. “Faithful to the Original: Fact Aware Neural Abstractive Summarization”. In: *Proc. of AAAI*. 2018, pp. 4784–4791.
[2] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proc. of NAACL-HLT*. 2019, pp. 4171–4186.
[3] Tobias Falke et al. “Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference”. In: *Proc. of ACL*. 2019, pp. 2214–2220.
[4] Han Guo, Ramakanth Pasunuru, and Mohit Bansal. “Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation”. In: *Proc. of ACL*. 2018, pp. 687–697.
[5] Shun Kiyono et al. “Reducing Odd Generation from Neural Headline Generation”. In: *Proc. of PACLIC*. 2018.
[6] Wojciech Kryściński et al. “Neural Text Summarization: A Critical Evaluation”. In: *Proc. of EMNLP-IJCNLP*. 2019, pp. 540–551.
[7] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *CoRR* abs/1907.11692 (2019).
[8] Ramakanth Pasunuru and Mohit Bansal. “Multi-Reward Reinforced Summarization with Saliency and Entailment”. In: *Proc. of NAACL-HLT*. 2018, pp. 646–653.
[9] Alexander M. Rush, Sumit Chopra, and Jason Weston. “A Neural Attention Model for Abstractive Sentence Summarization”. In: *Proc. of EMNLP*. 2015, pp. 379–389.
[10] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. “From Neural Sentence Summarization to Headline Generation: A Coarse-to-Fine Approach”. In: *Proc. of IJCAI-17*. 2017, pp. 4109–4115.
[11] Ashish Vaswani et al. “Attention is all you need”. In: *Proc. of NIPS*. 2017, pp. 5998–6008.
[12] Adina Williams, Nikita Nangia, and Samuel Bowman. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. In: *Proc. of NAACL-HLT*. 2018, pp. 1112–1122.