Vol.2019-NL-240 No.1
2019/6/13

含意関係に基づく見出し生成タスクの見直し

松丸 和樹[1,a)]　高瀬 翔[1,b)]　岡崎 直観[1,c)]

概要：見出し生成タスクでは，エンコーダ・デコーダモデルの高い性能が報告される一方で，記事内容から逸脱した見出しを生成してしまう問題が指摘されている．この原因のひとつとして，訓練データ中の記事に含まれる情報が不足しているため，記事中に書かれていない内容を無理に見出しに出力するような学習が行われていることが考えられる．そこで我々は，JApanese MUlti-Length Headline Corpus (JAMUL) の記事の先頭３文と記事全文それぞれが正解見出しを含意しているか調べた．その結果，先頭３文では 42.0%，全文でも 11.1%の事例で記事が見出しを含意していないことがわかった．続いて，Japanese News Corpus (JNC) の記事先頭３文で学習したエンコーダ・デコーダが出力した見出し文候補に対し，クラウドソーシングを用いて含意関係を付与し，51,027 件の含意関係データセットを作成した．構築したデータセットで含意関係認識器を学習し，その含意関係認識器で生成器が出力した見出しの含意割合を判定したところ，生成された見出しの多くは含意と判定されないことがわかった，さらに，含意関係認識器で訓練データをフィルタリングし，見出し生成器を学習する実験を行った．フィルタリングしない訓練データで学習した場合との比較により，訓練時の記事の情報不足が見出し生成器に悪影響を及ぼし，含意しない見出しを生成する一因になっていることが分かった．

キーワード：文書要約，見出し，含意関係認識，データセット

1. はじめに

新聞において見出しは「ショーウィンドウ」に喩えられる [1]．記事の内容を端的に表現することで読者の興味を引く．現在の新聞では，見出しが紙面の約３割を占めると言われており [2]，新聞の見出しだけを読んで幅広い情報を収集する「見出し読者」がいるほど，見出しが果たす役割は大きい．また，デジタルニュース配信やソーシャルネットワーキングサービス（SNS）が普及した現代では，記事のメール配信，SNS 上での記事のシェア，Open Graph Protcol (OGP)[*1] などで見出しが用いられている．米コロンビア大学とフランス国立情報学自動制御研究所（INRIA）の研究によると，Twitter でシェアされた URL のうち 59%は全くクリックされず，フォロワーがシェアされた URL を訪れる確率は 0.001%から 0.1%程度と推測されている [3]．見出しは記事へのアクセス数を左右するだけでなく，記事の代わりに流通するほど，重要かつ身近なものになった．

表 1　不適切な見出し生成の例.

記事	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

近年，新聞記事の本文から見出しを自動的に生成する研究，すなわち**見出し生成**が盛んに行われている．機械翻訳研究で sequence-to-sequence (seq2seq) や注意機構などのエンコーダ・デコーダモデルが発展したことを受け，新聞記事の本文を入力，見出しを出力と見なし，入力・出力の組を訓練データとして大量に与え，見出し生成のモデルを学習する手法が主流となった [4], [5]．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学習できる．

一方で，自動生成された見出しが元記事の内容から逸脱してしまう問題が報告されている [6], [7]．これは，報道という応用を考えたとき，深刻な問題を引き起こす．例えば，**表 1** では「あす投開票」という記述を含む見出しが生成されているが，正しくは「きょう投開票」であるので，国民主権の根幹を揺るがす大問題となる．ただ，人間が表 1 の記事本文を読んでも，投開票の日が今日なのか明日なの

[1] 東京工業大学情報理工学院
[a)] kazuki.matsumaru@nlp.c.titech.ac.jp
[b)] sho.takase@nlp.c.titech.ac.jp
[c)] okazaki@c.titech.ac.jp

[*1] Twitter や Facebook などの SNS 上でウェブサイトの URL を投稿したときに，そのウェブサイトのタイトルや概要，画像などを表示する仕様.

か判断できない．もし，見出し生成器の訓練データに入力（記事）から逸脱した出力（見出し）の事例が多く含まれるのであれば，入力に含まれない情報を出力するような「無理な」学習を見出し生成モデルに強いることになる．

本研究では，記事内容から逸脱した見出しが生成される原因の一つは，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．朝日新聞社から公開[*2]されている Japanese News Corpus (JNC) と JApanese MUlti-Length Headline Corpus (JAMUL) [8] を分析し，日本語の見出し生成の訓練データや評価データの見出しの中に，記事内容から逸脱したものがどの程度存在するのか調査する．その結果，JAMUL の 42.0%（記事先頭 3 文を入力としたとき）および 11.1%（記事全文を入力としたとき）に，記事内容から逸脱した見出しが存在することを報告する．

続いて，記事内容から逸脱した見出しを訓練データから除外したり，モデルが生成した見出しの中で記事内容から逸脱していないものを選び出すことで，記事内容からの逸脱を軽減できるのではないか，という仮説を検証する．この仮説の検証には，見出しが記事から逸脱しているか否か判定する処理が必要である．そこで，記事が見出しを含意するかどうかを推論する処理を含意関係認識と捉え，記事と見出しの組（51,027 件）に含意関係ラベルを付与したデータセットを構築した．これは，日本語の含意関係認識のデータセットとしては，過去最大の規模である．また，JNC の入力である記事の先頭 3 文は 49.9%しか見出しを含意していないことが判明した．ゆえに，JNC で学習した見出し生成器は記事内容から逸脱した見出しを生成するように誘導されている可能性がある．このデータセットを訓練データとして用い，含意関係認識器 [9] を学習したところ，その正解率は 79.5%であった．

最後に，構築した含意関係認識器を用い，JNC の訓練データから見出しを含意しない記事を除去（フィルタリング）し，記事内容から逸脱しない見出し生成器を学習する実験を行う．そして，このフィルタリング処理によるものと，JNC からランダムにサンプリングして訓練データ数を揃えた場合と比較する．その結果，フィルタリング処理を導入したモデルのほうが，含意する見出しを生成する傾向が見られることを報告する．

2. 見出しデータセットの調査

2.1 日本語見出しデータセット

まず，本研究で用いるデータセットを紹介する．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) が公開されている．JNC は朝日新聞の記事と紙面見出しのペア 1,831,812 件を収録したコーパスであり，記事は先頭から 3 文のみが収録されている．また，記事に対して異なる長さの見出しを収録し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．本研究では，JNC と JAMUL の紙面見出しを使って実験を行う．

[*2] https://cl.asahi.com/api_data/jnc-jamul.html

1. 1つの記事に付与されている 1 つの見出しに対し、記事が見出しを含意しているかどうかを判定する。
2. 各判定は以下の 3 つのうちのどれかとする。
   - A. 含意する
     - 記事の内容が見出しの内容を含んでいる場合。
     - 見出しが記事にない表現を使っていても、その表現の内容が記事から導き出される場合は「含意する」とみなす。
   - B. 含意しない
     - 見出しが記事と相反することを述べている場合。
     - 見出しが記事から確認できない情報、記事にない情報を述べている場合。
   - C. その他
     - 見出しが文法的に日本語として不完全で、判定のしようがない場合。
     - 全て人間が書いた見出しなので基本的にこの判定になることはないが、文字化けなどが発生していた場合はこれを選ぶ。

図 1 含意ラベリングのガイドライン

表 2 ワーカーに提示した例（青字は言い換え表現，赤字は逸脱している部分）

記事（一部）	見出し
含意する例	
派遣社員を雇い止めする「派遣切り」が今年、多発する可能性がある。	「雇い止め」急増の恐れ
…25 億 8600 万枚…	…25 億枚…
含意しない例	
鹿児島市で地場食材を使った料理教室があった。市内から 25 人が参加。市内五つの農産加工グループの 10 人が講師を務めた。	地場食材で新メニュー 鹿児島の加工グループが料理教室
井の頭池でボートに乗ったカップルは別れる-。都立井の頭公園で最も有名な都市伝説だ。いやいや、恐れるなかれ。	井の頭　愛を誓う　「お別れ伝説」なんのその
22 日に…廃炉を決める	きょう廃炉決定

2.2 記事が見出しを含意している割合

2.1 節で説明した JNC と JAMUL について，見出しが記事本文から逸脱しているか調査を行う．この調査を行う理由は，見出し生成器が記事から逸脱した見出しを出力するのは，訓練データにも記事から逸脱した見出しが多く含まれることが原因である，という仮説を検証するためである．

本研究では，見出しが記事内容から逸脱しているか否か，という問題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．見出し生成に当てはめると，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）を含意しているか否かのラベルを付与する．

最初に，JAMUL の記事全文が見出しを含意しているか，検証を行った．**図 1** および**表 2** に，作業者に提示したガイドラインと，例示した事例をそれぞれ示した．今回は，JAMUL の記事全文と見出し 1,000 件について 3 人の作業者に含意関係のラベル付けを行ってもらい，2 人以上が「含意する」と判定した事例を含意，2 人以上が「含意しない」と判定した事例を非含意とし，それ以外の事例[*3]は除去した．その結果，816 件が含意，102 件が非含意とラベル付けされた．ゆえに，JAMUL の記事の 11.1%では，見出しが記事を含意しないことが分かった．

含意しないと判定された 102 件について，その原因を調査した．なお，今回のアノテーションガイドラインでは，記事と見出し以外の情報（例えば記事が書かれた日付など）は利用しないこととした．

- JAMUL の誤り（52 件）：記事と見出しの対応付けが明らかに間違っているもの（朝日新聞デジタルで検索すると見出しが別の記事のものになっている）．
- 日付（31 件）：記事中では日付で述べられているが見出しでは「きょう」「来年度」などに言い換えられているもの．
- 注釈付き（5 件）：見出しの最後に「続報注意」や「訂正・おわびあり」と書かれているもの
- 湧き出し（5 件）：記事中で言及されていない情報が見出しに含まれるもの
- その他（7 件）

このうち，JAMUL の誤りはコーパスが構築されたときのバグであるため，この問題は将来的に修正される見通しである．したがって，この 52 件を除けば，JAMUL の見出しの約 95%は対応する記事を含意すると言える．

ところが，JAMUL に収録されている記事の数は少ないため，見出し生成器の学習には JAMUL が使えず，JNC を用いることになる．JNC コーパスでは，全ての記事はその先頭 3 文しか収録されていない．したがって，元々の記事は見出しの情報をカバーしていたとしても，JNC コーパスを訓練データとして用いる場合は，見出し生成器の入力である記事に情報の欠損があることに注意が必要である．

そこで，JAMUL の先頭 3 文についても同様に記事が見出しを含意しているか調査したところ，42.0%の見出しが記事から逸脱していることが判明した．従って，記事の先頭 3 文から見出しを生成するというタスク設定は，記事から逸脱した見出しの生成を助長している可能性がある．

3. 含意関係データセットの作成

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がわかった．したがって，記事内容から逸脱した見出しを訓練データから除外したり，モデルが生成した見出しの中で記事内容から逸脱していないものを選び出すことで，記事内容からの逸脱を軽減できるのではないか，という仮説が立てられる．この仮説の検証には，見出しが記事から逸脱しているか否か自動的に判定する処理が必要である．

そこで，記事が見出しを含意するかを推論する処理を含意関係認識としてとらえ，記事と見出しの組に含意関係ラベルを付与したデータセットを作成し，このデータセットを訓練データとすることで含意関係認識器を構築する．この含意関係データセットは，新聞社が執筆した記事本文と見出しに加え，SWAG データセット [10] のように自動的に生成した見出しから構成される．これには二つの理由がある．一つ目は，ラベリング作業の効率化である．記事と見出しの含意関係を判定するときは，両方の文章を読む必要があるが，記事の方が文章量が多いので，記事を読む時間が相対的に多くなる．もし，1 記事につき複数の見出しがまとめて提示され，見出しごとに含意関係のラベルを付与することができれば，記事を読む件数を減らすことができ，データ作成のコストを削減できる．二つ目は，生成された見出しの含意割合を調べたり，見出し生成器の出力候補を含意関係認識器のスコアによって入れ替えたり，Pasunuru ら [11] の研究のように生成器が出力した見出しに対する含意関係認識器のスコアに基づき，強化学習を行うアプローチへの利用も想定しているためである．

含意関係データセット構築の全体の流れを**図 2** に示す．まず，3.1 節で含意関係データセットに収録する見出しを生成するためのモデルについて説明する．次に，3.2 節でデータセットの仮説文の生成方法を説明する．3.3 節では，クラウドソーシングによる見出し文のラベリングについて述べる．最後に，構築したデータセットで含意関係認識器を学習する（3.4 節）．

3.1 見出し生成モデル

含意関係データセットの仮説文として利用する見出し生成モデルには，機械翻訳タスクで最高性能を達成している Transformer [12] を使用する．実装は fairseq[*4]を用いた．トークン，長さ埋め込み，隠れ層は 512 次元とし，エンコーダ，デコーダは共に 6 層とした．Attention Head 数を 8，順伝搬層を 2048 次元，Adam の初期学習率を 0.0005，β_2

[*3] 文字化け等が発生して正確に判定されなかったもの．

[*4] https://github.com/pytorch/fairseq

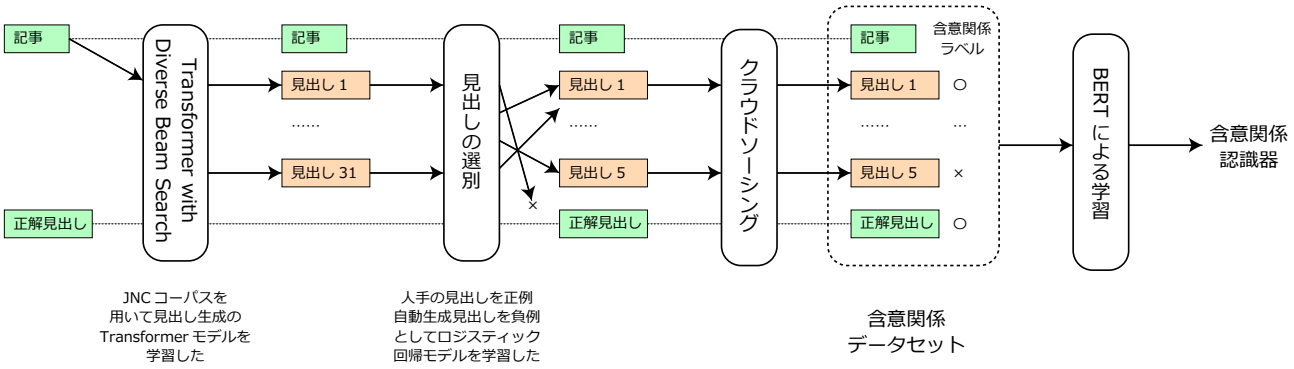

図 2 含意関係データセット構築の概要

を 0.98，Warming up を 4000 ステップ，Label smoothing の ϵ を 0.1，Dropout の確率を 0.3 に設定した．JNC の約 180 万件の記事と見出しの組から，3,000 件を開発データ，100,000 組をテストデータとして取り除き，残りを訓練データとして用いた．この訓練データの記事（先頭 3 文）を入力，見出しを出力として見出し生成モデルを訓練した．

3.2 仮説文の生成

前節で述べた見出し生成モデルによって，データセットの仮説文を生成する．Transformer を含むエンコーダ・デコーダモデルでは，文を出力する際にビームサーチが用いられる．ビームサーチが保持する出力文の候補をデータセットの仮説文として利用したいが，ビームサーチの候補同士の差異は助詞の変更などに留まり，多様性に欠けることが多い．そこで，本研究では Diverse Beam Search (DBS) [13] を利用した．DBS はビームサーチを拡張した手法で，各ステップで n 番目のビームのスコア計算時に $1...n-1$ 番目までのビームとの異なり度を加算する多様度関数（diversity function）を追加する．多様度関数は系列 y と系列の集合 Y の異なり度を測定する任意の関数である．本研究では多様度関数にハミング距離を使用した．なお，DBS にはグループという概念が存在するが，本研究ではグループ数とビーム幅を同じに設定している（つまり，グループが存在しない）ため，割愛する．多様度関数にかかる係数は 0.5 とした．テストデータから 12,000 記事を選び，3.1 節で述べた見出し生成モデル，および DBS を適用し，各記事に対して 31 種類の見出し候補を生成した．

次に，含意関係データセットに明らかに不自然な見出しが混入することを防ぐため，見出しの「自然さ」をスコア付けするモデルを用意する．具体的には，人間が書いた見出しを正例，機械が生成した見出しを負例として疑似的に訓練データを構築し，ロジスティック回帰モデルを学習する．ロジスティック回帰の素性には見出しとそれに対応する記事の unigram，bigram，trigram，さらにそれらを見出しの長さで割ったもの，見出しのパープレキシティを使う．パープレキシティの算出には訓練データの見出し文で学習した言語モデルを使用する．言語モデルには Transformer のデコーダ部分を使用する [14]．

ある記事に対して，見出し生成モデルが出力した 31 件の見出しのうち，最も DBS のスコアが高かった 1 文を選出する．残りの 30 件については，ロジスティック回帰モデルが出力した確率が高かった 4 件を選び，自動生成された 5 件の見出しを選んだ．これに加え，新聞社が実際に配信した見出しを合わせて，1 記事に対して 6 件の見出しを含意関係データセットの仮説とした．

3.3 クラウドソーシングによる含意ラベルの付与

3.2 節までの準備で，1 記事に対して 6 件の見出しを獲得し，記事が全部で 12,000 件あるので，トータルでは 72,000 件の見出しを収録したデータセットを構築した．続いて，記事と見出しの組について，記事が見出しを含意しているか否かのラベルをクラウドソーシングで付与した．クラウドソーシングのプラットフォームには『Yahoo! クラウドソーシング』[*5]を用いた．**図 3** にタスクの概要を示した．

作業者には一度に一つの記事と六つの見出しを提示し，記事内容から逸脱していないと判断した見出し文に全てチェックを付けてもらった．品質保持のため，チェック設問によって信頼できない解答をフィルタリングしながら，ひとつの記事につき 5 人の作業者からの解答を集め，見出し文は記事の内容から逸脱していないと 4 人以上が判定したものを「含意」，1 人以下しか判定しなかったものを「非含意」，それ以外を「不明瞭」とした．

この手順に従い，72,000 件の見出し文と記事の組について，含意関係ラベルを付与した．**図 4** に，5 人のワーカーのうち含意していると判定した人数毎に，見出しの件数を示した．4 人以上の票を得た見出しを含意，1 人以下だったものを非含意と見なすと，19,114 件（約 37.5%）の含意，31,913 件（62.5%）の非含意のラベルが付与された含意関係データセットを構築したことになる．このデータセット

[*5] https://crowdsourcing.yahoo.co.jp/

記事とその見出し文の候補を読んで、各見出し文が記事で書かれている内容から逸脱していないかをチェックするタスクです。

- 例
  以下の例では、4 つの見出しのうち、1 つ目と 3 つ目の見出しが記事の内容から逸脱していないのでチェックすることになります。
  記事：
  横浜市立平戸中学 (戸塚区) の 2 年生 4 人が 11 日、職業体験のため朝日新聞横浜総局を訪れた。校閲をしたり、記事をインターネットに載せたりして新聞記者の仕事を学んだ。また、近くの日本新聞博物館で、新聞製作に取り組んだ。
  見出し：
  (1) 平戸中 2 年 4 人、記者の仕事体験　朝日新聞横浜総局
  (2) 新聞記者が職業体験　朝日新聞横浜総局
  (3) 中学生 4 人、新聞製作を体験　朝日新聞横浜総局を訪問
  (4) 女子中学生が新聞製作　朝日新聞横浜総局
- チェックすべき見出し
  記事に書かれていることから逸脱していない見出しをチェックしてください。「訪れた」→「訪問」のような類似した意味の言い換えは逸脱とはみなさずチェックしてください。
- チェックしてはいけない見出し
  – 見出しが記事と相反することを述べているとき（例：「新聞記者が職業体験　朝日新聞横浜総局」）
  – 見出しが記事から確認できない事柄を述べているとき（例：「女子中学生が新聞製作　朝日新聞横浜総局」）
- 注意事項
  全ての選択肢をチェックすることになる設問や、ひとつもチェックすべき選択肢がない設問もあります。チェックすべき選択肢がない場合は「該当なし」を選んでください。

図 3　ラベル付けタスクの説明文

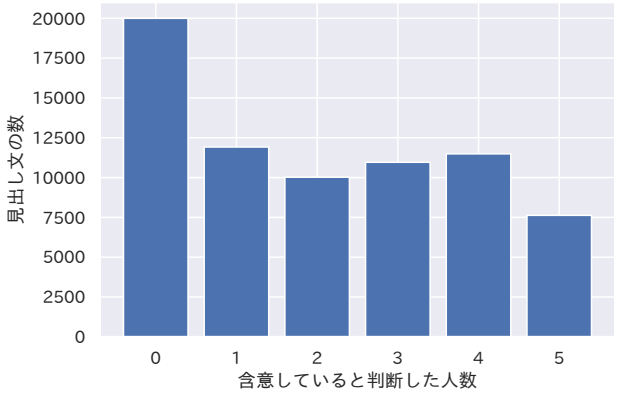

図 4　見出し文の含意判定の一致度の分布

のうち，3,021 件を開発データ，2,955 件をテストデータとし，残りの 45,051 件を訓練データとした．分割する際には，記事ごとに分割を行うことで，同一の記事が異なる分割に混入することが無いように配慮した．

なお，このデータセットを構築したことにより，JNC の記事の先頭 3 文が人間（新聞社）が書いた見出しを含意している割合を測定することができる．人間が書いた見出し 12,000 件のうち，最終的に「含意」とラベル付けされた事例は 4,199 件，「非含意」とラベル付けされた事例は 4,190 件であった．JAMUL での分析と同様に，半分くらいの見出しは記事の先頭 3 文以外の情報を使って書かれていることになり，見出し生成器の訓練データとして，記事の先頭 3 文にトリミングしてしまうと，記事内容から逸脱した見出しを生成するように誘導されてしまう恐れがある．

3.4　含意関係認識器

3 節で作成したデータセットを用いて，含意関係認識器の学習を行う．含意関係認識モデルには，BERT [15] を使用する．BERT は最初に大規模な生コーパスで事前学習したのちに，各タスクで fine-tuning するモデルである．fine-tune 時の BERT への入力は 1 文，文ペアもしくは文書であり，今回は含意関係認識タスクへの適用なので，入力には文ペア（記事と見出し）を与える．本研究では，事前学習済みのモデルとして日本語 Wikipedia を SentencePiece でトークン化して事前学習したもの [9] を利用する．fine-tuning 時には，記事が見出しを含意しているかどうかの二値分類を行うための全結合層を導入し，作成した含意関係データセットの訓練データで学習した．fine-tune 後にテストセットで含意関係認識の性能を計測したところ，正解率は 79.5%であった．

4.　記事から逸脱しない見出しの生成実験

本節では，まず通常の訓練データで学習した見出し生成器がどれくらい逸脱した見出しを生成しているかについて議論する（4.1 節）．次に，3.4 節で学習した含意関係認識器で JNC の訓練データから見出しを含意しない記事を除去（フィルタリング）し，フィルタリングされた訓練データでの見出し生成器の学習を行う（4.2 節）．そして，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器と，フィルタリングを行わずに学習したモデルを含意関係認識器の判定から比較し，後者よりも前者が生成する見出しの方が，含意関係認識器に含意と判定される割合が向上することを確認する．また，学習データの量を揃えた場合，後者よりも前者の見出し生成器の方が高い ROUGE スコアを示すことを報告する．

4.1　生成された見出しの含意割合

英語の大規模見出し生成データセットとして，Gigaword コーパスが有名である．Cao ら [6] は注意機構付きの Sequence-to-Sequence モデルとして学習された見出し生成器を，Gigaword コーパスのテストセットに適用し，その出力からランダムに 100 事例をサンプリングして人手評価したところ，記事に「忠実な」生成見出しが 68 例しか無かったことを報告している．

JNC で訓練したモデルでも同様の問題が発生するか確認するため，3.4 節で訓練した含意関係認識器を用いて記事と

生成された見出しの含意関係を推定した．その結果，JNC のテストデータのうち 3,000 件の記事から生成された見出しの 33.1%，JAMUL データセットの記事から生成された見出しのうち 35.2%は記事を含意しないと判定された．

4.2 フィルタリングした訓練データによる見出し生成

いよいよ本研究のメインの仮説 —見出し生成器が 4.1 節のように記事で述べられていない事柄を述べてしまう原因のひとつは，訓練データ中の記事に含まれる情報が不足しているため，記事中に書かれていない内容を無理に見出しに出力するような学習が行われている— を検証する．

JNC の訓練データの事例に含意関係認識器を適用し，記事と見出しの含意確率が 0.5 を上回った事例のみを抽出する．含意関係認識器には 3.4 節で構築したものを用いる．このフィルタリング処理により，1,728,812 組の訓練データが 659,602 件に減少した．このスクリーニングした訓練データを用いて見出し生成器を学習する．3.1 節と同様に，見出し生成器には Transformer を使用する．比較対象として，通常の訓練データで学習した見出し生成器と 659,602 件（フィルタリング後と同量）の事例をランダムにサンプルして学習した見出し生成器を用いた．テストデータには JAMUL の記事の先頭 3 文と紙面見出しを採用する．

表 3 に実験結果を示す．なお，含意割合は元記事と生成された見出しについて，3.4 節で構築した含意関係認識器が含意として判定した割合である．表 3 によると，訓練データに含意関係認識によるフィルタリングを適用した場合 (1) に，含意割合が最も高くなった（75.7%）．見出し生成の訓練データを見出しと記事の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ていることは興味深い．また，フィルタリングにより学習データの事例数が元々の 38.2%まで減少したため，フィルタリングを行わない設定 (3) と比較すると ROUGE スコアの低下がみられるが，学習データを同量に揃えた場合 (2) よりは高い ROUGE スコアを維持している．

この実験結果により，見出し生成器の訓練データを精錬することにより，記事から逸脱した見出しを抑制する可能性が示された．ただ，(1) と (3) の見出しのどちらの方が良いのか？　これを調査するには，見出し生成の応用先を考慮しながら，生成された見出しの人手で評価する必要がある．ROUGE スコアの数ポイントの差が読者にとって顕著な差があるのか，ROUGE スコアで記事と見出しの含意関係を成否を評価するのには無理があるのではないか，BERT で構築した含意関係認識器に正確性や頑健性がどのくらいあるのか，さらなる検証を進める必要がある．

5. 関連研究

エンコーダ・デコーダが記事と無関係な単語を見出し中で生成する問題について言及している研究としては，清野ら [16] のものがある．ただし，清野らの研究は逸脱した見出しの生成を抑制する手法の提案であり，逸脱した見出しが生成される原因については調査していない．

抽象的要約文生成において，生成文が記事と異なる事実を述べてしまう問題全般の改善について扱った研究としては，記事と生成された要約の含意スコアを含意関係認識器で算出し，それを強化学習の報酬として利用したもの [11] や，要約文生成と含意文生成のマルチタスク学習を行うもの [17]，元記事の主語，述語，目的語を予め抽出し，それらを出力に含めるように制約を課す（具体的にはそれらをエンコーダモデルの入力に追加する）もの [6] などが挙げられる．

含意関係認識のデータセットとして広く使われている英語の大規模データセットとして，Stanford Natural Language Inference (SNLI) [18] や Multi-Genre Natural Language Inference (MultiNLI) [19] がある．これらのデータセットでは前提文と仮設文のペアが与えられ，その 2 文の関係を含意，中立，矛盾の 3 値に分類するタスクとなっている．含意関係認識器のスコアを強化学習の報酬とする研究 [11] では含意関係認識器を SNLI や MultiNLI で学習しているが，これらのデータセットのドメインはニュース記事とは大きく異なるため，その効果には疑問が残る．特に，これらのデータセットは前提文も 1 文であることは見出し生成への適用の障壁となり得る．実際，Pasunuru ら [11] は，含意関係認識器に記事を前提文，生成要約を仮説文として入力することも試みたが，うまく機能しなかったので前提文に正解見出しを使ったと論文で報告している．

SNLI やその亜種は，前提文に画像のキャプションなどを使い，仮説文は人手によって作成し，前提文と仮設文のペアを改めてクラウドソーシングでラベル付けしている．これに対して，SWAG [10] は時刻 t の動画のキャプションを前提文として，時刻 $t+1$ の文を言語モデルによって生成し，安価に大量のデータを作成することに成功した．しかし，SWAG では基本的に人間が書いた動画のキャプションが正例，機械が生成した事例が負例となるため，機械が生成した文がどれか判別できてしまうと容易にタスクが解けてしまう，そこで，SWAG では本来のキャプションと機械が生成したキャプションを簡単な素性で識別する識別器を学習し，識別器が見分けられなかった事例について，クラウドソーシングでアノテーションしている．

日本語の含意関係データセットには，小谷ら [20] によるデータセットや，RITE-2 [21] などがあるが，これらのデータセットは評価用に構築されており，モデルを学習する用途には向かない．

6. おわりに

本研究では，まず既存の見出し生成の訓練データやタス

表 3 フィルタリングの有無による見出し生成の性能の比較

訓練データ	訓練事例数	ROUGE-1	ROUGE-2	ROUGE-L	含意割合
（1） 含意関係認識器によるフィルタリング	659,602	43.9	19.0	36.5	75.7%
（2）（1）と同量の訓練データ	659,602	41.6	16.9	34.8	49.8%
（3）通常の訓練データ	1,728,812	46.3	20.5	38.1	64.8%

ク設定の適切さを分析し，何も対処を行わない場合は，見出しに対して記事中に含まれる情報が不足していることを確認した．

次に，タスク設定が逸脱した見出しを生成する原因になっているという仮説の検証のために，含意関係データセットを作成した．このデータセットの際には，多様性のある見出しを自動生成し，その中で自然な見出しを上手くサンプリングし，クラウドソーシングで含意関係のラベル付けを行うなどの工夫を導入した．構築したデータセットで含意関係認識器を学習し，これを生成された見出しの検証に用いることで，記事を含意しない見出しが生成されていることを確認した．

そして，逸脱した見出しの生成を抑えるためにデータセットをフィルタリングし，含意関係が成立している事例だけで学習する実験を行った．その結果，全ての訓練データを使ったモデルよりも見出しが記事を含意する割合が向上することが分かった．より正確な方法で見出しの記事からの逸脱を計測することは今後の課題としたい．フィルタリングにより訓練事例が減少することで，ROUGE スコアの低下が起こるが，ランダムにサンプルした同量の訓練事例で学習した場合よりは，高い ROUGE 値を示すことも確認した．

今後は，記事から逸脱した見出しの抑制に向けて，さらなる検証を進めるとともに，逸脱した見出しを抑制するような見出し生成モデルやデコーディング手法を探求したいと考えている．

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] Paul, P.: *Handbook Of Print Journalism*, Lulu.com (2014).
[2] 奥 武則：見出しの誕生：新聞の視覚媒体的要素についての 一断章，社会志林，Vol. 55, No. 1, pp. 1–17 (2008).
[3] Gabielkov, M., Ramachandran, A., Chaintreau, A. and Legout, A.: Social Clicks: What and Who Gets Read on Twitter?, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 44, No. 1, pp. 179–192 (2016).
[4] Rush, A. M., Chopra, S. and Weston, J.: A Neural Attention Model for Abstractive Sentence Summarization, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 379–389 (2015).
[5] Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C. and Xiang, B.: Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290 (2016).
[6] Cao, Z., Wei, F., Li, W. and Li, S.: Faithful to the Original: Fact Aware Neural Abstractive Summarization, *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 4784–4791 (2018).
[7] 田川裕輝，嶋田和孝：スポーツ要約生成におけるテンプレート型手法とニューラル型手法の提案と比較，自然言語処理， Vol. 25, No. 4, pp. 357–391 (2018).
[8] 人見雄太，田口雄哉，田森秀明，菊田 洸，西鳥羽二郎，岡崎直観，乾健太郎，奥村 学：出力長制御を考慮した見出し生成モデルのための大規模コーパス，言語処理学会第 25 回年次大会，pp. 1225–1228 (2019).
[9] Kikuta, Y.: BERT Pretrained model Trained On Japanese Wikipedia Articles, `https://github.com/yoheikikuta/bert-japanese` (2019).
[10] Zellers, R., Bisk, Y., Schwartz, R. and Choi, Y.: SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 93–104 (2018).
[11] Pasunuru, R. and Bansal, M.: Multi-Reward Reinforced Summarization with Saliency and Entailment, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 646–653 (2018).
[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017).
[13] Vijayakumar, A., Cogswell, M., Selvaraju, R., Sun, Q., Lee, S., Crandall, D. and Batra, D.: Diverse Beam Search for Improved Description of Complex Scenes, pp. 7371–7379 (2018).
[14] Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L. and Shazeer, N.: Generating Wikipedia by Summarizing Long Sequences, *International Conference on Learning Representations (ICLR)* (2018).
[15] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, p. (to appear) (2019).
[16] 清野 舜，高瀬 翔，鈴木 潤，岡崎直観，乾健太郎，永田昌明：ニューラルヘッドライン生成における誤生成問題の改善，言語処理学会第 24 回年次大会 (NLP2018)，p. A11 (2018).
[17] Guo, H., Pasunuru, R. and Bansal, M.: Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 687–697 (2018).
[18] Bowman, S. R., Angeli, G., Potts, C. and Manning, C. D.: A large annotated corpus for learning natural language inference, *Proceedings of the 2015 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pp. 632–642 (2015).

[19] Williams, A., Nangia, N. and Bowman, S.: A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1112–1122 (2018).

[20] 小谷通隆, 柴田知秀, 中田貴之, 黒橋禎夫：日本語 Textual Entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識，言語処理学会第 14 回年次大会，pp. 1140–1143 (2008).

[21] Watanabe, Y., Miyao, Y., Mizuno, J., Shibata, T., Kanayama, H., Lee, C.-W., Lin, C.-J., Shi, S., Mitamura, T., Kando, N., Shima, H. and Takeda, K.: Overview of the Recognizing Inference in Text (RITE-2) at NTCIR-10, *Proceedings of the 10th NTCIR Conference* (2013).