DEIM Forum 2023 X99-99

スタンスに合わせたニュースタイトルの自動生成による ニュース記事本文へのアクセスを促す情報提示

江原　駿介†　　荘司　慶行†　　山本　岳洋††　　Martin J. Dürst†

† 青山学院大学 理工学部　〒 252 – 5258 神奈川県 相模原市 中央区 淵野辺
†† 兵庫県立大学 社会情報科学部　〒 651 – 2197 兵庫県 神戸市 西区 学園西町
E-mail: †ebara@sw.it.aoyama.ac.jp, ††{shoji,duerst}@it.aoyama.ac.jp, †††t.yamamoto@sis.u-hyogo.ac.jp

あらまし　本論文では，生成的言語モデルを用いて，ニュース記事を読者のスタンスに合わせて書き換える手法を提案する．同じ内容のニュースでも，タイトルが「新型ワクチンは 1 週間で 90%の予防効果」か「1 週間もかかるのに一割の人には無効」であるかによって，そのニュースをクリックしたくなるかが変わる．このように，トピックに肯定的な人にも否定的な人にも，同じニュースを読ませるために，個人のスタンスに合わせたニュースタイトルを生成する．実際に，大規模なニュースコーパスを用いて，記事本文と主題語を与えると，主題語に応じたタイトルを生成する GPT-2 言語モデルを学習した．また，生成されたタイトルについて，元のニュース記事の内容をどれだけ正しく含むかも考慮し，もっともらしい順にランキング可能にした．自由なスタンスで生成可能になったタイトルについて，その精度や有用性について，被験者実験で評価した．実験結果から，事前学習の重要性や，関連度に基づくランキングの重要性が明らかになると思われる．

キーワード　GPT-2，情報要約, 認知バイアス

1　はじめに

近年，インターネットや情報端末の普及に伴い，老若男女問わず，SNS（Social Networking Service）に触れる機会が増加してきている．総務省の調査によると，2022 年現在，日本国民の約半数（48.6 パーセント）が SNS を使用している．[1]．このように誰もが使うようになってきている SNS において，表示される情報を個人化することは，すでに一般的に行われている．例えば，個人にとって不快なメッセージを表示しないようにしたり，好きそうなトピックの投稿を目につきやすくしたりするようなパーソナライゼーションが，当たり前のように行われている．加えて，その人の興味のありそうな投稿やユーザを推薦するような，レコメンデーション機能が一般的に実用化されてきている．これらの機能は，投稿や広告など，様々な側面で，SNS 利用者に届く情報を操作している．

このような個人化技術は，個人の SNS 利用の満足度を上げている反面,「フィルターバブル」や「エコーチェンバー」といった，認知的，社会的課題を発生させる側面も併せ持っている．Pariser らは，過度な個人化が世間一般から隔絶する現象である「フィルターバブル」の危険性を指摘している [1]．アルゴリズムがユーザデータを分析し学習すると，個々のユーザにパーソナライズされた情報が優先的に表示されるようになる．これによって，利用者は価値観や観念に合わない情報から隔離され，自身の考え方や価値観の「バブル」の中に孤立すると言われる．

また，SNS においては必然的に同じような考え方を持つ人とのつながりが深まりやすい．そのため，意見を投稿すると，似たような意見が返ってくるという「エコーチェンバー」現象が発生することも指摘されている．これらの現象によって，SNS 利用者が個人の考え方をより孤立的に，極端にする事が危惧されている．Cinus ら [2] の研究によれば，これらの問題はレコメンデーション機能などによって顕著になると指摘されている．

SNS に代表されるように，インターネット上に一般的な利用者が意見交換できる場が増え，様々なトピックに関する議論が昼夜行われている．こういった気軽に参加可能な議論は，利用者に新たな知見や発見などをもたらす一方で，気軽であるがゆえに，個人の思い込みや間違った情報を発信させやすくする弊害ももっている．特に，議論の際に，根拠を提示することなく，ただ個人の主義信条などに基づいた，議論とも呼べない無意味な対話を展開している様が散見される．このような無意味な対話を生み出す原因の一つに，情報の偏りが挙げられる，現代では，個人がインターネット上で情報を収集する際に，個人化や推薦の弊害によって，個人の確証バイアスを満足させる極端な情報を好んで入手する場合がある．そのために，個人の発信する,「裏の取れていない（信頼性の低い）」情報を，信じやすくなっている可能性がある．

このような現象に関係する筆頭著者の個人的な体験として，新しいワクチンについて情報収集をしていた際のことを例として挙げる．ワクチンの賛成派も否定派も，同様に好きなように SNS 上に個人の意見を発信していた．この際，ワクチンに肯定的な意見を持つ人が良く引用しているニュース記事と，否定的な意見を持っている人が良く引用するニュース記事につい

1：普段利用しているインターネットサービス：
令和 3 年版情報通信白書，第 1 部，第 1 章，第 1 節：
https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/pdf/n1100000.pdf

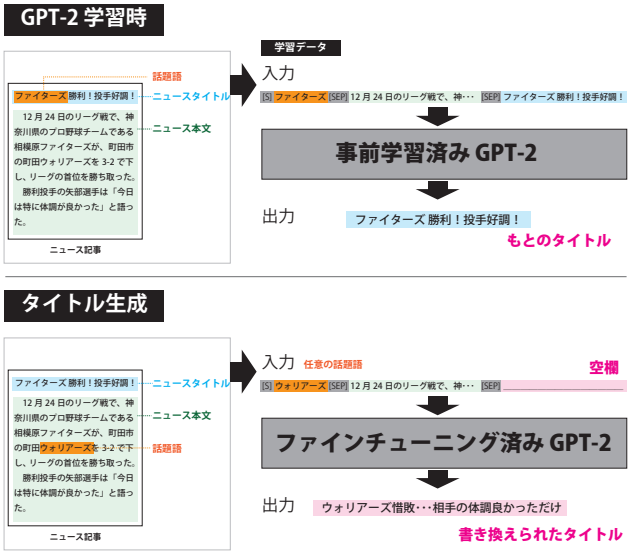

図 1　GPT-2 のファインチューニングによる実際のスタンスを考慮するニュースタイトル生成の例．学習時は「主題語と，本文の後に，タイトルが書かれる」というルールで学習させ，推論時は主題語と本文だけを与えて，タイトルを生成させる．

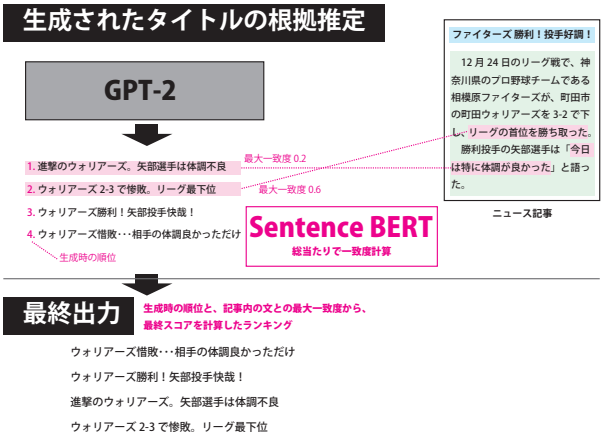

図 2　Sentence BERT を用いた生成された各タイトルの根拠となる記述部位の推定．生成時の順位と，記事内の文との最大一致度から，タイトルをランキングし出力する．

て，実際に記事の内容をそれぞれ読んでみた．すると，記事に書かれていることには，書きぶりにはスタンスにもとづく差があったが，含まれている情報自体は同一な場合が多かった．つまり,「ワクチンを打ってから抗体が得られるまでには 1 週間かかる」,「ワクチンは 9 割の人に効果がある」,「ワクチンは感染を完全に予防するものではない」という 3 つの事実があったとする．この場合，肯定的な人たちは「たった 1 週間で，ほとんどの人が抗体を得られる」と偏った受け取り方ができるし，否定的な人たちは「1 週間待っても，感染は防げないし，1 割の人にはまったく効かない」いう受け取り方ができる．これにより，人々が情報を選ぶ場合にその内容そのものではなく，それをどのように提示しているかが重要だと考えた．

こうした偏った情報の受け取り方に関する，人々のバイアスに関して，様々なアプローチで研究が行われてきている．そのうちの多くは，バイアスそのものをどうにかして軽減することを重視している．例として，Quentin [3] らは，推薦がどのようにしてユーザの行動に影響を及ぼし，フィルターバブルを形成するのかを分析している．更にその後に，独自のモデルを実装することによってフィルターバブルの効果を軽減しようとしている．しかし，ここで我々は，確証バイアスによってもたらされるこれらの問題は，人間が後天的に獲得したものではなく，誰もが陥ることのあるものであると考えた．そこで，バイアスを矯正してフィルターバブルやエコーチェンバーの発生を押さえつけるのではなく，それを利用できる可能性を考える．

個人のバイアスを利用してより多くの客観的な情報を読ませる技術として，本研究では，ニュースタイトルをスタンスに合わせて書き換える方法を提案する．先ほどの例に照らし合わせて説明する.「ある病気に対して開発されたこのワクチンは，90 パーセントの予防効果を発揮するが，そもそも効果を得られるまでに接種から 1 週間の時間を要する」という，客観的事実だけ書かれた記事があったとする．人々にはこのワクチンについて,「素晴らしいワクチンに違いない」や,「人の体に悪影響をもたらすものに違いない」と考えたりする．この場合に，記事についたタイトルが「新ワクチンは 90 パーセントもの効果を発揮！」または,「新ワクチンは効果発揮まで 2 週間もかかる！」だった場合を考える．ワクチンに対して肯定的なバイアスを持つ人は前者のタイトルの記事を閲覧したくなり，否定的なバイアスを持つ人は後者の記事を選択しがちである．

このように，ニュース記事本文を書き換えずに，ニュースタイトルだけをスタンスに合わせて書き換えることを考える．こうすることで，あるトピックについて肯定的な人にも，否定的な人にも，記事本文を読ませることができる．これにより，バイアス自体は改善されず，強化される可能性もあるが，少なくとも少しでも多くの客観的な情報を両方の利用者に持たせられる．この研究では，異なるバイアスを持った人が同じ情報を多く共有することで，スタンスを問わず，少しでも建設的に議論することを助けることを期待している．すなわち,「やった」「やらない」,「言った」「言わない」,「効果がある」「効果がない」などの二元的水掛け論から,「これは○○というメリットがある一方で，××というデメリットもある」というような議論に，議論の性質を変えられる可能性がある．

このような，スタンスに合わせたニュースタイトルをつけるために，本研究では，近年主流になりつつある大規模言語モデルを用いた文書生成技術を用いた．GPT-2 などの Transformer を用いた言語モデルでは，大規模なデータで事前学習したモデルを少量のデータでファインチューニングすることで，文書生成や要約など様々なタスクを解くことができる．ここで，本研究では，あるニュース記事本文から，ニュースタイトルを生成するように GPT-2 をファインチューニングする．ただし，この際に，ある任意の単語を与えると，その単語を主題としてタイトルを生成するように，工夫して学習する．

実際の学習の模式図を，図 1 に示す．図の例では，野球の試合結果を伝える記事について，タイトルを書き換える概要を示

している．もとのデータでは，ニュースの本文が「12 月 24 日のリーグ戦で，神奈川県のプロ野球チームである相模原ファイターズが，町田市の町田ウォリアーズを 3-2 で下し……」という記事に対して,「ファイターズ勝利！投手好調！」というタイトルがつけられている．ここで，このタイトルを,「ファイターズ」を主題語として，ニュース記事本文を要約したものであると見なす．そのために，GPT-2 の入力形式に従い，主題語とニュース記事本文とニュースタイトルを，連結して 1 つの文書とする．GPT-2 は与えられた文（プロンプト）の続きを書くことに特化しているため，このような形式で学習すると，主題語とニュース記事を与えると，その続きとしてニュースタイトルを生成する言語モデルができる．このようなファインチューニングを行うことで，実際に任意のキーワードを主題語とした状態で，複数のニュースタイトルを生成することができる．

この際，従来の情報要約タスクに比べ，ニュースタイトルの生成は内容の正確性が重要視される．そこで，生成されたニュースタイトルが正しいものであるかを判定する機構を設ける．具体的には，生成されたニュースタイトルについて，それと意味的に近い部分がニュース記事本文内に存在するかを判定する．図 2 に表されるように，Sentence BERT に代表される，文類似度比較用の大規模言語モデルを用いて，生成されたタイトルとニュース本文中の各文の類似度を計算する．これにより，生成された各タイトルが，根拠に基づいている度合いが数値化される．そして最後に，生成時の順位（すなわち，スタンスに基づいている度合い），根拠に基づいている度合いをもとに，生成されたタイトルをランキングする．

こうして作成されたそれぞれのタイトルについて，実際に被験者に記事本文と合わせて見せる実験を行った．実験では，タイトルの自然さ，タイトルが主題に沿っているか，タイトルが記事内容に対して正しいかという，生成精度の評価を行った．また，手法によって生成タイトルにスタンスを含ませることが出来たかについても評価した．

本論文の構成について説明する．本章では，研究の背景と，本研究が最終的に目指す目的を述べた．第 2 章では，本論文に関連した分野についてすでに行われている研究を紹介し，本研究の位置づけを示す．第 3 章では，本論文で用いる手法について説明し，第 4 章ではそれによってもたらされた結果を示す．第 5 章にて結果に関する考察を行い，第 6 章にてまとめと今後の展望について述べる．

2 関 連 研 究

本研究は，BERT，GPT-2 等の機械学習分野，前提とする確証バイアスを含む心理学分野，特定のクエリに焦点を絞った要約を作る分野の研究と関連する．

2.1 言語モデル

本研究では，ニュースタイトルという自然文を扱う上で，事前学習された大規模言語モデルをファインチューニングして用いる．この際，文の生成には GPT-2 を，文の意味の同一性の判定には BERTScore，SentenceBERT の 2 種類をそれぞれ用いた．

GPT-2 とは Generative Pre-trained Transformer 2 の略で，OpenAI によって開発された自然言語処理モデルであり，文章を入力と出力にしている．その使用用途は，入力した文章から逐次的に次の単語を予測することである．

GPT-2 を用いた言い換えに関する研究の例として，Hegde ら [4] は，事前に学習させた言語モデルを用いて教師なしの言い換えの生成を行っている．本研究では，こうした書き換えと，BERT を組み合わせて用いることで，正確性を保ちながらタイトルを書き換えることを目的としている．Koppatz ら [5] は，フィンランド語における GPT-2 のファインチューニングを行っている．さらにそれを応用して事実に則したニュースヘッドライン，すなわちタイトルの自動生成を行っている．この論文では，実際に生成したタイトルを評価するにあたり,「Language」,「Usable」,「Good」の 3 つを評価基準として設けており，それぞれが言語として成り立っているか，ニュースタイトルとして利用可能であるか，ニュースタイトルとして優れたものであるかに対応しており，本物のニュースを扱うジャーナリストによって評価を行っている．本論文では目的こそ少し違けれど生成物が同じニュースタイトルであるため，このような複数観点からの評価を行なった．

BERT は Bidirectional Encoder Representations from Transformers の略で，Devlin ら [6] によって提唱された自然言語処理モデルである．これは過去のモデルと比べ，ラベルのないテキストから深い双方向表現を事前にトレーニングできるという進化を遂げている．本論文では，Reimers ら [7] によって開発された Sentence-BERT および Zhang ら [8] が開発した BERTScore を用いる．この技術は BERT そのものを並列に使用することによって，2 つの文章の類似度を比較することができる．本論文ではこれを用い，生成されたタイトルの根拠となる記述がニュース記事内に存在しているかを推定する．これにより，生成したニュースタイトルの中から，生成する元となったニュース本文から意味が異なっているものを取り除くことができる．

2.2 確証バイアス

確証バイアスとは認知バイアスの一種である．社会心理学や認知心理学などの分野でよく用いられ，個人が主義信条を持っている場合，それを満足させる，裏付ける情報を優先して取り込んでしまい，客観的な判断力を鈍らせてしまうことを指す．

確証バイアスに関する研究の例として，White [9] は，人々が持つ主義思想などのバイアスと，ウェブ検索行動との関連性を示している．インターネットにおいて，確証バイアスが起こす問題の最たる例がフィルターバブルとエコーチェンバーである．フィルターバブル，エコーチェンバーに関する研究の例として，Rhodes [10] はフィルターバブルおよびエコーチェンバー現象の実験を行っている．この研究では，アメリカにおける 2 大政党である共和党と民主党それぞれに同意する価値観を持った被験者を対象にして，それぞれの政党にすり寄ったフェイク

ニュースとどちらにも寄っていないフェイクニュースを読ませ，どれほどの人がどのように信じるかを計測している．

本論文では，この問題そのものを真っ向から打ち砕くようなアプローチではなく，この問題が発生している状態の人物にもデマのないニュースを読ませることを目標にしている．

2.3 クエリに焦点を当てた生成

この研究が関連する研究群として，QFS（Query Focused Summarization）という技術が存在する．これは特定の単語，文字に対して焦点を置いた要約に関する研究である．Kulkarni ら [11] の行っている研究は，QFS の中でも，複数文書を要約するタスクに関して，このタスクを学習するための学習データを作る研究をしている．他にも，Krysci'nski ら [12] は本のような長い文章で要約を行う研究をしている．このように，既存の研究で焦点を絞った要約を行う場合，長い文章や複数文章の要約を行う場合が多い．しかしながら本論文は，この特定のクエリに着目するという考え方をより短いニュースに対して応用している．

3 生成的言語モデルを用いたタイトル変換手法

本章では，ニュースタイトルの自動生成のために，GPT-2 を用いる手法について提案する．はじめに，事前に収集した現実のニュース記事について，それぞれを本文とタイトルに分け，学習用データセットを作成する．作成したデータセットを用い，ベースラインと提案手法の 2 つのモデルを学習させ，生成したタイトルをさらに識別器によって選別する．

3.1 GPT-2 によるタイトル生成の学習

はじめに，ニュースコーパスを GPT-2 のモデルに学習させ，文の続きではなくタイトルを生成するようにトレーニングする．本論文で使用する GPT-2 モデルは，rinna 社の公開している GPT-2-japanese-medium である．このモデルは，文の続きを制し得するというタスクを通して事前学習されている．そのため，はじめに文の続きを生成するモデルから，ニュース本文からニュースタイトルを生成する，要約モデルに作り替える必要がある．そこで，GPT-2 が持つ「文の続きを推定する」機能をもとに，ニュース本文からニュースアタイトルを生成するための追加トレーニングを行う．

本研究では，GPT-2 の訓練に，[SEP] という特殊なトークン（セパレータトークン）を含んだ文字列を用いる．このトークンを文の途中に挿入することで，そこを境に別の文であることを明示的に GPT-2 に伝える．具体例として，対話形式で返答を返すように GPT-2 で学習を行いたい場合に，GPT-2 は単なる文章の続きではなく，発話者が切り替わったことを理解する必要がある．例えば，話者 A が「こんにちは」といい，話者 B が「どうも，いい天気ですね」と返答し，話者 A がまた「それではごきげんよう」と返したことを GPT に教えたいとする．こういった場合，“こんにちは [SEP] どうも，いい天気ですね [SEP] それではごきげんよう” というように，発話ごとに [SEP] 変換し学習させる．こうすることで，学習後のモデルに “こんにちは” を入力した場合に，GPT-2 は文の続きを生成しようとして，“こんにちは，いい天気ですね” などを返す．一方で，“こんにちは [SEP]” という文を入力すると “どうも” などの応答を自動生成するようになる．

このような GPT-2 に，ニュース本文からニュースタイトルを生成する，要約的なタスクを解かせることを考える．もっとも単純に GPT-2 に要約タスクを解かせる方法として，本文と要約文をそのまま [SEP] トークンでつないで学習させる方法がある．こうすることで，GPT-2 は,「[SEP] の続きを推定する」という，文の続きを予想しているつもりで，実際には要約文を生成するようになる．

本論文では，スタンスに合わせたタイトル生成を行うことを目的としている．そのため，実際の学習では本文とタイトルだけでなく，そのタイトルに含まれるスタンスに基づく語を考慮する必要がある．そこで，本研究では，元からタイトルに含まれている固有名詞に注目した．ある実在のニュース記事があった際に，そのニュース記事につけられたタイトルは，ある主題にフォーカスしながらニュース記事を要約したものとみなせる．今回は，ニュース記事タイトル中に含まれる固有名詞を，その主題語としてみなことで，主題語にフォーカスしたニュース生成の学習を行う．

具体例として,「ワクチンの予防効果は 90 パーセントであり，効果は 1 週間後に現れる」というタイトルがあった場合,「ワクチン」や「予防効果」といった単語を抜き出し，これとタイトル，本文をセットで学習させる．具体的には，“ワクチン [SEP] 記事本文 [SEP] ワクチンの予防効果は 90 パーセントであり，効果は 1 週間後に現れる” などのように，主題語，記事本文，タイトルの順で文を並べ，GPT-2 に学習させる．こうすることで，例えば “副作用 [SEP] 記事本文 [SEP]” とだけ学習済みモデルに与えると，このモデルは「副作用」という単語に注目しながら，記事本文を要約する．

3.2 タイトルの根拠となる記事中の記述に注目した生成タイトルの選別

単なる要約タスクと異なり，ニュースタイトルの生成は，内容に根拠があることが重要である．いかに読者のスタンスに合致している要約であっても，ニュース記事本文と矛盾した，間違ったタイトルは避ける必要がある．たとえば，ニュース本文には「巨人が勝った」と書いているにもかかわらず，生成されたタイトルに「巨人が負けた」という書いてあってはならない．

このために，生成されたニュースタイトル中に本文で一切書かれていない内容，または本文と違う内容が含まれないようにする必要がある．そこで本研究では，任意のニュース記事と主題語が与えられた際に，はじめに多数のニュースタイトルを一度生成し，その中から根拠に乏しいものを除去するアプローチをとる．そのために，生成されたニュースタイトルと記事本文中の文の意味的類似度を計算し，生成されたニュースタイトルを再ランキングする．

2 つの文章がいかに意味的に近いものであるかを測る指標として，翻訳や要約における情報の同一性を表す評価指標がある．

具体的には，翻訳タスクの評価に用いられる BLEU（BiLingual Evaluation Understudy）や，要約タスクによく使われる ROUGE（Recall-Oriented Understudy for Gisiting Evaluation）などがある．また，翻訳にも要約にも使われる METEOR（Metric for Evaluation of Translation with Explicit ORdering）などの指標もある．一方で，これらの指標は N-gram に基づくテキストの完全一致で同一性を計算する．そのため，「A because B」と「B because A」のような因果関係の逆転を見逃しやすい点や，意味的な一致よりも文面の一致を優先しやすいなどの欠点があることが指摘されている [8]．また，ROUGE スコアは実際に人間が評価を行った場合との相関が弱いことも報告されている [13]．そのため，本研究のように，生成モデルを用いた，スタンスを考慮する要約には適さない可能性がある．

そこで本論文では，生成タイトルの再ランキングに，意味的類似度に比較的強い Transformer を用いた自然文の同一性指標を，2 つ用いた．ひとつめの指標は SentenceBERT [7] である．これは，2 つの BERT モデルがそれぞれ 1 つの文章をベクトル化することによって，2 つのベクトルがどれだけ似ているかを算出するコサイン類似度による類似度の算出を可能にする．ふたつめの指標は，BERTScore [8] である．BERTScore では文章を単語レベルに分割し，それぞれを前後の文脈の意味を含んだベクトルに変換する．そうした後で，総当たりでコサイン類似度をとることによってどの単語とどの単語が最も似た意味，文脈を含んでいるかを数値化する．そして，最後に IDF を用いた重みを使用することにより，2 つの文章がどれほど意味的に近いかを数値化して表す．

これら 2 つの指標は，同程度の精度を有しているとされる（BERTScore がわずかに精度が高いという報告もある [14]）．一方で，アルゴリズムを見ると，文章全体を重視するか，意味的に一致する部分が含まれるかを重視するかなどの点で，異なる．そのため，本研究ではこの 2 つの手法をどちらも起用することにした．

今校庭では，生成されたニュースタイトルがニュース本文と合致しているかである．しかし，これら 2 つは，文としての長さが全く違うため直接比較することが難しい．そこで，ニュース本文を句点で分割し文ごとに分け，その分割した文それぞれに対し生成したタイトルと比較した．本文と全く共通点のない突飛な文章が生成された場合，それは本文のどの文章とも意味的に近くないため，スコアは自然と低くなる．

この際，意味的類似度に関する考え方として，「タイトルとよく合致する記述がニュース記事内に 1 か所でもあればよい」という考え方と，「タイトルはニュース全体を反映して意味を保っていないといけない」という考え方がある．そのため，それぞれの識別機に対して，本文と比較した類似度の平均値を用いる方法と，最大値を用いる方法をどちらも行った．

最終的な値のランキング方法として，今回の実験に用いた実装では，生成時の順位と，意味的類似度の尺度のスコアをかけ合わせて最終的なランキングを決定した．GPT-2 はモデル内で生成された文章をモデル内での順位に応じて出力する機能があるため，それぞれの尺度の値に順位に応じた重みをかけることで，モデル側の評価値と識別機の評価値をどちらも反映できるようにした．

4 評　　価

本論文では，ニュースを主題語に注目しながら要約することで，スタンスに合わせたニュース生成を行っている．そのために，ニュースタイトルの日本語としての質，正しさ，スタンスが反映されているか，有用性について，それぞれ評価する必要がある．

4.1 生成タイトルの評価

生成されたタイトルが正しいかどうか，人手でラベル付けして評価した．ラベル付けの際には，

- **言語的正確性**: 日本語の文章として成り立っているか，
- **タイトルとしての正確性**: その文章がニュース本文を要約するタイトルとして間違っていないか，そして，
- **主観的評価の非同一性**: 人の受け取り方の違い

という 3 つの観点に基づいて項目を設定した．実際の評価時のラベル付け項目を図 2 に示す．

言語的正確性，タイトルとしての正確性は，単純に「この文は日本語として正しいですか」および「この文章は本文に即したタイトルになっていますか」と問えば，簡単にラベル付けできる．一方で，主観的観点の同一性の評価項目では，「あなたが仮にこの主題語に興味のある読者だったとして，このタイトルをどう思うか」という，ロールプレイを含むラベル付けを行った．

4.2 学習用データセット

GPT-2 のファインチューニングのために，ニュースデータを収集し，学習に用いた．ニュースコーパスの作成には，GoogleNews を用いた．GoogleNews は yahoo ニュースや新聞各社のオンライン版など，数々の掲載元から集約されたニュースを，API を用いて検索可能である．複数の掲載元から集約しているため，同じ事件について，新聞社ごとに異なるタイトルをついている場合があり，本研究に適している．

一方で，複数サイトから集めたものなので，タイトルや記事フォーマットに違いがあるため，事前にクレンジングを行った．例えば，連続するニュースの速報記事の場合，「ワールドカップ」などのトピック名だけの試合の記事が多く含まれた．こうしたデータは，ルールベースで取り除いた．また，本文が別サイトや次ページに記載されているニュースなども同様に，ルールベースで除去した．

4.3 比 較 手 法

比較対象の手法として，4 通りの提案手法を含む，11 個の手法をそれぞれ実装した．実際の手法の一覧を表 3 に示す．

本研究は，通常の文書要約タスクに，主題語の考慮と，根拠による再ランキングをそれぞれ組み込んだものである．そのためベースラインとなるモデルは，タイトルと本文のセットを学習しただけの，何も考えずに本文を要約するモデルである．こ

のモデルでは，主題語を用いておらず，GPT-2 から最も高いスコアで出されたものをそのまま生成結果として用いる．

次に，主題語の考慮の有無を比較するために，主題語を考慮するファインチューニングの有無で手法を 2 分した．主題語を考慮するが，根拠推定を行わない手法を「トピック FT のみ」とした．そして，根拠による再ランキング手法が 4 種類あるので，合計 8 種類の組み合わせの手法を作成した．「ベースライン」で用いた言語モデルに，4 種類のテキスト間類似度指標で根拠に基づく再ランキングをした 4 つの手法,「トピック FT のみ」で用いた言語モデルに，4 種類の類似度指標で再ランキングをした 4 つの手法をそれぞれ比較する．

また，最後に，人手で作成したタイトルの例として，もとの記事タイトルそのものについても評価対象に加えた．このため，1 件のニュース記事に対して，合計 11 個のニュースタイトルがラベル付けの対象となる．

4.4 評価タスク

それぞれの手法で生成されたタイトルが実際に正しく，スタンスを反映し，興味を引くかを判定するため，被験者実験を行った．被験者は学部 3 年生で，人数は 2 名である．被験者数を 2 名としたのは，このタスクがある程度客観的に判断可能なラベル付けタスクだと考えたためである．

はじめに，2023 年 1 月 9 日から 1 月 11 日の間に GoogleNews において収集したニュースから，なるべくニュースカテゴリが多用になるように，30 件のニュースを抽出した．そして，各ニュース記事について，タイトルには含まれない別スタンスの単語を，ニュース記事本文から人手で 1 つ選定した．こうして，1 つのニュース記事に対して，重複を除去して最大 11 件のニュースタイトルが生成された．

被験者は，1 件の記事本文と，11 件のタイトルについて，それぞれのタイトルが各評価項目に当てはまるかどうか，ラベル付けした．具体的には，表 2 に示される評価項目として，日本語として適当な文であるか，タイトルとして適当な形の文であるか．内容と本文に齟齬がないか，指定したクエリに興味がある人向けであるか，指定したクエリに興味があると仮定した場合にクリックしたくなるタイトルであるかの 5 項目について，各項目独立に 5 段階の評価値をつけた．これを 30 記事分，繰り返した．バイアスを防ぐため，タイトルをランダムな並び順に変換したものを使用し，本来のタイトルが含まれていることも伏せた．実際に選定したニュースの元のタイトルと，主題語となる主題語の一覧を表 1 に示す．ニュース記事はカテゴリがなるべく被らないように人手で選定した．

4.5 実 験 結 果

本節では，被験者に評価してもらった評価データをもとに手法の優位性を確かめるための t 検定の結果と，被験者の評価の信頼性を示すカッパ係数を求めた結果を示す．

はじめに，本実験は被験者を用いた実験であるため，被験者の信頼性を確かめる必要がある．実験で用いたそれぞれの被験者の評価データに対し，信頼性評価によく用いられるコーエンのカッパ係数を求めた．本実験では 1 から 5 までの五段階評価を行ったため，値の差の絶対値を反映させることができる二次の重みを付けたカッパ係数を用いた．結果を表 2 に示す．

表 1　実験に使用した記事の元タイトルと選んだクエリ（主題語）．

元のタイトル	選択した主題語
来議会占拠に酷似、ブラジル前大統領の支持者「票が盗まれた」と訴え暴徒化	ブラジル
川南町の養鶏場でニワトリ死ぬ 鳥インフル感染の疑い	鳥インフル
韓国の中国大使館 中国訪れる韓国人への短期ビザ発給を停止	新型コロナ
ウクライナ東部 ロシア軍が大量の砲撃 激しい攻防続く	ゼレンスキー
東京都 新型コロナ 28 人死亡 7462 人感染確認 6 日より前週下回る	重症
サンマ 去年の水揚げ量 4 年連続で過去最低	全国さんま棒受網漁業協同組合
気温上昇 1.5 度に抑えても＂今世紀末までに約半数の氷河消滅＂	海面上昇
巨人移籍の松田 DeNA 佐野ら「チーム熟男」が自主トレ開始 宮崎	ソフトバンク
宮崎 河野知事が陳謝 初詣の事実を感染確認後に新聞に修正依頼	厳重注意
岸田首相 イタリアで首脳会談 外務・防衛の協議枠組み立ち上げ	G7
Steam が同時ゲーム内プレイヤー数 1000 万人を達成、新型コロナパンデミックから増加傾向	オンライン同時接続者数
将棋王将戦「世紀の対局」は藤井聡太五冠が勝利…掛川市では「御城印」など大盛り上がり！一方で望ましくない事態も…	羽生九段
イギリス ヘリー王子の自伝発売 父の再婚反対など赤裸々に	暴力
「どうする家康」１５・４％　ＮＨＫ大河の初回視聴率	関東地区
ファーストリテイリング、年収最大４割上げ　３月改定　「ユニクロ」展開	新入社員
明日の株式相場に向けて－米ＣＰＩ待ちも材料株物色は花盛り	日経平均株価
新型プリウス発売、燃費は上級モデルでリッター２８・６ｋｍ…ＰＨＶは３月頃に投入	トヨタセーフティセンス
サービス終了「東方ダンマカグラ」のクラファンに約 2 億円集まる　目標の 1288 %　買い切り版リリースへ	開始から 30 分
林外相、中国ビザ停止「極めて遺憾」　抗議し撤廃要求	措置の撤廃
ロシア軍、要衝バフムト近郊の町に砲撃８６回…ウクライナ軍「東部戦線で最も困難な状況」	ドネツク
AMD/Intel 製 GPU の AV1 エンコーダーも利用可能 ー「OBS Studio 29.0」が正式版に	使い勝手向上
吉村真晴 五輪２連覇 中国・馬龍に大金星！世界卓球出場権を獲得【世界卓球 アジア大陸予選】	ダーバン
チップレットになった「第 4 世代 Xeon SP」、性能向上の鍵は AMX と 4 つのアクセラレータ	Intel
「マニュアル借のギア操作誤った」歩道に車突っ込み男性 (40) 死亡　スポーツクラブ事務員の男逮捕　東京・町田市	27 歳の男を逮捕
大阪 遺額堀川でニホンウナギ生息確認 学術的調査で初めて	研究所などの調査
日英、部隊往来の円滑化で協定署名へ　豪に続き２カ国目	岸田文雄
遺留物からＤＮＡ型採取　ゴミ箱にペットボトル―宮台さん襲撃事件・警視庁	襲撃
春節＂21 億人大移動＂直前―日本の観光地「期待と不安」　中国は「ビザ」で対抗措置か	ビザの発給
ネットで検索したイラスト、「学校だより」に無断掲載―小学校長「無料素材と思い込み」	著作権
ツインズがコレアと６年２億ドルで再契約合意　最大 10 年２億 7000 万ドル　メッツと契約成立ならず	身体検査

表 2　実際のラベル付け項目と，被験者間の信頼性評価（κ 係数：0.21 以上で Fair な合意）．

項目名	二次重み付きカッパ係数
日本語として正しい文か	0.247
タイトルとして正しい文か	0.469
内容に齟齬がないか	0.595
クエリに対し興味のある人向けか	0.583
興味がある人だとしたら，クリックしたくなる文か	0.488

カッパ係数の評価について，本実験では良く用いられる Landis ら [15] によって提唱された基準を用いる．この基準によれば日本語として適切であるかの項目は Fair（ある程度の一致）と評価され，その他の項目は 1 段階上の Moderate（適度な一致）として評価されることがわかる．5 段階評価であること，また日本語としての完成度をどのように評価するかは人それぞれであることを鑑みるに，十分な一致がとれたと言える．

次に，実際に被験者がタイトルをどのように評価したのかを，候補者と記事の二つの面から平均を取り，手法と項目のみで比較できるようにしたものを表 3 に示す．全体的に，主題語を用いたファインチューニングによって，主題語によって興味がある人向けになっており，クリックしたくなる傾向がみられた．また，根拠推定の有無により，内容の正しさだけでなく，日本語やタイトルとしての質も向上した現象がみられた．また，最もニュースタイトルとして正確で，内容を表しているのは元のニュースタイトルであるが，一方で興味を反映しているか，主題語に興味のある人がクリックしたくなるかという項目では，多くの場合で主題語に基づくファインチューニングを行った手法が高評価であった．特に評価が高かったのは，主題語に基づくファインチューニングを行ったうえで，BERTScore でニュース中の各文との意味的類似度を算出し，全文の類似度の平均を用いたトピック FT+BSAVG の手法であった．

5 考　　察

はじめに，本論文で提案した手法についての総合的な評価を行う．比較するのは表 3 における，ベースラインと下 4 つの提案手法である．提案手法は識別機の計算方法に応じ 4 種類が

表 3　手法，項目ごとの平均評点（5 段階中）．ファインチューニング（FT）の有無と，根拠推定に用いた手法ごとの被験者評点（5 段階）の，評価項目ごとの集計．下 4 つの手法が提案手法（ベースラインと比べて ** $p < 0.01$, * $p < 0.05$）．

手法名	トピック FT	根拠推定	日本語	タイトル	内容	興味	クリック
元のタイトル	なし	なし	**4.82	**4.88	**4.07	**2.42	**2.47
ベースライン	なし	なし	4.05	4.10	1.95	1.32	1.47
BSMAX のみ	なし	BSMAX	*4.40	4.28	**2.92	**2.15	*2.10
BSAVG のみ	なし	BSAVG	4.33	*4.43	*2.47	**1.90	*1.95
SBMAX のみ	なし	SBMAX	*4.38	4.37	**2.75	**2.13	*2.20
SBAVG のみ	なし	SBAVG	4.27	4.42	**2.50	**1.98	**2.10
トピック FT のみ	あり	なし	4.35	4.42	**2.80	**2.63	**2.68
トピック FT+BSMAX	あり	BSMAX	*4.57	*4.60	**3.57	**2.60	**2.60
トピック FT+BSAVG	あり	BSAVG	*4.53	*4.55	**3.38	**3.05	**3.10
トピック FT+SBMAX	あり	SBMAX	**4.53	**4.70	**3.63	**2.85	**2.80
トピック FT+SBAVG	あり	SBAVG	**4.58	*4.53	**3.23	**2.83	2.82

あるが，そのいずれもベースラインに対し，五つの項目すべてにおいて有意差が認められることがわかる．これにより，ただニュース記事を学習させただけの GPT-2 よりも，特定のクエリに焦点を当てた学習と，生成後の識別を挟んだモデルの方が精度の高いタイトルを生成できることが分かった．

次に，提案手法の一つである識別機の評価を行う．本実験における識別機の役割は，本文と全く意味の異なる生成タイトルを取り除くことにある．よって，ここでは評価項目のうち，日本語，タイトル，内容の 3 点に絞って考察を行う．まずはベースラインとの直接比較として，識別器のみの 4 つの手法を上げる．日本語の項目では BSMAX のみの手法に，タイトルの項目では BSAVG の手法に，内容の項目では SBAVG のみの手法に有意差が認められている（日本語：$p = 0.032$，タイトル：$p = 0.029$，内容：$p < 0.01$）．他の項目ではこの傾向はみられないことから，識別器が内容の同一性を担保する効果があることがわかる．

次に，提案手法の 1 つであるトピック FT の評価を行う．はじめに，トピック FT は読者のタイトルへの関心を，特定の方向性を持たせることによって引き上げることを目的としている．そのため，ここでは主に後半の評価項目である興味，クリックの 2 点に絞っての考察を行う．まずはベースラインとの直接比較として，表 3 より，トピック FT のみの手法が，ベースラインに対して内容，興味，クリックの 3 つの評価項目において有意差が認められることがわかる．次に，BSMAX のみの手法に対するトピック FT+BSMAX の手法のように，同じ識別機を挟んでいるがトピック FT のあり，なしの差がある手法についても考察する．表 4 に，興味，クリックの 2 つの項目に絞った t の結果を示す．BSMAX 以外のすべての識別手法において，トピック FT のあるなしで有意差が認められていることがわかる．仮説通り，特定のクエリに焦点を置いたファインチューニングを行うことによって生成されるタイトルにそのクエリに関する興味を刺激するような要素を含めることができた，すなわちスタンスを含ませられたと言ってよいだろう．また，提案手法の一部は，スタンスに関する評価指標である，指定したクエリに興味がある人向けであるか，指定したクエリに興味があると仮定した場合，クリックしたくなるタイトルであるかの 2 項目において，元のタイトルに対し有意差が認められた．これにより，元のタイトルから異なるスタンスのタイトルを生成できたと言える．

表 4　同じ識別手法を用いた手法同士の t 検定結果

識別器の種類	興味	クリック
識別器無し	0.000	0.000
BSMAX	0.074	0.053
BSAVG	0.001	0.001
SBMAX	0.017	0.029
SBAVG	0.002	0.005

表 5　ブラジルでの大統領選について，元タイトルと異なり,「ブラジル」という単語に焦点を置いて生成した各手法のタイトル

手法	生成タイトル
元のタイトル	米議会占拠に酷似、ブラジル前大統領の支持者「票が盗まれた」と訴え暴徒化
生成のみ	9・6 日本との国交 50 周年... 日米との連携強化を強調
BSMAX のみ	ブラジル大統領選に異議を唱えるジャイル・ボルソナロ氏支持者らが連邦議会襲撃、警察に拘束される
BSAVG のみ	ブラジル大統領選ボルソナロ氏が勝利の立役者だった?
SBMAX のみ	ブラジル大統領選に異議を唱えるジャイル・ボルソナロ氏の支持者らが「ブラジル大統領を逮捕せよ」... 大統領府襲撃も襲撃
SBAVG のみ	ブラジル大統領選の衝撃冷めやらぬまま... 早朝からボルソナロ氏が各地で抗議、裁判所占拠を襲撃
トピック FT	大統領選で「敗北認めず」、露の支持者がバスで襲撃... 支持者らが軍施設の前などで抗議続ける... 治安当局「冷え込む」
トピック FT+BSMAX	ブラジル大統領選、混乱の続くボルソナロ氏が敗北認めず... 支持者らは軍施設の前などで抗議
トピック FT+BSAVG	ブラジル大統領選、現職のボルソナロ氏が優勢... 支持者らは軍施設の前などで抗議継続、対立も
トピック FT+SBMAX	ブラジル大統領選、ボルソナロ氏が敗北認めず... 支持者らは軍施設の前などで抗議継続、支持者らは軍施設の前などで抗議
トピック FT+SBAVG	ブラジル大統領選、混乱の続くボルソナロ氏が敗北認めず... 支持者らは軍施設の前などで抗議

ここからは，様々なカテゴリから用意した 30 種類の記事について，記事の内容やクエリの選定が実験結果にどのような影響をもたらしたかについて考察していく．いずれの評価項目においても，手法に関わらず高得点を挙げた「ブラジル」の記事についての生成結果を表 5 に示す.「ブラジル」はブラジルで行われた大統領選についての記事であったが，この記事についてはベースラインでも高い精度での生成が行われており，これは手法というより GPT-2 そのものの特徴によるものであると考察できる．では，この記事を軸として GPT-2 によるタイトル生成について考察を行う．この記事は大別すると政治的なカテゴリに入る．本実験で使用したデータは，収集した時期は違うがどちらも GoogleNews であり，学習データにも政治的な記事は多分に含まれていた．しかしながら同じ政治系の記事の中でもこの記事のみが異様に高い評点を出している．これは，政治系というありふれた分野でありながら，ブラジルにフォーカスするという日本ではあまり見かけない記事であるために，GPT-2 内でタイトルに含まれる可能性のある単語が小さく絞られていたためではないかと考えられる．この例から，GPT-2 には文脈はありふれた，しかし中に入る単語はありふれていないという記事との相性がいいのではないかと考察できる．

前述のように GPT-2 との相性の良い記事がある反面，日本語やタイトルとしての文章の質は悪くなくとも，内容とは齟齬があったり，クエリを用いても方向性を誘導できなかったりした記事も存在する．ここではその最たる例である図中の「オンラインゲーム同時...」と図中の「ダーパン」について考察していく.「オンラインゲーム同時...」の記事は大手の PC ゲームプラットフォームである Steam についての話題であった．この記事のタイトルは，そもそも内容が正しいものが極端に少なく，生成タイトルに全く関係のない話題が混ざってしまっていた．「ブラジル」の記事と同じく，この記事もあまりニュースには出てこない単語が多かった半面，ゲーム分野という学習データの少ない記事であったためにこういう結果になったのだと思われる．次に「ダーパン」の記事について，これは卓球の吉村真晴の試合に関しての記事であり，一見すると学習データの多いスポーツカテゴリの記事である．しかしながら，この記事に関する生成結果を見ると,「吉村長官」「吉村知事」といった誤生成が行われていた．別の記事でも「中国の報道官」が「中国の

外交官」になるなど，人名や役職名といった普遍的な単語は誤生成を招きやすいことが分かった．

最後に，本研究では提案手法を識別器の使い方によって 4 種類に分けている．SentenceBERT を用いた手法は日本語，タイトルのスコアが高く，BERTScore を用いた手法ではないよう，興味，クリックのスコアが高くなるという傾向が得られた．しかしながら，表 3 より，これらはいずれにしてもベースラインと比較し高いスコア，有意差を示しており，その手法同士に明確な差はなかった．このようなタスクに使用するのであれば，どちらを利用しても研究に大きな差は出ないと考えられる．

6 まとめ

本研究では，ニュース本文の内容を要約するニュースタイトルとしての機能を保持したまま，GPT-2 を用い，スタンスを変更した言いかえを行うことで，より多くの人々に興味を持ってもらうことを目的としたタイトルの自動生成手法を提案した．その目標を達成するにあたって，本文中のクエリを本文とは別に入力として与えることによって，本来のタイトルとは異なるスタンスによって書かれたタイトルを生成した．また，機械学習による生成の弊害とされる正確性の低さを補うため，BERTScore と SentenceBERT を用い，本文との一致度の低い結果をはじくことによってタイトルとしての質を担保する手法を提案した．

評価実験は被験者実験を通じて行い，2 人の被験者には元のタイトルを含めた最大 11 種類のタイトルをランダムな順番で提示し，日本語として成り立つか，タイトルとしてふさわしい文の形か，本文との間に内容の齟齬がないか，指定したクエリに興味のある人向けであるか，興味があるとしたらクリックしたくなる文であるかの 5 項目での評価を行った．

実験の後に，被験者の評価データを解析した結果，被験者同士の回答の信頼性を求めるコーエンのカッパ係数には十分と言える一致がとれた．また，提案手法のファインチューニング方法と識別器のそれぞれに対して，設定した仮説を満たす有意差が認められた．これにより，クエリを用いたトピック FT によって特定のスタンスを持つ人がタイトルから受ける印象が変わること，また，識別器を用いたタイトルの選別によってよりニュース本文との齟齬の少ない精度の高いタイトルが生成可能であることが証明された．

今後の課題として，5 章にて言及した記事のカテゴリやクエリの選び方に関する考察は，あくまで評価データの傾向などをもとにした推測の域を出ない．本研究では汎用的にニュースタイトルを生成する方法を提案したが，今後は細分化された記事のカテゴリごとのファインチューニングなどを行うことで，更に精度の高いタイトル生成が可能になることが期待される．

謝辞

本研究の一部は JSPS 科研費 21H03775，21H03774，22H03905 による助成，ならびに 2022 年度国立情報学研究所共同研究 22S1001 の助成を受けたものです．ここに記して謝意を表します．

文献

[1] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* penguin UK, 2011.

[2] Federico Cinus, Marco Minici, Corrado Monti, and Francesco Bonchi. The effect of people recommenders on echo chambers and polarization. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, No. 1, pp. 90–101, May 2022.

[3] Quentin Grossetti, Cédric du Mouza, and Nicolas Travers. Community-based recommendations on twitter: Avoiding the filter bubble. In Reynold Cheng, Nikos Mamoulis, Yizhou Sun, and Xin Huang, editors, *Web Information Systems Engineering – WISE 2019*, pp. 212–227, Cham, 2019. Springer International Publishing.

[4] Chaitra Hegde and Shrikumar Patil. Unsupervised paraphrase generation using pre-trained language models. *arXiv e-prints*, pp. arXiv–2006, 2020.

[5] Maximilian Koppatz, Khalid Alnajjar, Mika Hämäläinen, and Thierry Poibeau. Automatic generation of factual news headlines in Finnish. In *15th International Natural Language Generation Conference (INLG)*, Seattle, United States, July 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[7] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

[8] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTscore: Evaluating text generation with BERT. *arXiv e-prints*, pp. arXiv–1904, 2019.

[9] Ryen White. Beliefs and biases in web search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, p. 3–12, New York, NY, USA, 2013. Association for Computing Machinery.

[10] Samuel C. Rhodes. Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation. *Political Communication*, Vol. 39, No. 1, pp. 1–22, 2022.

[11] Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv e-prints*, pp. arXiv–2010, 2020.

[12] Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. Booksum: A collection of datasets for long-form narrative summarization. *arXiv e-prints*, pp. arXiv–2105, 2021.

[13] Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, 2020.

[14] Andrei Paraschiv and Dumitru-Clementin Cercel. Upb at germeval-2020 task 3: Assessing summaries for german texts using BERTscore and sentence-BERT. In *SwissText/KONVENS*, 2020.

[15] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, Vol. 33, No. 1, pp. 159–174, 1977.