DEIM Forum 2023 3a-2-2

手続き的知識源を用いた方法を表すクエリからの目的抽出

工家　昂之† 　山本　岳洋† 　荘司　慶行††

† 兵庫県立大学 大学院情報科学研究科　〒 651–2197 兵庫県神戸市西区学園西町 8-2-1
†† 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 5–10–1
E-mail: †ad22l024@gsis.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††shoji@it.aoyama.ac.jp

あらまし　目的を達成するために手順や方法を知りたいと思った際，しばしばウェブで検索をする．既存の検索システムでは，入力したクエリに関する手順や方法が提示されるが，それ以外の方法や手順が提示される可能性は低い．例えば,「車のタイヤ　チェーンを巻く」というクエリからは「雪道を自動車で走る」という目的があるが，これは「轍を走る」という方法でも達成することができる．しかし，このような方法を既存の検索システムで調べる場合,「轍を走る」のようにクエリを変更する必要がある．そこで，入力されたクエリから，ユーザの目的やほかの方法を自動で取得し提示するために，クエリから目的を抽出することを提案する．例にあるような，何かをする際のやり方に関する知識のことを手続き的知識と呼ぶ．この手続き的知識源として，ある目的について複数の解決方法を集めたまとめサイトである wikiHow を用いた．目的を抽出するため，本研究のアプローチとして手続き的知識源を用いた，文生成モデルのファインチューニングを行った．

キーワード　情報検索，Web 検索，信憑性，確証バイアス

1　は じ め に

今日では，目的を達成するためのあらゆる方法や手順を，ウェブで検索するようになってきている．検索エンジンの発達で，入力したクエリに関連するキーワードを自動で取得し，ユーザにそれを提示することが可能になっている．例えば，雪が積もる路面を走行するために，車のタイヤにチェーンを巻くことを考える．タイヤチェーンを手に，これからタイヤに巻くのだがそのやり方がわからない．ユーザが手順について検索する場合,「車のタイヤにチェーンを巻く」をクエリとして入力すると考えられる．このとき，既存の検索システムではタイヤチェーンを巻く手順が示される．また，これと同時にクエリ候補が示され，それを選択することでユーザが求める情報を検索することができる．

しかし，ここで見つかる方法は，数ある方法の中の一つであることが多い．また，既存の検索エンジンではユーザの入力したクエリに関する情報が多く提示される．では,「車のタイヤ　チェーン　巻く」と検索したユーザが何を達成したかったのかを考える．入力されたクエリよりこのユーザが達成したいこととして，車のタイヤにチェーンを巻くことで，雪道を自動車で走るということが考えられる．雪道を自動車で走るためには，車のタイヤにチェーンを巻く方法以外にも轍を走る方法が存在する．しかし，既存の検索システムではこのような方法を「車のタイヤにチェーンを巻く」と入力したユーザに提示することは難しい．

そこで本研究では，多様な情報をユーザに提示するために,「車のタイヤ　チェーン　巻く」のような方法を表すクエリから「雪道を自動車で走る」という目的を抽出することを提案する．ユーザの目的を抽出しその目的で検索することで，図 1 に示すような複数の解決方法を提示することができると考えた．本研究のアプローチは，文生成モデルのファインチューニングを行い，入力された方法を表すクエリから目的を生成することである．本研究では文生成モデルとして T5 と BERT2BERT を用いる．また，これらのモデルによる生成では，1 つの入力に対して複数の生成を行うと，類似した生成結果が連続して生成される．この問題を解決するために生成後に結果の多様化を行った．

ファインチューニングを行うために手続き的知識の大規模なデータセットが必要になるが，既存のデータセットが存在しないためデータセットを作成する必要がある．そこで本研究ではデータセットを作成するために wikiHow を用いた．wikiHow とは，ある目的について複数の解決方法を集めたまとめサイトである．提案手法の有効性を評価するために，wikiHow データセットより作成したテストデータと，自作のデータセットより作成したテストデータの 2 種類のテストデータを用いて評価を行った．自作のデータセットによる評価では，wikiHow データセットには存在しない日常生活で実際に検索されそうなクエリに対する汎用能力を評価する．

提案手法の有効性の評価には，自動評価と人手評価の 2 種類の評価方法で行った．自動評価では正解との類似度を BERTScore で数値化しその平均を計算する評価と，MRR による評価の 2 種類の評価を行った．人手評価では正解との類似度や，入力と生成結果の関係を考慮し評価した後，DCG@10，P@10 のそれぞれの平均を計算し評価した．なお，多様化の有無による影響を検証するために，多様化の前後でそれぞれ同様の評価を行った．これらの評価基準に従って生成結果を評価した結果，すべての評価において T5（多様化なし）が最も高い評価結果を達成した．

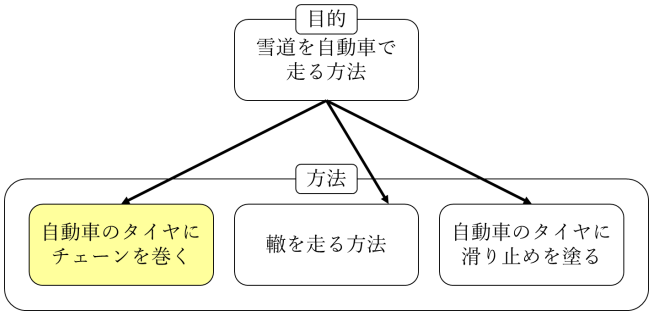

図 1　方法と目的の関係．図中の「目的」雪道を自動車で走るは,「方法」自動車のタイヤにチェーンを巻く，轍を走る，自動車のタイヤに滑り止めを塗る，の目的になっている．

2　関 連 研 究

本研究では，方法から目的を抽出するという問題に取り組む．本研究と関連が深い，タスク検索および，クエリ推薦に関する研究について述べる．

2.1　タスク検索

ある目的を達成するために必要な方法の検索をタスク検索と呼ぶ．例えば，雪道を自動車で走りたいユーザは「自動車のタイヤにチェーンを巻く」というクエリで検索をする．しかし，同じ目的を達成する別の解決策として「轍を走る」方法も存在するが，ユーザはこの解決策の存在に気付くことができるのかという問題がある．もしこの解決策に気づけない場合，ユーザはこれらの解決策を考慮せずに「車のタイヤにチェーンを巻く」という解決策を選択する．この問題を解決するために，Pothirattanachaikul らは特定のクエリの代替アクションをコミュニティQ&A サイトからマイニングする方法について報告している [7].

加藤らはあるタスクを達成するために必要なタスクの集合を Web から発見する「タスク検索」を提案している．加藤らの提案手法では，実行することでユーザが入力したタスクを達成できるタスクを Web 検索により発見し，一連のタスク集合を形成する．その後，形成したタスク集合について，実行することでタスクを達成できる度合いのランキングを行うというものである [14].

タスク検索においてあるタスクを完了するために必要なサブタスクが存在する．このサブタスクとは，完了すると目的のタスクが完了するようなタスクである．このサブタスクのマイニングについて Yamamoto らは，検索連動型広告を利用したクエリのクラスタリングによる取り組みについて報告している．彼らはあるタスクとそのサブタスクについて，search goal と subgoal という概念を提案している [10]．本研究において目的が search goal，方法が subgoal と対応している．

Hassan らは，タスクを完了するために必要な手順をユーザに提供するために，タスク間の関連性を自動的に発見する手法を提案している [1].

これらの研究は，図 1 に示す「雪道を自動車で走る」から「自動車のタイヤにチェーンを巻く」や「轍を走る方法」を提示する研究である．しかし，本研究ではこれらの研究とは逆に，「自動車のタイヤにチェーンを巻く」や「轍を走る方法」から「雪道を自動車で走る」を提示することに取り組む．

2.2　クエリ推薦

Zhang らは手続きに関する知識はタスク検索クエリに応答する際に重要であると報告している [12]．Boldi らはクエリ修正の分類について報告している．あるクエリが与えられ，そのクエリが修正されたとき，修正前後の関係が特化であるのか，あるいは汎化であるのかの分類を行っている [3]．本研究では汎化に分類される生成を行う．大石らは検索エンジンに与えるクエリを改善するクエリ拡張を行うために，重要な語の近くに出現する単語は重要であるという考えに基づいた，センテンス間の距離に注目した関連単語抽出アルゴリズムの提案を報告している [15]．Bhatia らはクエリログが少ない状況でクエリ推薦を行う場合，クエリログを使用せずに確率論的メカニズムでクエリ推薦を行い，既存の試みよりも高品質なクエリの提案が行えたと報告している [2].

田麥らは，ユーザの目的推定に関する研究として，サービスの検索において，入力されたクエリに対してどの種類のサービスを利用したいかの推定について報告している．なお，クエリとサービスの種類の結び付けには検索連動型広告を利用している [16].

Zamani らは，クエリを提示する検索システムの提案を行っている．彼らは，ユーザが入力するクエリは短く，抽象的であるという仮説を立て，ユーザ実験を実施した．実験ではクエリの明確化をシステムが行い，それをユーザに提示するというユーザ実験である．実験結果では「検索しやすくなった」や「商品検索においてストレスが軽減された」とアンケート結果が得られたと報告している [11].

本研究で取り組む，方法からの目的の抽出という問題は，クエリ推薦における汎化の 1 種である．しかし，一般的なクエリ推薦における汎化では，クエリ中の単語の削除や，単語の上位語化などで汎化が実現できるが，本研究で取り組む「自動車のタイヤにチェーンを巻く」のようなタスク検索では，一般的な汎化では難しい．

3　方法を表すクエリからの目的抽出

本節では，本研究で扱う方法と目的の概念について定義し，その後問題定義を行う．

3.1　方法と目的の定義

本研究で扱う方法と目的についての定義を雪道を自動車で走行したいユーザを例に説明する．雪道を自動車で走行したいユーザがその手段として，車のタイヤにチェーンを巻くという手段を選択する．そのユーザが車のタイヤにチェーンを巻くや

り方について検索する場合，入力されるクエリとして「車のタイヤ　チェーン　巻く」が考えられる．この，何かをする際のやり方に関する検索で入力されるクエリを本研究では「方法」と呼ぶ．本来「車のタイヤ　チェーン　巻く」と入力したユーザには，雪道を自動車で走行するという目標があった．つまり，このユーザには車のタイヤにチェーンを巻くことで，最終的に雪道を自動車で走ることが目標である．このような，ある方法を達成すると同時に達成される方法を本研究では「目的」と呼ぶ．

3.2 問題定義

本研究における問題定義は以下のとおりである．

- 入力: 方法を表すクエリ
- 出力: 入力の達成と同時に達成される方法，すなわちユーザの目的

入力した方法以外の方法を見つけるために，まずユーザの目的を抽出することを考える．本研究では方法を表すクエリを入力し，その方法の達成と同時に達成される目的を出力するモデルを考える．

4 手続き型知識源を学習に用いた文生成による方法からの目的抽出

まず，提案手法で行う，文生成モデルのファインチューニングの概要について述べる．その後，ファインチューニングのためのデータセットの構築方法について述べる．

4.1 文生成モデルのファインチューニング

方法を表すクエリから目的を抽出するための手法として，クエリ中のキーワードを削除することや上位語に置き換えるという手法がある．この手法を用いると，例えば,「自動車のタイヤにチェーンを巻く」というクエリの場合,「乗り物のタイヤにチェーンを巻く」のような置き換えとなる．また，キーワードの削除では,「タイヤににチェーンを巻く」のようになる．しかし，この手法では本研究が目指す「雪道を自動車で走る」という目的を抽出することはできない．よって，クエリ中のキーワードの削除や上位語へ置き換えることだけでは不十分であると考えた．

そこで，文生成モデルのファインチューニングにより，方法を表すクエリから目的を生成することで目的抽出を実現する．文生成の手法には図 2 に示す Encoder-Decoder と呼ばれるモデルを用いる．このモデルは文の情報を集約する Encoder モデルと，Decoder と呼ばれる集約した情報から文を生成するモデルが組み合わさったモデルである．具体的には，本研究では事前学習済みモデルとして，T5 [8] および BERT2BERT [9] を用いてそれぞれのモデルをファインチューニングし，性能を比較する．

これらのモデルの損失関数には交差エントロピーを用いた．出力はビームサーチにより，上位 n 件を抽出する．

4.1.1 T5

T5（Text-To-Text Transfer Transformer）は様々な自然言語処理タスクを自然言語で解くモデルである．T5 の事前学習には次の 3 種類のタスクが検討された．

- Prefix language modeling
- BERT-Style
- Deshuffling

それぞれのタスクについて説明する．Prefix language modeling は文の出だしが表示され，その後にどう続くかを予測するタスクである．BERT-Style は文章中の 15%の語を伏字にした後，その 10%をランダムな語に置き換えて元の文章を復元させるタスクである．また，Deshuffling 語の順番を入れ替えて元の文章を復元するタスクである．上記 3 つのタスクでは BERT-Style のスコアが最良だったことが報告されている [8]．

本研究で用いる T5 は事前学習に日本語版 Wikipedia のダンプデータ (2020 年 7 月 6 日時点) を使用している，sonoisa/t5-base-japanese を用いる [1]．

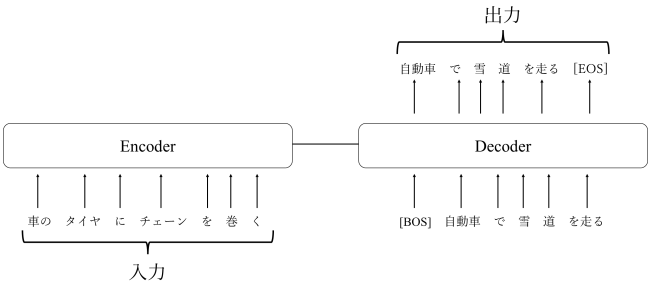

図 2　Encoder-Decoder モデルを用いた方法を表すクエリからの目的抽出.

4.1.2 BERT2BERT

BERT2BERT は，Encoder モデルである BERT（Bidirectional Encoder Representations from Transformers）モデルのパラメータを Decoder にも用いた Encoder-Decoder モデルである．なお，BERT の事前学習は，以下のタスクで実行される．入力した語の 15%を次の 3 つのうちどれかに置き換える．80%を [MASK] に，10%をランダムな語彙に，10%は原文のままに置き換える．その後，置き換えられた単語が何かを予測するタスク (Masked LM)．2 つの文章を入力し入力した文章が連続したものかそうでないものかを予測するタスク (Next Sentence Prediction) により学習を行う [5]．なお，本研究では BERT モデルの事前学習済みモデルとして東北大学の事前学習済みモデルを用いた．

4.2 wikiHow を利用した訓練データの構築

提案手法である文生成モデルのファインチューニングには大量のデータが必要になる．本研究で必要とするデータは何かをする際のやり方に関する知識で，このような知識は一般に手

1：https://huggingface.co/sonoisa/t5-base-japanese（2022 年 11 月 24 日閲覧）

図 3　実際の wikiHow のページ.

続き的知識と呼ばれている．しかし，このような手続き的知識を集めた大規模なデータセットは存在しない．人手でこのようなデータセットを作る手段もあるが，大量に用意することはできない．この問題を解決するために本研究では wikiHow[2]というウェブページに注目した．wikiHow は何かをする際のやり方に関する知識を集めたまとめサイトである．図 3 に実際の wikiHow[3]のページを示す.

4.3　出力の多様化

本研究で用いる文生成モデルでは，1 つの入力に対して複数の生成結果を得ることができるため，この機能を利用する．この理由は，1 つの入力に対して複数の生成を行う場合，最上位の生成結果が正解と類似していなくても，下位に正解と類似する生成結果が存在する可能性が考えられるからである．しかし，本研究で用いる文生成モデルでは，文生成時にビームサーチを用いているため，その性質上確率が高いものから順に抽出される．例えば,「初心者がヨガを学ぶ」,「初心者がヨガを始める」のように類似した文が続けて出力される．この問題を解決するために出力を Maximal Marginal Relevance [4](MMR) を用いて多様化する.

4.3.1　MMR

MMR は式 (1) を用いることで，r 位に順位づける出力結果を決定する．式 (1) 中の D は文集合，s は生成された順位の逆順位，S^{r-1} すでにランキングされた文集合，sim は文間の類似度，$\lambda \in [0, 1]$ である．また，文間の類似度の計算には SentenceBERT を用いた.

$$d^r = \underset{d \in D \setminus S^{r-1}}{\operatorname{argmax}} [\lambda s - (1-\lambda) \max_{d' \in S^{r-1}} \mathrm{sim}(d, d')] \quad (1)$$

4.3.2　SentenceBERT

SentenceBERT は BERT モデルをファインチューニングしたモデルである．このモデルは 2 つの類似した文を BERT モデルにそれぞれ入力したとき，モデルから出力されるベクトルも類似するようにファインチューニングされている.

本研究では SentenceBERT の学習済みモデルとして sonoisa/sentence-bert-base-ja-mean-tokens-v2[4]を用いた．また，類似度の計算はコサイン類似度で算出した.

表 1　wikiHow より作成したデータの具体例.

見出し	記事タイトル
スイカの苗を植える	スイカを栽培する
地面を覆うツタの除去	ツタを取り除く
木を覆うツタの除去	ツタを取り除く
酔い止め薬を服用する	乗り物酔いに対処する
電圧計でバッテリーを点検する	自動車の点検

5　評価実験

提案手法の有効性を検証するために評価実験を行った．まず，本研究で用いるデータセットについて説明し，精度評価について説明する.

5.1　データセット

本節では本研究で用いるデータセットについて説明する．本研究では大きく 2 種類のデータセットを扱う．wikiHow よりクローリングで集めたデータセットと日常生活で入力されると考えられるクエリより独自に作成した非 wikiHow データセットである．非 wikiHow データセットを作成する理由は，wikiHow 上に存在しないクエリを入力した際の出力も得たいたからである.

5.1.1　wikiHow データセット

図 4 に wikiHow の構造を示す．図 4 のように記事のタイトルを出力，複数の解決方法を入力としてデータを作成することで，手続き的知識源として wikiHow を用いることができる．図 3 に示すページを例に説明すると,「姿勢を良くする」を出力とし,「いい姿勢で立つ，歩く」を入力とする.

本研究では wikiHow データセット作成のためのデータ収集には，クローリング[5]でデータの収集を行った．クローリングでは合計 1,685 件のページよりデータの収集を行った．クローリングで得られた wikiHow のページより，記事タイトルと見出しを抽出し，図 4 のようにペアを作る．この操作を収集したすべてのページで実行する．なお，収集したデータについて，図 4 のような構造でないページについてはデータセットから除外した．最終的に，4,977 件の入力と出力のペアを作成した．また，作成したデータの例を表 1 に示す.

5.1.2　自作データセット

本研究で取り組む課題は wikiHow にあるものがすべてではなく，日常生活で行う検索では必ずしも wikiHow に答えが存

2：https://www.wikihow.jp/ （2023 年 1 月 6 日閲覧）
3：https://www.wikihow.jp/姿勢を良くする （2023 年 2 月 23 日閲覧）
4：https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2 （2023 年 1 月 9 日閲覧）
5：https://book.st-hakky.com/docs/text-summary-dataset-wikihow/（2023 年 2 月 23 日最終閲覧）

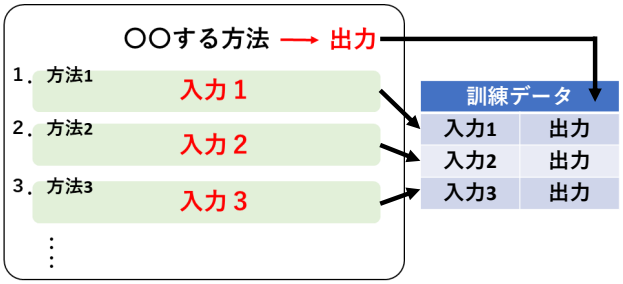

図 4　wikiHow の構造を利用した訓練データ構築.

表 2　方法を表すクエリとその目的.

方法を表すクエリ	考えられる目的
アイロンのかけ方	服のしわを伸ばす方法
お風呂のお湯を洗濯に使う	水を節約する
携帯のバッテリーを交換する	充電を長持ちさせる方法
小顔ローラーの使い方	顔を小さく見せる方法
サボテンを育てる	観葉植物を部屋に置く

在するわけではない．よって wikiHow より作成したデータだけでなく，日常生活で実際に検索されると考えられるクエリを用意した．作成したデータの例を表 2 に示す．用意したデータは図 2 のように，方法を表すクエリと考えられる目的がペアになっているデータである．このようなデータを 100 件用意した．このデータセットを用いて評価することで，モデルの汎用性を評価する．

本研究で作成した，方法を表すクエリとその目的の一部を表 2 に示す.

5.2　比 較 手 法

本研究では，方法を表すクエリから目的を抽出するための手法として T5 と BERT2BERT の 2 種類の文生成モデルを用いる．また，これらのモデルでは 4.3 節で述べたように，1 つの入力に対して複数の生成を行うと類似する生成結果が連続するため，生成結果の多様化によりこれを解消する．以上より，本研究で比較する手法は以下の手法である.

- T5（多様化あり）
- T5（多様化なし）
- BERT2BERT（多様化あり）
- BERT2BERT（多様化なし）

これらの手法を 5.3 節で説明する評価基準に従って比較する.

5.3　評 価 尺 度

本研究では評価指標として正解との類似度を自動で数値化する BERTScore による自動評価と，入力と生成結果が方法と目的の関係にあるかどうかを人手で判別し P@10，DCG で評価する 2 種類の評価方法で評価を行った.

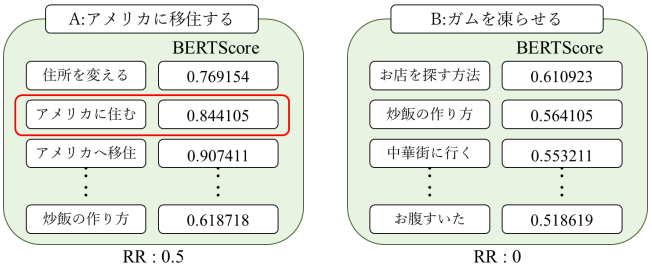

図 5　逆順位を用いた評価.

5.3.1　自 動 評 価

a）　BERTScore

生成結果を自動で評価するために要約や翻訳の自動評価で用いられる BERTScore を用いて正解データとの類似度を算出した．正解の目的と，出力が類似しているほど高いスコアになると考えられる．BERTScore は事前学習された BERT モデルから得られるベクトル表現を利用して，文章間の類似度を計算する [13]．本研究では BERT の学習済みモデルに，bert-base-multilingual-cased を使用している．訓練データとして 104 言語の Wikipedia のデータが使われている [6].

b）　上位 1 件の出力に対する自動評価

まず，モデルに 1 件生成させる．この生成結果と正解との類似度を BERTScore により類似度を数値化し，その平均値を計算する.

c）　ランキング結果の自動評価

図 5 に MRR を用いた評価について示す．この評価ではしきい値とする BERTScore を設定し，生成結果を評価した BERTScore がしきい値を超えている場合，入力に対する正解とみなし，その RR を算出する．図 5 ではしきい値となる BERTScore が 0.8 の場合について示す．なお，すべての BERTScore を参照した結果，しきい値に満たない BERTScore のみの場合はその RR を 0 とした．この評価で得られた RR の平均を計算し，MRR としてそれぞれの比較手法で計算した.

5.3.2　人手での評価

BERTScore による自動評価では正解と類似する生成結果が高く評価されてしまう．学習時には 1 対 1 の関係で，基本的に正解とする出力は 1 つである．しかし，実際の検索では入力した方法に対して様々な目的を考えることができる．例えば,「自動車のタイヤ　チェーンを巻く」の目的として考えられるのは「雪道を自動車で走る」のほかに「砂浜を自動車で走る」も考えられる．よって，このような生成結果を評価するために人手による評価も行う.

wikiHow データセットより作成したテストデータには，本研究で対象としない入力と正解の関係を持った文のペアが存在す

6：https://huggingface.co/bert-base-multilingual-cased（2022 年 12 月 5 日閲覧）

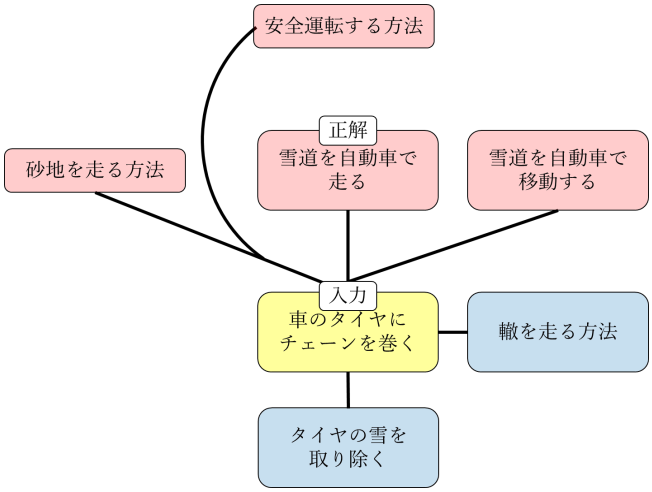

図 6　予想される生成結果.

表 3　自動評価で用いたデータの量.

wikiHow データセット			非 wikiHow データセット	合計
訓練	検証	テスト	テスト	
4,480	249	248	100	5,077

表 4　人手の評価で用いたデータの量.

wikiHow データセット	非 wikiHow データセット	合計
30	30	60

る．例えば，入力が「はじめに」，正解が「カホンを作る」のようなデータが存在する．人手による評価ではこのようなデータは評価しないため，評価するテストデータの中からクエリを選んだ．さらに，MMR による多様化前後それぞれの生成結果上位 10 件を選び評価した．

評価基準は図 6 に従って評価する．図 6 は，図中の黄色のクエリを入力した場合，予想される生成パターンを示したものである．図 6 中の赤色で塗られた生成結果には得点 1 を与え，青色で塗られた生成結果の場合は得点 0 を与える．

図 6 中の青色で塗られたクエリにおいて得点 0 を与える理由について説明する．図中の「轍を走る方法」は入力に対する目的ではなく,「タイヤの雪を取り除く」は入力の手順を出力しており，これもまた入力に対する目的ではないからである．

図 6 の評価基準に従って，入力と生成結果が方法と目的の関係にあるかどうかを人手で判別した後，P@10，DCG，を計算しその平均値を評価した．

5.4　実 験 設 定

5.4.1　使用するデータセット

4.2 節にて説明した wikiHow より作成したデータセットと独自に用意した非 wikiHow データセットについて，その総量と訓練，検証，テストデータの内訳を表 3 にまとめた．

モデルの学習には表 4 の訓練，検証を用いて学習し，自動評価には表 4 のテストを用いて評価した．また，人手の評価では表 4 のテストデータより，それぞれ 30 件ずつ抜粋し評価した．

表 5　ファインチューニングの学習条件.

項目	T5	BERT2BERT
バッチサイズ	256	256
学習率	3.0×10^{-4}	2.0×10^{-6}
最適化関数	AdamWeightDecay [6]	AdamWeightDecay [6]

5.4.2　学 習 条 件

生成モデルの学習条件について説明する．T5 と BERT2BERT の事前学習済みモデルとしてそれぞれ，sonoisa/t5-base-japanese, 東北大学の事前学習済みモデルを用いた．

本研究では早期終了により学習を終了した．早期終了を用いた学習を終了する条件は 2 種類ある．1 つ目はあらかじめ学習を繰り返す最大値を決定し，最大値まで損失が減少し続けた場合．2 つ目は学習を進める過程で損失が減少しなくなったとき，学習を終了する場合である．なお，早期終了の判断は検証データを用いて行う．本研究では 2 つ目の終了条件において，10 回学習する間に損失が減少しなければ学習を終了した．

その他の学習条件を表 5 に示す．

5.5　自動評価による評価結果

本節では，まずファインチューニングを行った生成モデルに対して，2 種類のテストデータを入力した際の生成結果を示す．その後，自動評価の結果を示す．

5.5.1　生 成 結 果

表 6 にテストデータを入力した際の生成結果の一部を示す．この表に示している生成結果は，多様化前の生成結果における 1 番目に生成された結果を選んだ．表のように生成結果が正解と意味的に似ているものや，正解と完全に一致している結果が存在した．

5.5.2　上位 1 件の出力に対する自動評価結果

BERTScore で評価した結果を表 7 に示す．表に示す値は各テストデータを入力した際の出力結果と正解との類似度を BERTScore により数値化し，すべての BERTScore の平均値を計算したものである．表よりテストデータによる評価では，BERT2BERT よりも T5 の方が平均の BERTScore が高くなった．

5.5.3　ランキング結果の自動評価結果

5.3.1 節で説明した逆順位による評価結果について表 9 に示す．本稿では MMR の多様化前の結果を示す．表よりしきい値を高くするほど T5 の方が BERT2BERT と比較して高い値になった．

5.6　人手での評価による評価結果

本節では，まずファインチューニングを行った生成モデルに対して，2 種類のテストデータを入力した際の生成結果を示す．その後，自動評価の結果を示す．

表 11 と表 12 にテストデータを入力した際の生成結果を人手で評価した結果を示す．表 11 の数値は 5.3.2 節で述べた基準で生成結果を評価した後，DCG の平均を計算したものである．また，表 12 は表 11 と同様の基準で評価した後，P@10 の平均

表 6　テストデータの生成結果.

入力	BERT2BERT	T5	正解
テントを買う	夏休み中の暇をつぶす	キャンプをする	キャンプに必要なものを買う
アメリカに移住する	児童書を書く	カナダへ移住する	アメリカで仕事を見つける
バターをめん棒で伸ばす	バターを柔らかくする	バターを柔らかくする	バターを柔らかくする
牛乳を飲む方法	牛乳を加熱する	牛乳からミルクを作る	牛乳嫌いを克服する方法
無駄遣いを減らす方法	節約して生きる	お金を管理する	貯金する方法

表 7　wikiHow より作成したテストデータを入力した際の BERTScore.

BERT2BERT	T5
0.808	0.826

表 8　非 wikiHow データセットの BERTScore.

BERT2BERT	T5
0.733	0.739

表 9　wikiHow データセットの MRR.

	BERT2BERT		T5	
評価手法	MMR 前	MMR 後	MMR 前	MMR 後
Top100($\theta = 0.80$)	0.498	0.482	0.554	0.530
Top100($\theta = 0.85$)	0.396	0.381	0.443	0.423
Top100($\theta = 0.90$)	0.323	0.301	0.400	0.383

表 10　非 wikiHow データセットの MRR.

	BERT2BERT		T5	
評価手法	MMR 前	MMR 後	MMR 前	MMR 後
Top100($\theta = 0.80$)	0.177	0.161	0.215	0.208
Top100($\theta = 0.85$)	0.0608	0.0477	0.0900	0.0806
Top100($\theta = 0.90$)	0.0209	0.0200	0.0175	0.0150

表 11　人手の評価結果（DCG）.

	BERT2BERT		T5	
データセット	MMR 前	MMR 後	MMR 前	MMR 後
wikiHow	1.21	0.810	1.28	0.998
非 wikiHow	0.203	0.156	0.290	0.239

表 12　人手の評価結果（P@10）.

	BERT2BERT		T5	
データセット	MMR 前	MMR 後	MMR 前	MMR 後
wikiHow	0.210	0.117	0.223	0.160
非 wikiHow	0.0300	0.0200	0.0533	0.0467

を計算したものである．表 11 より，BERT2BERT，T5 ともに多様化前の方が上位に適合文書が存在していることがわかる．また，表 12 より，BERT2BERT，T5 ともに多様化前の方が適合文書数が上位にランキングされていることがわかる．これらの結果より，本研究で比較した手法の中で T5（多様化なし）が方法を表すクエリから目的を抽出することに有効であるといえる．

6　まとめと今後の課題

本研究では方法を表すクエリが入力された際に，クエリを入力したユーザの目的を抽出する問題に取り組んだ．目的達成のために必要な方法を検索する場合，様々な方法を検討するためにユーザが入力した方法から目的を抽出する必要があると仮説を立て，方法を表すクエリからの目的抽出を提案した．4.1 節では提案手法について説明し，5.1 節では本研究で扱うデータセットについて説明した．このデータセットを用いて T5，BERT2BERT のファインチューニングを行うことで，方法を表すクエリから目的を生成するモデルを構築した．また，5.5 節の実験結果では 5 で説明した 4 種類の提案手法，T5（多様化あり），T5（多様化なし），BERT2BERT（多様化あり），BERT2BERT（多様化なし）の評価を行った．自動評価及び人手評価の 2 種類の評価を行い，人手の評価ではそれぞれのデータセットにおいて 30 件ずつ評価した．その結果，自動評価では T5 が最も高い BERTScore，を達成した．また，多様化前後における MRR の結果より，ランキング結果の評価より，多様化しない方が高い MRR を達成した．人手評価では T5 が最も高い DCG@10，P@10 を達成した．また，こちらの評価でも多様化しない方が高い DCG@10，P@10 を達成した．

今後の課題として，非 wikiHow データセットによる生成で精度を向上させることが課題である．

謝辞　本研究は JSPS 科学研究費助成事業 JP21H03904，JP21H03775，JP22H03905，による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] Hassan Ahmed and White Ryen, W. Task tours: helping users tackle complex search tasks. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 1885–1889, 2012.

[2] Sumit Bhatia, Debapriyo Majumdar, and Prasenjit Mitra. Query suggestions in the absence of query logs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 795–804, 2011.

[3] Paolo Boldi, Francesco Bonchi, Carlos Castillo, and Sebastiano Vigna. From “dango” to “japanese cakes”: Query reformulation models and patterns. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pp. 183–190, 2009.

[4] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual inter-*

national ACM SIGIR conference on Research and development in information retrieval, pp. 335–336, 1998.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[7] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Sumio Fujita, Akira Tajima, and Katsumi Tanaka. Mining alternative actions from community Q&A corpus for task-oriented web search. In *Proceeding of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 607–614, 2017.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, Vol. 21, No. 140, pp. 1–67, 2020.

[9] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 264–280, 2020.

[10] Takehiro Yamamoto, Tetsuya Sakai, Mayu Iwata, Chen Yu, Ji-Rong Wen, and Katsumi Tanaka. The wisdom of advertisers: mining subgoals via query clustering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 505–514, 2012.

[11] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pp. 418–428, 2020.

[12] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. Query understanding via intent description generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1823–1832, 2020.

[13] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019.

[14] 加藤龍, 大島裕明, 山本岳洋, 加藤誠, 田中克己. タスクの汎化と特化に着目した web からのタスク検索. 第 6 回データ工学と情報マネジメントに関するフォーラム, C1–6, 2014.

[15] 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸. 関連単語抽出アルゴリズムを用いた web 検索クエリの生成. 電子情報通信学会論文誌 D, Vol. 92, No. 3, pp. 281–292, 2009.

[16] 田麥節治, 赤星祐平, 是津耕司, 木俵豊, 田中克己. Web サービスを対象とした検索意図推定とその応用. 第 4 回データ工学と情報マネジメントに関するフォーラム, F4–3, 2012.