クエリ指向要約におけるクエリと要約の統合的な生成

服部 翔　Youmi Ma　岡崎 直観
東京工業大学
{kakeru.hattori@nlp.,youmi.ma@nlp.,okazaki@}c.titech.ac.jp

概要

クエリ指向要約は，ある特定のクエリ（質問）に対して要約を生成するタスクである．しかし実際には，ユーザが未知の文書に対してクエリを考えることは難しい．そこで本研究では，クエリ指向要約の発展形として，クエリも含めて自動生成する**クエリ推薦付き要約**を提案し，具体的なタスクと評価方法の設計を行う．次に，設計したタスクに対し，文書の特定の部分（スパン）からクエリ・要約を生成する手法などを提案する．実験では，スパンやそれに代わる機構が多様なクエリの生成に有効だが，スパンの予測精度が全体のボトルネックとなることを確認する．また，文書の不要な部分を事前に予測して除去する手法も提案し，その有効性を実証する．

1　はじめに

自動要約のタスク設定の 1 つとして，**クエリ指向要約**（Query-Focused Summarization）[1] がある．クエリ指向要約では，ある特定のトピック・内容に関するクエリ（質問）が与えられ，そのクエリに対する要約を生成する．ユーザ（読者）の興味や疑問に合わせてクエリを与えることで，議事録のような長い文書に対しても，ユーザが真に求める情報を要約として的確に提供できる．

一方，クエリ指向要約の既存研究では，クエリが事前に与えられることを前提としている．しかし実際には，ユーザが内容を知らない文書に対してクエリを考えることは難しい．もし，システムが与えられた文書に対して適切なクエリを推薦できれば，この課題を解決できる．例えば，ウェブ検索エンジンなどではクエリ推薦機能の提供例[1)]があるなど，社会的な需要も高いと見込まれる[2)]．

そこで本稿では，クエリ指向要約を発展させた研究として，クエリの推薦と対応する要約の統合的な生成，すなわち**クエリ推薦付き要約**を提案する．クエリ推薦付き要約では，与えられた文書の重要なトピックや内容の把握に役立つクエリと，対応する要約のペアを複数生成し，それら全体をユーザに提供する（図 1）．入力としてクエリが与えられる既存のクエリ指向要約とは異なり，クエリ推薦付き要約では，クエリも含めて全て自動生成する．

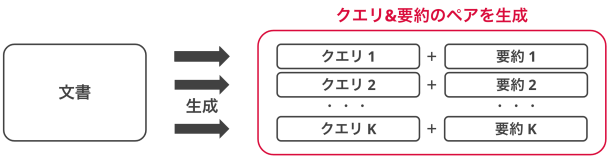

図 1　クエリ推薦付き要約

クエリ推薦付き要約は，ユーザが興味を持ちそうなクエリと，その特定のトピックに詳しく踏み込んだ内容の要約のペアを複数生成するため，ユーザが真に求める情報や，文書の重要な内容を網羅すると期待される．また，要約全体がクエリと要約のペアの集合という形で構造化されるとともに，ユーザが自分の興味に合わせて要約の一部を選択的に読むことができるため，可読性の観点でもユーザ・エクスペリエンスが向上する．

本研究の主な貢献は以下の通りである．

- クエリ推薦付き要約の自動生成タスクを定義し，生成したクエリと要約の集合全体を適切に評価するための評価指標を設計・選定した．
- 文書の特定の区間（スパン）に着目してクエリ推薦付き要約を生成する手法などを提案した．実験では，スパンやそれに代わる機構が多様なクエリ生成に重要であることを確認した．
- スパンを明示的に指定する手法では，スパンの予測精度の低さがクエリ・要約の生成精度のボトルネックとなっていることを確認した．
- 事前に文書の不要な部分を予測し，除去してからスパンを指定して生成することで，クエリ・要約の生成精度が向上することを確認した．

1)　海外の Google 検索では People Also Ask と呼ばれる機能に対応する．日本でも「他の人はこちらも質問」として，近年表示されるようになった．
2)　いわゆる FAQ なども，身近なクエリ推薦の一種である．

2　本研究で取り組むタスク

2.1　使用するデータセット（QMSum）

本研究では，データセットとして **QMSum** [2] を使用する．QMSum はマルチドメインの議事録を対象にした，クエリ指向要約向けの英語のデータセットであり，232 件の会議と 1,802 件のクエリ・要約の組を収録している．各会議に対する複数のクエリは，その会議の重要なトピックをカバーしているとされる．また，各クエリに関連するスパン（文書の特定の一部分）がターン[3)]単位で注釈付けされている．

2.2　タスクの目的

本研究の目的は，1 節で提案したクエリ推薦付き要約として適切なクエリ・要約の組の集合を生成することである．その際，クエリ・要約の組の集合が満たすべき性質として，以下の 2 つを考慮したい．

- 多様なクエリで構成されている．
- 文書の重要な内容を網羅的にカバーしている．

「多様なクエリ」とはクエリ同士の語彙的な重複や意味的な重複が少ないことを意味する．一方，「文書の重要な内容を網羅的にカバー」するというのは，本質的には一意に解が定まらない難しい問題であるが，今回は用いるデータセット（QMSum）に収録されているクエリが，重要なトピックをカバーする良いクエリであるとされている [2] ことから，データセットのクエリと要約の組の集合をそのまま全て再現することを本研究の目標とする．

2.3　タスクの定義

前述の目的を踏まえ，具体的な生成タスクを次のように定義・設計する．ある文書 D に対して，K 個のクエリ・要約の組で構成される集合 $S = \{(q_1, a_1), (q_2, a_2), \cdots, (q_K, a_K)\}$ を生成する．1 つの文書 D に対し，生成目標はデータセットで用いられた全てのクエリ・要約の組とする．また，その個数 K は定数として事前に与えるものとする．

2.4　タスクの評価方法

前述の目的を踏まえ，生成したクエリ・要約の組の集合を，(i) 多様性と (ii) 内容一致性の 2 つの側面で評価する．具体的な評価指標は 4.1 節で記述する．

3)　同一の話者によるひとまとまりの発話．

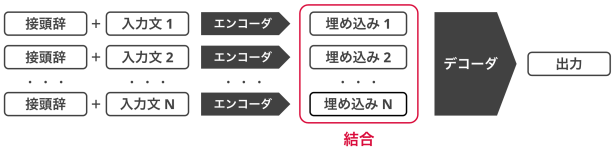

図 2　SegEnc（Fusion-in-Decoder）のアーキテクチャ

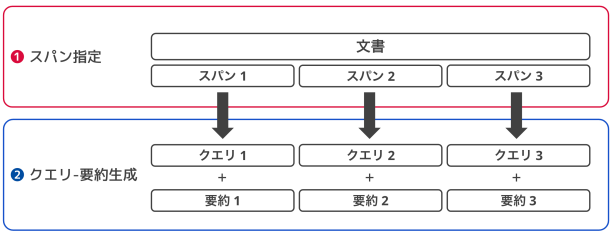

図 3　スパンを指定する手法による生成

3　クエリ推薦付き要約の自動生成

文書からクエリや要約を生成する際には，Transformer [3] をベースとした seq2seq モデルを用いる．本研究では特に，Fusion-in-Decoder [4, 5, 6] のアーキテクチャを取り入れ，QMSum のクエリ指向要約タスクで最先端の性能を達成している SegEnc [7] を主に採用する．SegEnc は図 2 のように入力文を一定の長さでチャンクとして分割し，別々のエンコーダに入力したものをデコーダ側で結合する．また，シーンに応じて各チャンク共通の適切な接頭辞を入力に追加することで効果的な文生成を実現している．本稿では，SegEnc を用いたクエリや要約の生成手法として，以下の 2 つを検討する．

3.1　文書全体を入力する手法

まず，クエリ生成器に文書全体と生成するクエリ数[4)]を入力として与え，クエリの集合を区切り文字で繋いだ文字列として一括で生成する．その後，要約生成器に文書全体と生成したクエリの 1 つを入力として与え，対応する要約を個別に生成する．

3.2　スパンを指定する手法

文書全体ではなく，特定の区間（スパン）を入力として，クエリや要約を個別に生成する．図 3 に示すように，同じ文書内でも毎回異なるスパンを与え，入力情報を差別化することで，同一の文書に対して複数の異なるクエリ・要約を容易に生成できるようになる．また，クエリ・要約生成の際に明示的にスパンを指定するため，生成結果に対する説明性の向上も期待できる．

4)　ただし，生成するクエリの個数は完全には制御できない．

1 つの文書に対して複数のクエリ・要約の組を生成する際に，互いに異なるスパンをどのように指定するかは難しい問題であるが，本稿ではシンプルな方法として，**均等割り当て**を検討する．これは，図 3 のように文書全体にスパンを均等に割り当てるというものである．単純ではあるが，文書全体をカバーするスパンから，文書の内容を広くカバーするクエリ・要約の生成が期待できる．

ターン・フィルタ　前述の均等割り当てに対する工夫として，ターン・フィルタを検討する．これは，文書内のあまり重要ではないターンを事前に予測し，文書から除去することで，スパンの予測精度向上を目指すものである．具体的には，データセットで文書にアノテートされている複数のクエリに対応するスパンに 1 つでも含まれているターンを 1，そうでないターンを 0 とラベル付けした教師データを準備し，文書の各ターンを二値分類する．その後，0 とラベル予測したターンの一部を除去して，スパン指定を行う．実際には，BIO 記法 [8] を採用し，単語ごとに BIO のラベルを予測する分類器を構築し，ターンに含まれる全ての単語の予測ラベルの多数決でターンのラベルを決定する．

4　実験

提案手法の性能を検証するため，クエリ推薦付き要約の自動生成を QMSum [2] 上で実験した．基本的に，クエリ・要約は SegEnc で「クエリ→要約」の順で生成し，ターン・フィルタは無効とした．

スパン指定の有無　文書全体を入力する手法と均等割り当てでスパンを指定する手法との性能を比較し，スパン指定がクエリ・要約の生成に与える影響を検証したい．また本項目では，SegEnc に加えて LED-large[9] をクエリ生成器とした場合も実験し，モデル構造とクエリの多様性の関係についても検証した．

ターン・フィルタの有無　均等割り当てにおけるターン・フィルタの有効性を検証したい．また，ターン・フィルタを有効・無効にした場合に加え，データセットの正解のスパンを与えた場合も比較し，スパンの予測精度がクエリ・要約の性能に与える影響について検証した．

クエリ・要約の生成方法　スパンを指定する手法において，クエリ・要約の生成順序や方法の違いがもたらす影響について調査したい．具体的には「クエリ→要約」「要約→クエリ」「クエリ＋要約」「要約＋クエリ」の 4 通りを検証した．最初の 2 つは個別の生成器でクエリ・要約を順に生成するが，生成順序を入れ替える．残りの 2 つは 1 つの生成器でクエリ・要約の組を一括で生成するが，生成目標の順序を入れ替える．

4.1　評価指標

本稿では，スパンの重複度・クエリの多様性・クエリと要約の内容一致性を文書単位で評価する．

Span-F1　スパンをターン番号の配列とみなし，正解のスパン K 個と生成したスパン K 個との間の全てのペア K^2 個について，重複度を F1 スコアで評価する．そしてスコアが高いペアの組み合わせを最大二部マッチングで決定し，それらのペアの F1 スコアの平均値を最終スコアとする．

distinct-n　生成したクエリ K 個の間の語彙多様性を評価する指標として Li ら [10] が使用した distinct-n を導入する．distinct-n は K 個のクエリに含まれる [n-gram の種類数] / [n-gram の総数] で定義される．本稿では {1,2}-gram について算出する．

ROUGE　正解および生成したクエリ・要約との一致度を ROUGE-{1,2,L(sum)} [11] で評価する．Span-F1 と同様の要領でペアを決定し[5]，それらのスコアの平均値を報告する．なおクエリについては，QMSum のスキーマ部分の一致を過剰に評価しないようにするため，事前にルールに基づいてスキーマ部分を除去し，ROUGE を計算する（付録 B 参照）．

4.2　実験結果

実験結果を表 1 と 2 に示す．提案手法は，適切なスパン指定もしくはモデル構造によって，1 つの文書に対して多様なクエリを提示し，それらに対する要約を生成できた．また，クエリ・要約の内容に関する精度も一定の水準に達したが，こちらは今後より改善の余地があると思われる．詳細な結果を以降で分析する．なお，具体的な生成例は，本研究で制作したデモサイト[6]上で公開している．

スパン指定の有無　表 1 の 2〜5 行目に示すように，スパンを指定する手法での distinct-n は，クエリ生成器が SegEnc・LED-large いずれの場合も，データセットの正解のクエリ（0.49, 0.73）には及ばないが，一定水準の語彙多様性を達成した．一方で文書全体を入力する手法での distinct-n は，SegEnc の場

5）　マッチング基準は ROUGE-{1,2,L(sum)} の平均値とする．
6）　https://qss-demo.vercel.app/

表 1　スパン指定およびターン・フィルタの有無による性能の比較

質問生成器	スパン	フィルタ	Span-F1	distinct-n		ROUGE（クエリ）			ROUGE（要約）		
				1	2	1	2	L	1	2	L
SegEnc	正解	-	*100.00*	*0.4061*	*0.6055*	*35.38*	*15.29*	*32.83*	*37.43*	*13.32*	*32.64*
LED-large	なし	-	-	0.2781	0.3874	29.47	9.83	28.08	30.91	8.10	27.16
SegEnc	なし	-	-	**0.5431**	**0.8759**	**32.05**	11.78	**29.85**	31.91	8.53	27.79
LED-large	均等	なし	33.55	0.3502	0.5313	30.05	10.66	28.93	32.23	8.85	28.22
SegEnc	均等	なし	33.55	0.3993	0.5960	31.11	11.96	29.12	32.13	8.84	28.13
SegEnc	均等	あり	**34.22**	0.3840	0.5692	31.46	**12.80**	29.84	**32.77**	**9.18**	**28.72**

表 2　クエリ・要約の生成方法による性能の比較

質問生成器	スパン	フィルタ	Span-F1	distinct-n		ROUGE（クエリ）			ROUGE（要約）		
				1	2	1	2	L	1	2	L
質問→回答	均等	なし	33.55	0.3993	0.5960	31.11	11.96	29.12	32.13	8.84	28.13
回答→質問	均等	なし	33.55	**0.4468**	**0.6696**	31.56	12.18	29.48	32.11	8.84	28.00
質問+回答	均等	なし	33.55	0.3808	0.5597	**32.10**	**12.81**	**30.57**	31.80	8.69	27.81
回答+質問	均等	なし	33.55	0.4400	0.6607	31.13	11.78	29.18	**32.39**	**9.01**	**28.18**

合は極端に高く，正解のクエリすら上回ったのに対し，LED-large の場合は極端に低く，実際にはほぼ同じクエリの反復となっていた．SegEnc は 3 節で述べたようなモデル構造によって，明示的なスパンが無くても，実質的に均等割り当てに近い方法で文書の一部分に着目しているといえるが，LED-large には同様の機構が存在しない．これらの結果から，文書の特定の区間に着目する何らかの機構が多様なクエリの生成に重要ではないかと考えられる．ただし，クエリ・要約の ROUGE にはいずれも明確な優劣関係が確認できなかったことから，多様なクエリであっても内容的に良いクエリ・要約の組が生成できているとは限らない．多様性と内容面での精度向上の両立は今後の課題である．

ターン・フィルタの有効性　表 1 の 5, 6 行目に示すように，ターン・フィルタを有効にした場合は，クエリの語彙多様性はわずかに低下したものの，スパン・クエリ・要約の予測・生成精度は全面的に向上し，本稿の提案手法の中では概ね最良のスコアが得られた．スパンの予測精度向上に伴い，クエリ・要約を生成する精度も改善されたと考えられ，ターン・フィルタの有効性が確認できた．ただし，表 1 の 1 行目に示すように，正解のスパンを与えた場合のクエリ・要約の ROUGE は他よりも大幅に高く，スパン予測の精度の低さがクエリ・要約生成のボトルネックであることが示唆される．より適切なスパンの推定手法，およびスパンを指定しない生成手法を構築することは，今後の課題である．

クエリ・要約の生成方法の影響　表 2 に示すように，クエリの distinct-n はクエリを先に生成する方法（「クエリ→要約」「クエリ+要約」）と要約を先に生成する方法（「要約→クエリ」「要約+クエリ」）で大幅な差があり，後者の方が高い結果となった．また，クエリの ROUGE はいずれも「クエリ+要約」が最も高い結果となった．このように生成方法はクエリの生成に影響を与え，特に生成順序がクエリの語彙多様性に大きな影響をもたらすことがわかった．一方で，要約の ROUGE は手法間での差が小さく，クエリの ROUGE との相関も確認できなかった．

5　おわりに

本稿では，クエリ指向要約の発展形として，クエリも含めて自動生成するクエリ推薦付き要約を提案し，具体的なタスクと評価方法の設計を行った．次に，設計したタスクに対して，文書全体を入力する手法およびスパンを指定する手法を提案し，実験では文書の特定の区間に着目する機構が多様なクエリ生成に重要であることがわかった．また，スパンを指定する手法では，文書の不要な部分を事前に除去することで，スパンの予測精度やクエリ・要約の生成精度が向上した．ただし，スパンの予測精度は依然として低く，クエリ・要約生成のボトルネックとなることが示唆された．さらに，クエリと要約の生成方法について検証し，主に生成の順序がクエリの語彙多様性に大きな影響を与えることがわかった．

今後は，クエリ・要約の組の多様性と内容面の向上の両立のため，より高度なスパンの推定手法，およびスパンを指定しない生成手法の構築を検討したい．また，日本語を含めたデータセットの作成，より良いクエリ推薦付き要約の評価指標，要約の可読性の定量的な評価方法についても研究を進めてゆきたい．

謝辞

本研究成果は，国立研究開発法人情報通信研究機構（NICT）の委託研究「自動翻訳の精度向上のためのマルチモーダル情報の外部制御可能なモデリングの研究開発」（課題 225）により得られたものです．

参考文献

[1] Hoa Trang Dang. Overview of DUC 2005. In **Proceedings of the document understanding conference, volume 2005**, pp. 1–12, 2005.

[2] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. QMSum: A new benchmark for query-based multi-domain meeting summarization. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5905–5921, Online, June 2021. Association for Computational Linguistics.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[4] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering, 2020.

[5] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering, 2020.

[6] Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Sebastian Riedel, and Edouard Grave. A memory efficient baseline for open domain question answering, 2020.

[7] Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. Exploring neural models for query-focused summarization. In **Findings of the Association for Computational Linguistics: NAACL 2022**, pp. 1455–1468, Seattle, United States, July 2022. Association for Computational Linguistics.

[8] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In **Third Workshop on Very Large Corpora**, 1995.

[9] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. **arXiv preprint arXiv:2004.05150**, 2020.

[10] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[11] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics.

A　より詳細な実験設定

全ての実験では，1 つの実験条件（生成モデル・方法・順序，スパンの指定方法，ターン・フィルタの有無）に対して，seed 値を変更して 5 回ずつ実験を行い，それらの平均値を結果として報告した．また，事前学習済みモデルは HuggingFace[7] で公開されているものを使用し，微調整の際は，Huggingface の examples の実装コード[8] を適宜改変した Python ファイル・スクリプトファイルを使用した．

A.1　ターン・フィルタ

事前学習済みの Longformer-base [9] を微調整してモデルを作成した．QMSum の文書は非常に長いため，一度にモデルに入りきらない場合は，4,096Token ごとに区切って入力を行った．その際，Vig ら [7] の研究を参考に，入力区間は 50%オーバーラップさせ，入力の端付近の予測をなるべく使わないように配慮した．なお，モデルの予測に基づいて文書からターンを除去する割合は最大でも文書全体の 25%とし，残すターンがなるべく連続した区間となるように実装した．また，割り当てるスパンのターン数自体は，同一の文書内で通常の均等割り当ての場合と揃えることで，Span-F1 によるスパンの重複度評価の公平性を保つこととした．このことにより，ターン・フィルタを有効とした場合は，通常の場合と異なり，割り当てたスパン間で区間の重複が発生している．

A.2　クエリ・要約生成器の学習

割り当てたスパンのテキストから SegEnc[7] もしくは LED-large[9] でクエリ・要約を生成した．SegEnc は BART-large[12] をバックボーンモデルとし，チャンクサイズを 512，最大チャンク数を 32，入力のオーバーラップを 50%とした．LED-large は入力長上限が 16,384token のものを使用した．また生成器は，データセットの正解のクエリ・要約および対応する関連スパンからクエリ・要約を予測するように学習させた．なお，学習率は 5e-6，エポック数は 10 とした．

B　クエリ評価時のスキーマ除去

本稿では，生成したクエリを評価する際，QMSum のスキーマ部分の一致を過剰に評価することを防止するため，事前にルールベースでスキーマ部分を除去した上で ROUGE を計算した．具体的には，QMSum の論文 [2] に記載がある 13 種類の Specific Query Schema List および，条件を緩めた独自定義のスキーマを，Python の正規表現ライブラリで検出し，スキーマ部分以外のみを残すように置換した．スキーマとその置換方法の一例を表 3 に示す．

表 3　クエリのスキーマとその置換の一例

置換前	置換後
Summarize the discussion about [内容]?	[内容]
What did [人物] think of [内容]?	[人物] [内容]
What did the (meeting/group/team) [内容]?	[内容]

C　生成結果のデモサイト

本稿で提案した，クエリ推薦付き要約の自動生成モデルの出力例を，ブラウザ上で簡単に確認できる Web サイトを制作した．サイトは以下の URL および QR コードから閲覧できる．

https://qss-demo.vercel.app/

7）https://huggingface.co/
8）https://github.com/huggingface/transformers