知識ベースの検索を伴う Video QA タスクの提案

黒田佑樹 [1]　中島悠太 [2]
[1] 大阪大学大学院情報科学研究科
[2] 大阪大学データビリティフロンティア機構
{y-kuroda@is., n-yuta@}ids.osaka-u.ac.jp

概要

ユーザの過去の発言や行動等の知識をもとに質問に答えられる質問応答システムの登場が期待される．Knowledge-Based Video Question Answering (KBVQA) では，テレビドラマを実世界に見立ててこの問題に取り組んでいる．KBVQA では，質問に対して必要な知識が紐付けられて利用されている．本研究では質問と知識の紐付けを行わず，全体の長大な知識ベースからの知識検索を伴う VideoQA タスクである MAGQA を提案する．また，動画像字幕からの自動知識ベース作成，知識検索及び質問応答を行うベースライン手法を提案し，本タスクの難しさと提案手法の有効性を示す．

1　はじめに

ユーザの過去の発言や行動に関する質問応答システムは高度なアシスタント機能の実現に不可欠であり，例えば，

> *What is the name of the person whom I talked with at one of the past editions of NLP?*

のような質問に回答することが期待される．これは，ユーザの発言や行動の長大なログ（映像）から質問に関連する部分（知識）を見つけ出し，その結果に基づいて回答を生成するタスクと考えることができる．

このタスクに関連するタスクとして，Knowledge-Based Video Question Answering（KBVQA） [1] が提案されている．このタスクでは，複数のエピソードから構成されるテレビドラマを対象として，同様の問題に取り組んでいる．質問に答えるための知識ベースとして，作問者によるアノテーション [1] やインターネット上の要約 [2]，動画像から作成されたシーン説明 [2]，映像に付属する字幕から自動作成した要約 [3] を利用する手法が提案されいる．

これらの手法では，各質問に対して対応する知識がアノテーションとして提供されている，または質問がどのエピソードに関するものか，などのように知識の範囲を限定できるものとしている．例えば，現在 KBVQA において最も精度の良い手法 [3] は，図 1（左）のように予め質問と紐付けられたエピソードの字幕要約のみを用いており，回答を生成する際には要約全文を入力としている．前述のいずれの手法も，学習時には質問に対して必要な知識（または．少なくとも必要な知識の真値）が紐付いていること前提としている．

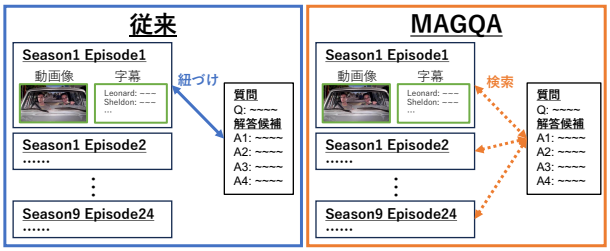

図 1　従来の KBVQA タスクと提案する MAGQA タスクの概要．後者は質問に答えるために必要な情報を紐付けず，検索して得る．

質問と知識の紐付けには，質問にアノテーションを付与する必要がある．ユーザの発言や行動のように長大なログ，すなわち大量の知識が提供されるタスクにおいて，質問に答えるための知識等に関するアノテーションは高コストで，例えばドメイン毎の（もしくは巨大なデータセットに対する）アノテーション付与は難しい．

そこで本稿では，図 1（右）のように，映像から質問応答に必要な知識を検索し，利用するタスクである MAGQA を提案する．このタスクでは，検索対象となる知識を限定するためのアノテーションを利用しない．従って，モデルはまず質問をクエリとして知識ベースを検索し，その結果に基づいて回答を生成する．学習では，検索と回答生成を合わせて最適化する必要がある．本稿では既存データセットである KnowIT VQA [1] を用いて MAGQA を提案す

る．本タスクのベースラインとして字幕というテキストモダリティに注目した手法を提案する．

2 タスク定義

KnowIT VQA はテレビドラマ Big Bang Theory に関する KBVQA のデータセットで，20 分程度の映像からなる 207 のエピソードから構成されており，各エピソードは 12,264 のシーンに分割され，各シーンについて 2 件，合計で 24000 件を超えるサンプルが提供されている．テレビドラマはユーザの発言や行動を記録した映像の代替に適したものであると考える．KnowIT VQA は多肢選択式のタスクとなっており，サンプルは，質問文 q，回答候補 $a = \{a_1, a_2, a_3, a_4\}$，対応するシーン $s = \{v, t\}$（ただし，v と t はぞれぞれ映像と字幕），人手で作成された回答に必要な知識 k，および正解 $c \in a$ から構成される． このタスクでは q，a，s が与えられたときに，全てのサンプルの k の集合である知識ベース $\mathcal{K} = \{k\}$ から適切な知識を選択し，これを元に a から回答を選択する．このタスクでは，学習時に k を知識の真値として利用できる．

MAGQA はこのデータセットを元にしており，人手で作成された知識ベースである $\mathcal{K}$ に代えて，全ての映像の集合 $\mathcal{S} = \{(v, t)\}$ を知識ベースとし，学習時にも k を真値として利用しない．つまり，q，a，s が与えられたときに，$\mathcal{S}$ から必要な知識を見つけ出し，a から回答を選択する．自動音声認識技術により，字幕テキスト t の利用は実用上の問題とならないと考える．このタスクにより，冒頭で示したユーザの発言や行動に基づいて質問に回答するシステムを模擬的に評価する．

3 提案手法

文献 [3] では，提案タスクと同様に $\mathcal{K}$ を利用せず映像から必要な知識を獲得するアプローチが採用されている．具体的には，質問文 q のシーン s が含まれるエピソードを e とすると，そのエピソードの映像 $\mathcal{S}_e \subset \mathcal{S}$ から言語モデルを用いて物語文 n を生成し，q，a，n から回答 $c^\star$ を生成する．

提案するベースライン手法でも文献 [3] と同様に物語文を利用する（図 2）．ただし，質問文とエピソードが紐付いていないことから，全てのシーンについて物語文を生成し，その集合 $\mathcal{N}$ を事前に構築する．さらに，Open Domain QA [4] における密ベクトルによる文書検索 [5] を援用し，質問文 q，回答候補 a をクエリとして，知識検索モジュールで $\mathcal{N}$ から知識の検索を行う．最後に検索して得られた知識 $n^\star$，q，および a を質問応答モジュールへの入力として回答 $c^\star$ を得る．なお，知識検索モジュールと質問応答モジュールの学習は独立して行う．

3.1 知識ベース生成

知識ベース生成モジュールでは，映像集合 $\mathcal{S}$ から知識ベースである物語文集合 $\mathcal{N}$ を生成する．本稿では，文献 [3] と同様に，映像 v と字幕 t のうち t のみを利用する．

字幕は質問への回答に必要な多くの情報を含むが，対話形式であることから代名詞を多用するなど，その一部分のみを見ると検索性や了解性が低く，質問応答の性能に悪影響を及ぼすことが懸念される．そこで提案手法では，要約等の自然言語処理タスクにゼロショットで高い精度を示す LLM である GPT-4 [6] を用いて字幕を物語文へと変換することで，言語モデルがより扱いやすい知識ベースを生成する．

具体的には，図 3 のようなテンプレートを用い，ビデオのシーン s の字幕 t を GPT-4 にプロンプトとして与えることで，物語文 n_s を生成する．付加的な情報として，そのシーンの文脈を明らかにするためにエピソード内で時間的に手前のシーン s' の物語文 $n(s')$，及びそのシーンの場所を表すテキスト p [2] もプロンプトとして入力する．n_s を一定の長さ L に分割したものをそれぞれ n_{sl}（ただし．$n_s = n_{s1} + \cdots$）とすると，知識ベース $\mathcal{N}$ は次式により得られる．

$$\mathcal{N} = \{n_{sl} | s \in \mathcal{S}, l = 1, \dots\} \tag{1}$$

3.2 知識検索

知識ベース $\mathcal{N}$ の検索には，Open Domain QA タスク [4] で採用されている密ベクトルによる文書検索 [5] を用いる．まず質問 q と回答候補 $a = \{a_i | i = 1, \dots, 4\}$ からクエリ

$$w_i = \texttt{[CLS]} + q + \texttt{[SEP]} + a_i + \texttt{[SEP]} \tag{2}$$

を生成する．クエリと知識のエンコーダを E_W，E_K とすると，知識検索モジュールは，w_i と $n \in \mathcal{N}$ の類似度 $\mathrm{sim}(w_i, n) = E_W(w_i)^\top E_K(n)$ を計算し，上位 5 件の知識 $n^\star(w_i) = \{n_j^\star | j = 1, \dots, 5\}$ を出力する．

エンコーダ $E_\mathrm{W}, E_\mathrm{K}$ は，BERT [7] ベースの Spider [8] を初期パラメータとし，DPR [5] に基づく手法を

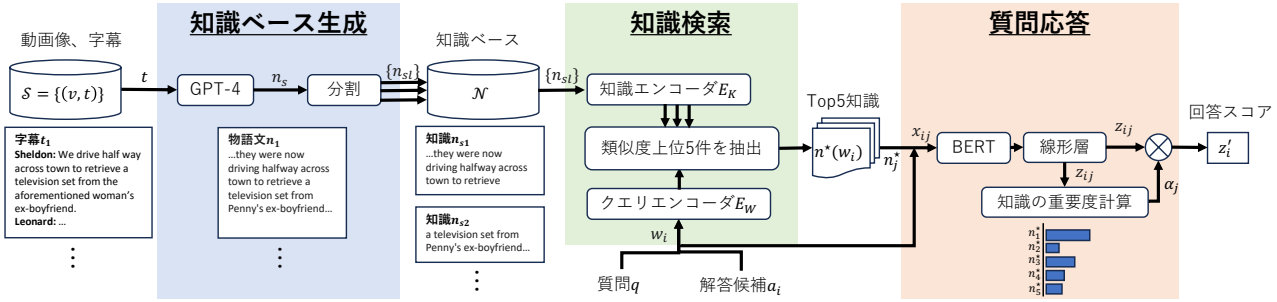

図 2　提案手法の概要.

```
You are a highly skilled AI trained in language comprehension
and summarization.
Please summarize the [Dialogue] from the [Scene] in the TV drama,
following the guidelines below:

1. Resolve co-references and avoid using pronouns as much as possible.
2. Use the same words as in the [Dialogue] whenever possible,
   to avoid losing information.
3. You may use the information from the [Context] if necessary.
[Scene]
-------------------------------------------------------
[Context]
-------------------------------------------------------
[Dialog]
-------------------------------------------------------
```

図 3　物語文生成のプロンプト．[Scene] 下部には場所を表すテキスト．[Context] 下部には前のシーンの物語文．[Dialogue] 下部にはそのシーンの字幕が入る．シーン冒頭では [Context] は空である．

用いて，本タスク向けにファインチューニングする．具体的には，質問 q とその回答に必要な知識 n^+，無関係な知識の集合 $\{n_b^-\}_b$ を用いて，正解 c に対応するクエリ w_c と n^+ の類似度が大きくなるよう次式により対照学習を行う．

$$\ell(w_c, n^+, \{n_b^-\}) = -\log \frac{e^{\mathrm{sim}(w_c, n^+)}}{e^{\mathrm{sim}(w_c, n^+)} + \sum_b e^{\mathrm{sim}(w_c, n_b^-)}} \tag{3}$$

本タスクでは q の回答に必要な知識 n^+ についてのアノテーションが存在しない．そのため，w_c と，対応するエピソード内で w_c との BM25 [9] スコアが最も高い知識を擬似的な n^+ とする．また，$\{n_b^-\}_b$ はバッチ内の別の質問に対する n^+ を用いる．

3.3　質問応答

質問応答では [3] を参考に，検索した知識 $n^\star(w_i)$ を BERT [7] ベースのニューラルネットワークの入力として回答を生成する．まず，質問 q の回答候補 a_i と知識 $n_j^\star \in n^\star(w_i)$ ごとに BERT と線形層を用いたスコア z_{ij} を計算する．BERT への入力 x_{ij} は

$$x_{ij} = \texttt{[CLS]} + n_j^\star + \texttt{[SEP]} + q + \texttt{[SEP]} + a_i + \texttt{[SEP]} \tag{4}$$

で得られる．スコア z_{ij} は次式で与えられる．

$$z_{ij} = \mathbf{W}^\top \mathrm{BERT}(x_{ij}) + \mathbf{b} \tag{5}$$

ただし，$\mathbf{W}$ と $\mathbf{b}$ は線形層のパラメータとする．

次に，質問文 q に対する知識 n_{*j} の重要度を計算する．これには，回答候補 a_i それぞれについて，スコア z_{ij} を最大とする n_{*j} が大きければ重要であると考え，温度パラメータ T を用いて次式により与える．

$$\alpha_j = \mathrm{softmax}\left(\max_i z_{ij}/T\right) \tag{6}$$

最終的な回答候補 a_i のスコア z_i' は．

$$z_i' = \sum_j \alpha_j z_{ij} \tag{7}$$

この z_i' と正解 c とのクロスエントロピーロスにより質問応答モジュールを学習する．

4　実験

実験では，検索範囲と回答精度の関係，および知識検索モジュールの回答精度への影響を調査した．知識の分割数 L は予備実験で良好な性能を示した 80 に固定した．これにより．知識ベースに含まれる知識数は 11,964 となる．また，質問応答モジュールでは，BERT-base-uncased [7] を用いた．

4.1　検索範囲と回答精度の関係

本研究で提案する MAGQA は，学習時に質問への回答に必要な知識の真値が与えられず，推論時にも検索範囲が限定されない点が従来研究 [3] と異なる．例えば文献 [3] の手法では，各質問のシーンが含まれるエピソードが既知であるとし，学習時，推論時のいずれもそのエピソード全体の対話全体の要約を入力している．そこで，知識の検索範囲が精度にどの程度影響を与えるかを評価した．

表 1 検索範囲を限定した場合の回答精度の比較．平均知識数は各検索範囲が含む知識数の平均．

検索範囲	平均知識数	精度
全体	11964	0.663
シーズン	1329	0.717
エピソード	58	**0.772**
シーン	4	0.712

実験では，検索対象の知識ベースをシーズン単位[1)]，エピソード単位，シーン単位で限定する．すなわち，ある質問文 q が紐付けられたシーンを s，s が含まれるエピソードのシーン集合を $\mathrm{Epi}(s)$，s が含まれるシーズンのシーン集合を $\mathrm{Sea}(s)$ で表すと，検索範囲をシーズン単位とする場合には知識ベースを $\mathcal{N}_{\mathrm{Season}}(s) = \{n_{s'l} | s' \in \mathrm{Sea}(s), l = 1, \ldots, \}$，エピソード単位とする場合には $\mathcal{N}_{\mathrm{Episode}}(s) = \{n_{s'l} | s' \in \mathrm{Epi}(s), l = 1, \ldots, \}$，シーン単位とする場合には $\mathcal{N}_{\mathrm{Scene}}(s) = \{n_{sl} | l = 1, \ldots, \}$ とする．知識検索モジュールの学習時には，常に全シーンを含む知識ベース $\mathcal{N}$ を利用し，質問応答モジュールの学習，および推論時は，知識検索モジュールで限定された知識ベースから得られる知識を利用した．

表 1 にそれぞれの構成での精度をまとめる．検索対象をシーズンやエピソードに限定すると精度が上昇することがわかる．これは検索対象の減少により．検索が容易になったためであると考えられる．シーンまで限定すると回答精度は減少する．これは検索対象の減少により，回答に必要な情報が含まれなくなったことが原因であると推察できる．知識検索モジュールが仮に必要な知識を必ず見つけ出せると仮定すると，検索範囲が全体の場合の精度がエピソードに限定した場合の精度を超える可能性も考えられる．MAGQA において高い回答精度を達成するためには，検索範囲を広くしつつ検索性能を向上させる必要があり，本タスクの難しさがわかる．

また．エピソード単位で知識を利用する文献 [3] の手法では，同じ知識ベースで性能を評価すると平均知識数 26 で回答精度 0.762 となり，提案するベースライン手法の方が高い性能を示した．この結果から，検索範囲を限定した場合でも提案手法が有用であるといえる．

1) シーズンはエピソードをまとめたもので，KnowIT VQA では The Big Bang Theory シリーズの 9 つのシーズンの映像を含む．それぞれのシーズンに含まれるエピソード数は異なるものの，1 シーズンは 20 エピソード程度で構成される．

表 2 知識検索モジュールごとの回答精度．KnowIT VQA で提供されている質問の分類ごとの回答精度も合わせて示す．

	Vis.	Text	Temp.	Know.	ALL
QA のみ	0.476	0.471	0.500	0.543	0.518
ランダム	0.447	0.496	0.558	0.541	0.516
BM25	0.561	0.620	0.733	0.664	0.639
Spider	0.551	0.612	0.628	0.657	0.627
提案手法	**0.588**	**0.649**	**0.756**	**0.687**	**0.663**

4.2 知識検索モジュールによる性能比較

前節の結果から，知識検索モジュールの性能が回答精度を大きく左右すると類推できる．そこで，提案ベースライン手法の知識検索モジュールの替わりに，ランダムな知識を選択する場合．BM25 [9]，エンコーダ E_W, E_K の事前学習モデルである Spider [8] で回答精度を比較した．また．知識を用いず．質問と回答候補のみを質問応答の入力として用いる場合 (QA のみ) も比較対象とする．

表 2 に結果をまとめる．この結果から，どのような種類の質問に関しても提案ベースライン手法の知識検索モジュールを用いた際の回答精度が最も高いことがわかる．疎な表現による検索方式である BM25 と密ベクトルの学習を用いる方法である Spider を比較するとわずかに BM25 が上回るが，Spider を本タスク向けにファインチューニングした提案手法はこれを上回る精度を達成した．これより，知識検索モジュールには質問のドメインに対するファインチューニングが必要であり，知識のアノテーションを利用しない MAGQA のようなタスク設計の重要性が示唆される．

5 おわりに

ユーザの過去の発言や行動に関する質問応答の模擬タスクとして，知識を必要とする Video QA で大きな知識ベースから必要な知識を検索し，質問に回答するタスクである MAGQA を提案した．また，LLM による知識ベース作成．密ベクトルを用いた知識検索．言語モデルによる質問応答によるベースライン手法を提案した．実験では，タスクの困難さとベースライン手法の有効性を示した．一方で，ベースライン手法は基礎的な物となっており，質問応答モジュールからの誤差逆伝播による知識検索モジュールの学習等，さらなる研究が期待される．

謝辞

本研究は JST 創発的研究支援事業 JPMJFR216O の助成による.

参考文献

[1] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. KnowIT VQA: Answering knowledge-based qaquestions about videos. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, pp. 10826–10834, 2020.

[2] Noa Garcia and Yuta Nakashima. Knowledge-based video question answering with unsupervised scene descriptions. In **European Conference on Computer Vision (ECCV)**, pp. 581–598. Springer, 2020.

[3] Deniz Engin, François Schnitzler, Ngoc QK Duong, and Yannis Avrithis. On the hidden treasure of dialog in video question answering. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 2064–2073, 2021.

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan, editors, **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics(ACL)**, pp. 1870–1879, 2017.

[5] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, 2020.

[6] OpenAI. GPT-4 technical report, 2023.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(NAACL-HLT)**, pp. 4171–4186, 2019.

[8] Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. Learning to retrieve passages without supervision. In **Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**, pp. 2687–2700, 2022.

[9] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at trec-3. **NIST Special Publication SP**, Vol. 109, p. 109, 1995.