様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ２ 日現在

機関番号：１４４０１
研究種目：基盤研究(B)（一般）
研究期間：2018～2021
課題番号：１８Ｈ０３２６４

研究課題名（和文）知識ベースを活用した視覚情報に関する質疑応答システムの実現

研究課題名（英文）Visual Question Answering System with a Knowledge Base

研究代表者
中島　悠太（Yuta, Nakashima）

大阪大学・データビリティフロンティア機構・准教授

研究者番号：７０６３３５５１

交付決定額（研究期間全体）：（直接経費）　　13,200,000円

研究成果の概要（和文）：VQAは、DNNの登場により飛躍的に発展したマルチモーダル（自然言語と視覚情報）なデータを扱う分野のひとつである。しかし、真に実用的なシステムとするためには、現状の統計的なアプローチを超えた新たな枠組みが必要である。本研究では、VQAは推論を要するような未知の質問にも対応できるかという挑戦的な学術的「問い」を核心とし、視覚情報に関する質疑応答における知識の利用（Knowledge-based Visual Question Answering; KBVQA）の実現を目的として研究を実施した。映像の記述方法についての検証を行いつつ、知識を利用するモデルを構築し、KBVQAの可能性を示した。

研究成果の学術的意義や社会的意義
本研究では、KBVQAの実現に向けて、モデルの評価のためのデータセットを構築し、その上でKBVQAのプロトタイプシステムを構築した。データセットは、今後のKBVQAの発展に大きく貢献するものであり、学術的に非常に価値が高いものであると考える。また、プロトタイプシステムでは、KBVQAの実現に際して問題となる映像記述とモデルの転用可能性について検証した。特に映像記述については、一般に広く利用されている高次元ベクトルによる記述が不十分であることを示し、新たな映像記述を提案している。

研究成果の概要（英文）：Visual Question Answering (VQA) is an interdisciplinary field, lying on the vision and natural language fields, which is recently advanced drastically due to deep learning. Current techniques for VQA rely on rather a statistics approach, where the distribution of the training set solely matters. We need to go beyond this to make VQA more practical. Our core research question is: “Can VQA systems can answer questions that require inference?”, and we have been committed to building a system that uses knowledge for visual question answering (knowledge-based visual question answering; KBVQA), while also exploring an effective video representation.

研究分野：コンピュータビジョン、パターン認識、自然言語処理

キーワード：質疑応答　知識ベース　深層学習

１．研究開始当初の背景

視覚情報に関する質疑応答（Visual Question Answering; VQA）は、コンピュータビジョンや自然言語処理などの研究分野における重要なタスクとなっている。既存の VQA システムの多くは、質問文-画像-回答の大量のサンプルからなるデータセットを利用して DNN を学習するアプローチを採用している。このアプローチでは、説明文に対して与えられた画像の条件の下で統計的に回答を選択するため、データセットに含まれないタイプの問題に対して尤もらしい回答を生成できず、現状の DNN を用いた VQA は深層学習のトイ・プロブレムという印象を拭えないものであった。

一方で、知識ベースを利用した質疑応答（Knowledge-based Question Answering; KBQA）についても自然言語処理やデータベース分野の研究者によって広く研究が進められている。このアプローチでは、例えば２つの要素間の関係で表現される知識ベースに対して、質問文から生成されたクエリによって検索を実施し、関連度の高い項目を回答として与える。ウェブなどから収集され常に更新され続ける知識ベースに基づく質疑応答システムであることから、任意の質問に対してもっともらしい回答を選出できる可能性を持つ一方、与えられた視覚情報による回答に対する条件付けのための枠組みは存在せず、VQA への単純な適用はできない。

２．研究の目的

VQA は、DNN の登場により飛躍的に発展したマルチモーダル（自然言語と視覚情報）なデータを扱う分野のひとつである。しかし、真に実用的なシステムとするためには、現状の統計的なアプローチを超えた新たな枠組みが必要である。本研究では、VQA は推論を要するような未知の質問にも対応できるかという挑戦的な学術的「問い」を核心とし、DNN を利用した VQA と KBQA の融合によって、視覚情報に関する質疑応答における知識の利用（Knowledge-based Visual Question Answering; KBVQA）の実現を目的として研究を実施する。

３．研究の方法

本研究では、上記研究目的の達成のために、以下の項目について研究を実施した。

（１）　テキストにおけるパラフレーズの検出
（２）　KBVQA のためのデータセット構築とベースラインモデルの提案
（３）　視覚情報の表現と知識の獲得

また、上記に加えて本研究において重要な要素である映像の表現、データセットのバイアス、及び学習済みモデルの転移利用可能性に関して研究を実施した。

（４）　映像の表現に関する実験的考察
（５）　知識の置き換えによる学習済みモデルの再利用可能性の検証

４．研究成果

（１）テキストにおけるパラフレーズの検出

本研究では、モデルに対して自然言語やグラフの形で知識を与えることを考える。グラフの場合であっても基本は自然言語の単語などで表現されるエンティティ間の関係を表すものであることから、いずれにしても自然言語の単語による表現が基盤となる。質問への回答に際しては、与えられた知識から必要なものを検索する。このとき、表層的な記述の違い（例えば、同義語の利用など）から単純な検索では必要な知識が得られない可能性がある。そこで、本研究では自然言語による表現の正規化を目指し、視覚情報と 2 つのフレーズが与えられたときに、それらのフレーズが視覚情報中の同一のエンティティを表すもの（Visually grounded paraphrase; VGP）かを判定するモデルを構築した（図 1）。このモデルでは、それぞれのフレーズから言語特徴量を抽出するとともに、視覚情報からも特徴量を抽出する。視覚情報特徴量と

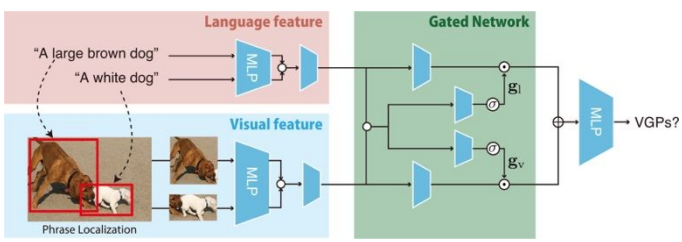

図 1: パラフレーズ検出のためのモデルの概要

しては、まずそれぞれのフレーズに対応する画像中の領域を Visual Grounding のための手法を利用して抽出し、それぞれの領域から特徴量を抽出する。このタスクは、与えられた 2 つのフレーズが表層的に類似している（フレーズ中の多くの単語が共通するなど）場合には、言語特徴のみを利用することで VGP の判定ができる。一方、特に視覚情報が与えられている場合、表層的に全く異なるフレーズが同一のエンティティを表すことも考えられる（自転車レースに参加する人物を「Bicyclist」と「Competitor」のいずれでも表すことができるなど）。この場合、視覚情報特徴量によるエンティティの特定が必要となる。提案手法では、表層的に類似する場合には言語特徴量のみを、そうでない場合には視覚情報特徴量を援用するように、それぞれの特徴量に重み付けを行う。結果として、提案手法では F1 スコアで 86.5%と他のベースライン手法などを超える性能が得られた（表 1）。この成果は、国際論文誌や国際会議などで発表している。

表 1: 提案手法とベースライン手法との比較。Chu et al. (2018)は[3]。

	F1	Prec.	Rec.
Chu et al. (2018)	84.16	82.71	85.67
Word-overlap	61.25	74.15	52.18
Phrase-only	85.66	84.72	86.61
Visual-only (PL-CLC)	57.73	51.86	65.09
Visual-only (DDPN)	66.36	60.92	72.87
BoundingBox-overlap (DDPN)	73.43	73.83	73.05
Ours (PL-CLC)	85.10	83.36	86.91
Ours (DDPN)	86.48	**85.81**	87.16
Ours+BBox (DDPN)	**86.50**	84.92	**88.15**

（2）KBVQA のためのデータセット構築とベースラインモデルの提案

本研究における中心的課題である、特に映像を対象とした KBVQA には公開されたデータセットが存在しない。そこで、提案モデルの学習や評価のためのデータセットを構築した。インターネとからの知識の獲得しやすさや、映像データの入手しやすさ、また知識を要求する質問の作りやすさを考慮し、映像ドメインとしてドラマを選択した。クラウドソーシングサービスを利用し、元映像から抽出した 12,000 本以上の短時間の映像クリップに対して、24,000 件以上の質問、回答、必要な知識の組を得た。表 2 に既存の関連するデータセットと本データセットの比較を示す。また、当該データセットのためのベースライン手法として、クラウドソーシングで得られた回答に必要な知識をデータベースとし、映像と質問が与えられた際に、データベースから知識を検索して回答に利用するシステムをベースライン手法として構築した（図 2）。ベースライン手法は 64%の正解率で回答が可能であることを実験的に示した。一方で、回答に必要な知識については問題に対応するものを与えた場合の正解率は 73%であった。ここから、本ベースライン手法の知識の検索が不十分であることがわかる。また、当該ドラマをよく知る人が回答する場合の正解率は 90%であった。この差から、ベースライン手法の性能が人間には全く及ばないことが明らかとなった。Ablation Study では、映像特徴量の有無が最終的な正解率に与える影響は限定的であることが示された。このような傾向は、多くの視覚情報と言語のタスクで共通しており、今後のさらなる研究を要する。関連する内容は、人工知能分野の著名な国際会議である AAAI をはじめとした国内外の国際会議、ワークショップで発表している。

表 2: 既存のデータセットと本研究で構築したデータセットの比較。Answers の列の MC_N は N 個の回答候補からの選択式、Word は自由記述を表す。Vis.、Text.、Temp.、及び Know.はそれぞれ視覚情報理解、言語理解、時間的変化の理解、知識を必要とする質問が含まれることを表す。

Dataset	VQA-Type	Domain	# Imgs	# QAs	Answers	Vis.	Text.	Temp.	Know.
MovieQA (Tapaswi et al. 2016)	Video	Movie	6,771	14,944	MC_5	✓	✓	✓	-
KB-VQA (Wang et al. 2017)	KB	COCO	700	2,402	Word	✓	-	-	✓
PororoQA (Kim et al. 2017b)	Video	Cartoon	16,066	8,913	MC_5	✓	✓	✓	-
TVQA (Lei et al. 2018)	Video	TV show	21,793	152,545	MC_5	✓	✓	✓	-
R-VQA (Lu et al. 2018)	KB	Visual Genome	60,473	198,889	Word	✓	-	-	-
FVQA (Wang et al. 2018)	KB	COCO, ImgNet	2,190	5,826	Word	✓	-	-	✓
KVQA (Shah et al. 2019)	KB	Wikipedia	24,602	183,007	Word	✓	-	-	✓
OK-VQA (Marino et al. 2019)	KB	COCO	14,031	14,055	Word	✓	-	-	✓
KnowIT VQA (Ours)	VideoKB	TV show	12,087	24,282	MC_4	✓	✓	✓	✓

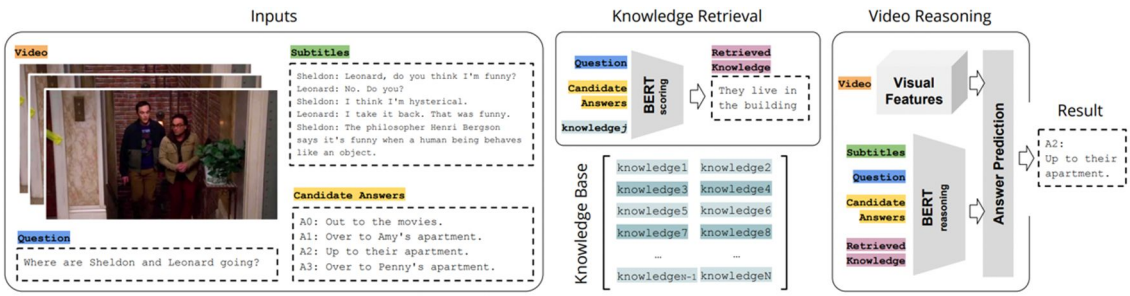

図 2: 知識を利用する質疑応答システムのベースライン手法の概要

（3）視覚情報の表現と知識の獲得

上記（2）で述べたとおり、ベースライン手法は映像特徴量の有無による正答率の差が小さく、映像理解に関する性能が不十分であった。そこで、多くの既存手法で採用されている、CNN などから得られるベクトルを映像特徴量とするアプローチとは異なるアプローチを模索した。また、ベースライン手法では知識として作問者が入力した知識を検索して利用したが、現実的にはこのような形で必要な知識を入手することができない。そこで、まず映像特徴量について、視覚情報全体から得られるベクトルでは十分な記述ができていないためであると考え、映像からそのシーンの場所やシーン中の人物、さらには人物や周辺の物体との関係などを検出し、検出されたすべての要素の間の関係をグラフにより記述するシーングラフを構築した。このとき、シーングラフを直接 Graph Neural Network などに入力することで映像特徴量を得る方法が考えられるが、学習に使えるデータ数が比較的少数であることから、十分な学習ができない可能性がある。そこで、シーングラフをルールベースで自然言語テキストに変換し、このテキストを巨大なデータセットで事前学習された BERT に入力することで映像特徴量を得る。これにより、テキスト内の単語の関係（検出された要素間の関係）がより反映された特徴量となることが期待できる。また、知識については、インターネット上から当該ドラマのそれぞれのエピソードのまとめが記述されテキストを抽出し、このテキストから必要な知識を獲得する。提案手法の概要を図 3 に示す。この手法により、正解率は 72%となり、ベースライン手法に比べて高い性能が得られた。この成果は、コンピュータビジョンに関するトップ国際会議の一つである ECCV で発表している。

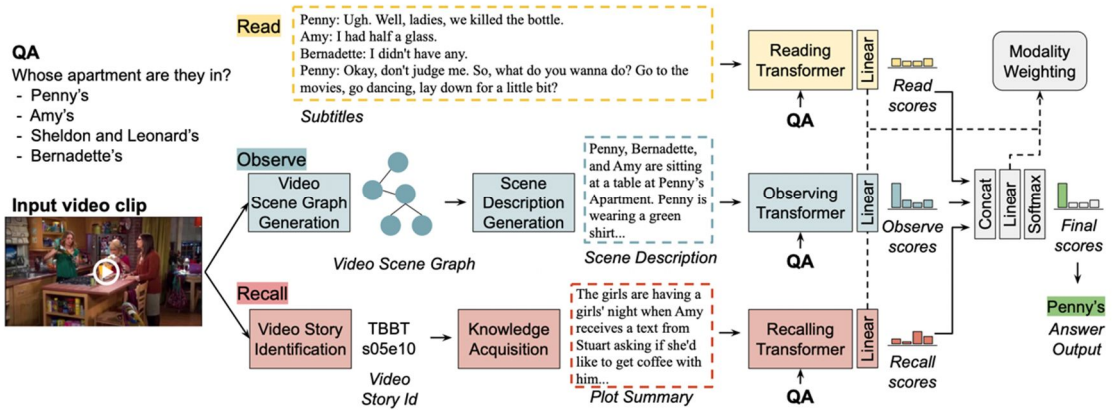

図 3: テキストによる映像記述と知識のインターネットからの収集によるモデルの概要

（4）映像の表現に関する実験的考察

前述の通り、特に視覚情報と自然言語に関わるタスクにおいては、視覚情報が役に立たないという問題が広く知られている。そこで、（3）に先立ち、映像から抽出したベクトルに変えて、映像に対して物体検出手法を適用することで得られるクラスラベル（に対応するフレーズ）をテキストであると捉え、そのテキストを BERT などの事前学習済みのモデルに入力して得られるベクトルを映像特徴量として利用するアプローチについて、その性能を映像に関する質疑応答タスクで検証した。結果、あるデータセットでは既存手法を超える正解率が得られた。このことから、CNN などから得られるベクトルでは、映像についての十分な記述ができない可能性が示唆された。この結果は国際論文誌や国際会議にて報告している。

（5）知識の置き換えによる学習済みモデルの再利用可能性の検証

本研究では、映像に関する質疑応答における知識の利用について研究を進めている。ここでの知識は、基本的に自然言語テキストによって与えられるものであり、あるドメインのための知識を別のドメインの知識に置き換えることが容易にできる。一方で、上記（2）や（3）では、知識の検索も学習によって実現する。学習時に知識にアクセスできることを前提としており、固有名詞や単語の偏りなどが学習されている可能性がある。そこで、あるドメインのデータセットで学習したモデルを別のモデルで利用できるかを検証するため、（2）で構築したものとは異なるデータセットを構築し、2 つのデータセットを利用してモデルの再利用可能性について実験した。新たに構築したデータセットは、（2）とは別のドラマをドメインとしており、固有名詞を中心に多く利用される単語が異なる。質問、回答、必要な知識の収集は、（2）と同一の方法でクラウドソーシングを利用して実施し、21,412 件のサンプルを得た。固有名詞や語彙の違いによる影響を抑えるために、固有名詞が人物の場合は学習時に当該固有名詞を「...固有名詞, a person,...」のように変更する、また英語からドイツ語に機械翻訳した後に再度英語に機械翻訳するバック

トランスレーションと呼ばれるデータ拡張手法の一種を利用して学習データに含まれる文のバリエーションを増加させる。実験では、一方のデータセットで学習したモデルを、他方のデータセットのサブセットで転移学習した場合の知識の検索性能を評価した。結果、転移学習では直接対象のデータセットで学習した場合には及ばないものの、少数のデータセットで学習した場合には、転移学習、及び上記のデータ拡張手法が有効であることが明らかになった。これは、あるドメインで学習したモデルを、別のサンプル数の少ないドメインで利用できる可能性を示す。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Otani Mayu、Chu Chenhui、Nakashima Yuta	４．巻 404
２．論文標題 Visually grounded paraphrase identification via gating and phrase localization	５．発行年 2020年
３．雑誌名 Neurocomputing	６．最初と最後の頁 165～172
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neucom.2020.04.066	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yang Zekun、Garcia Noa、Chu Chenhui、Otani Mayu、Nakashima Yuta、Takemura Haruo	４．巻 445
２．論文標題 A comparative study of language transformers for video question answering	５．発行年 2021年
３．雑誌名 Neurocomputing	６．最初と最後の頁 121～133
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.neucom.2021.02.092	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Mayu Otani, Chenhui Chu, and Yuta Nakashima	４．巻 -
２．論文標題 Visually grounded paraphrase identification via gating and phrase localization	５．発行年 2020年
３．雑誌名 Neurocomputing	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Noa Garcia, Benjamin Renoust, and Yuta Nakashima	４．巻 9
２．論文標題 ContextNet: Representation and exploration for painting classification and retrieval in context	５．発行年 2019年
３．雑誌名 International Journal on Multimedia Information Retrieval	６．最初と最後の頁 17-30
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s13735-019-00189-4	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計21件（うち招待講演　0件／うち国際学会　14件）

１．発表者名
Yuta Kayatani、Zekun Yang、Mayu Otani、Noa Garcia、Chenhui Chu、Yuta Nakashima、Haruo Takemura

２．発表標題
The Laughing Machine: Predicting Humor in Video

３．学会等名
2021 IEEE Winter Conference on Applications Computer Vision（国際学会）

４．発表年
2021年

１．発表者名
Mayu Otani、Yuta Nakashima、Esa Rahtu、Janne Heikkila

２．発表標題
Uncovering Hidden Challenges in Query-Based Video Moment Retrieval

３．学会等名
31st Biritish Machine Vision Conference（国際学会）

４．発表年
2020年

１．発表者名
Noa Garcia、Mayu Otani、Chenhui Chu、Yuta Nakashima

２．発表標題
Knowledge-Based Visual Question Answering in Videos

３．学会等名
2020 Conference on Computer Vision and Pattern Recognition Workshops（国際学会）

４．発表年
2020年

１．発表者名
Noa Garcia、Chentao Ye、Zihua Liu、Qingtao Hu、Mayu Otani、Chenhui Chu、Yuta Nakashima、Teruko Mitamura

２．発表標題
A Dataset and Baselines for Visual Question Answering on Art

３．学会等名
2020 Workshop on Computer Vision for Art（国際学会）

４．発表年
2020年

１．発表者名
Noa Garcia、Yuta Nakashima

２．発表標題
Knowledge-Based Video Question Answering with Unsupervised Scene Descriptions

３．学会等名
European Conference on Computer Vision（国際学会）

４．発表年
2020年

１．発表者名
Mayu Otani、Yuta Nakashima、Esa Rahtu、Janne Heikkila

２．発表標題
What We All Need Are Non-trivial Baselines and Sanity Checks

３．学会等名
第23回 画像の認識・理解シンポジウム

４．発表年
2020年

１．発表者名
Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura

２．発表標題
BERT representations for video question answering

３．学会等名
IEEE Winter Conference on Applications of Computer Vision（国際学会）

４．発表年
2020年

１．発表者名
Noa Garcia, Chenhui Chu, Mayu Otani, and Yuta Nakashima

２．発表標題
KnowIT VQA: Answering knowledge-based questions about video

３．学会等名
AAAI Conference on Artificial Intelligence（国際学会）

４．発表年
2020年

１．発表者名
Mayu Otani, Chenhui Chu, and Yuta Nakashima

２．発表標題
Adaptive gating mechanism for identifying visually grounded paraphrases

３．学会等名
Multi-Discipline Approach for Learning Concepts（国際学会）

４．発表年
2019年

１．発表者名
Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila

２．発表標題
Rethinking the evaluation of video summaries

３．学会等名
IEEE Conference on Computer Vision and Pattern Recognition（国際学会）

４．発表年
2019年

１．発表者名
Noa Garcia, Benjamin Renoust, and Yuta Nakashima

２．発表標題
Context-aware embeddings for automatic art analysis

３．学会等名
ACM International Conference on Multimedia Retrieval（国際学会）

４．発表年
2019年

１．発表者名
Noa Garcia, Chenhui Chu, Mayu Otani, and Yuta Nakashima

２．発表標題
Video meets knowledge in visual question answering

３．学会等名
第22回 画像の認識・理解シンポジウム

４．発表年
2019年

１．発表者名
Mayu Otani, Kazuhiro Ota, Yuta Nakashima, Esa Rahtu, Janne Heikkila, and Yoshitaka Ushiku

２．発表標題
Collecting relation-aware video captions

３．学会等名
第22回 画像の認識・理解シンポジウム

４．発表年
2019年

１．発表者名
Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura

２．発表標題
Video question answering with BERT

３．学会等名
第22回 画像の認識・理解シンポジウム

４．発表年
2019年

１．発表者名
萱谷 勇太，大谷まゆ，Chenhui Chu，中島 悠太，竹村 治雄

２．発表標題
コメディドラマにおける字幕と表情を用いた笑い予測

３．学会等名
2019年度 人工知能学会全国大会

４．発表年
2019年

１．発表者名
Noa Garcia, Benjamin Renoust, and Yuta Nakashima

２．発表標題
Understanding art through multi-modal retrieval in paintings

３．学会等名
Language and Vision Workshop（国際学会）

４．発表年
2019年

１．発表者名
Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila

２．発表標題
Rethinking the evaluation of video summaries

３．学会等名
IEEE Computer Society Conference on Computer Vision and Pattern Recognition（国際学会）

４．発表年
2019年

１．発表者名
Noa Garcia, Benjamin Renoust, and Yuta Nakashima

２．発表標題
Context-aware embeddings for automatic art analysis

３．学会等名
ACM International Conference on Multimedia Retrieval（国際学会）

４．発表年
2019年

１．発表者名
Chenhui Chu, Mayu Otani, and Yuta Nakashima

２．発表標題
iParaphrasing: Extracting visually grounded paraphrases via an image

３．学会等名
27th International Conference on Computational Linguistics（国際学会）

４．発表年
2018年

１．発表者名
Mayu Otani, Chenhui Chu, and Yuta Nakashima

２．発表標題
Phrase localization-based visually grounded paraphrase identification

３．学会等名
第21回 画像の認識・理解シンポジウム

４．発表年
2018年

１．発表者名
Mayu Otani, Chenhui Chu, and Yuta Nakashima

２．発表標題
Visually grounded paraphrase extraction via phrase grounding

３．学会等名
Workshop on Language and Vision at CVPR

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

KnowIT VQA Paper
https://knowit-vqa.github.io
Knowledge VQA
https://www.n-yuta.jp/project/knowledge-vqa/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	金　進東 (Kim Jin-Dong) (40536893)	大学共同利用機関法人情報・システム研究機構（機構本部施設等）・データサイエンス共同利用基盤施設・特任准教授 (82657)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関