DialDoc 2021 Shared Task:
Goal-Oriented Document-grounded Dialogue Modeling

Song Feng
IBM Research AI
sfeng@us.ibm.com

Abstract

We present the results of Shared Task at Workshop DialDoc 2021 that is focused on document-grounded dialogue and conversational question answering. The primary goal of this Shared Task is to build goal-oriented information-seeking conversation systems that can identify the most relevant knowledge in the associated document for generating agent responses in natural language. It includes two subtasks on predicting agent responses: the first subtask is to predict the grounding text span in the given document for next agent response; the second subtask is to generate agent response in natural language given the context. Many submissions outperform baseline significantly. For the first task, the best-performing system achieved 67.1 Exact Match and 76.3 F1. For the second subtask, the best system achieved 41.1 SacreBLEU and highest rank by human evaluation.

1 Introduction

Goal-oriented conversational systems could assist end users to query information in documents dynamically via natural language interactions. Meanwhile, there is a vast number of documents in which individuals and organizations choose to present their interests and knowledge to the world for broad applications. Thus, it attracts a lot of attentions from researchers and practitioners from different fields. There have been significant individual research threads that show promises in handling heterogeneous knowledge embedded in the documents (Talmor et al., 2021), including (1) unstructured content such as text passages (CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), ShARC (Saeidi et al., 2018), DoQA (Campos et al., 2020), Doc2Dial (Feng et al., 2020)); (2) semi-structured content such as tables or lists (SQA (Iyyer et al., 2017), HybridQA (Chen et al., 2020)); (3) multimedia such as images and videos with associated textual descriptions (RecipeQA (Yagcioglu et al., 2018), PsTuts-VQA (Colas et al., 2020), MIMOQA (Singh et al., 2021)) Despite these recent advances, the challenge remains for handling multi-turn queries of complex dialogue scenarios (Ma et al., 2020; Feng et al., 2020) and then respond based on the most relevant content in documents of various types from wide domains. As a step forward, we propose a shared task and competition to invite researchers to bring their individual perspectives and advance the field in joint effort.

We introduce DialDoc 2021 Shared Task, which focuses on building goal-oriented information-seeking dialogue that are grounded in textual content. In particular, the goal is to develop a dialogue system to comprehend multi-turn queries and identify the most relevant knowledge in the associated document for generating agent responses in natural language. It includes two subtasks for predicting agent response. The first subtask (Subtask 1) is to predict the grounding text span in the given document for next agent response; the second subtask (Subtask 2) is to generate agent response in natural language given the contexts. The dataset used for the task is a goal-oriented document-grounded dialogue dataset Doc2Dial (Feng et al., 2020). We hosted the leaderboards for Dev-Test and Test phase on eval.ai for two subtasks respectively. There are a total of 23 teams that participated Dev-Test phase. For final test phrase, 11 teams submitted to the leaderboard of Subtask 1, and 9 teams submitted to the leaderboard of Subtask 2. For the first task, the best system achieved 67.09 Exact Match and 76.34 F1. For the second subtask, the best system achieved 41.06 sacrebleu and rank the best by human evaluation.

In this work, we first describe the dataset and the two subtasks. Then, we provide a summary of the evaluation results from participating systems.

Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 1–7
August 5–6, 2021. ©2021 Association for Computational Linguistics

2 Dataset

We use Doc2Dial dataset [1] introduced in Feng et al. (2020), which contains 4793 goal-oriented dialogues and a total of 488 associated grounding documents from four domains for social welfare: `dmv`, `va`, `ssa`, and `studentaid`. In this dataset, dialogues contain the scenarios when agent ask follow-up questions for clarification or verification based on dialogue-based and document-based context. Each turn is annotated with (1) grounding span from the associated document, (2) dialogue act, e.g., *query*, *respond* and (3) speaker role, either *agent* or *user*.

For developing models, we divide the data into training, validation and test split based on the number of dialogues. For evaluating the models, we provide a dev-test set which contains about 30% test dataset. The final test set also includes dialogue and document data from an unseen domain *cdccovid* that is not in the training, validation or dev-test set. The dialogues of unseen domain were collected in the same data collection process as published Doc2Dial dataset. Table 1 presents the number of dialogues ('dials'), total turns ('turns') of all dialogues and total turns for prediction ('predicts') in each data split.

#	train	val	test-dev	test
dials	3474	661	198	787
turns	44149	8539	1353	5264
predicts	20431	3972	198	787

Table 1: Statistics of dialogue data of different data splits.

3 Task Description

This Shared Task focuses on building goal-oriented information-seeking dialogue systems. The goal is to teach a dialogue system to identify the most relevant knowledge in the associated document for generating agent responses in natural language. It includes two subtasks on predicting agent response. The agent can either provide an answer or ask follow-up question. Here we only consider the cases that use queries are answerable.

3.1 Subtask 1

This subtask is to predict the grounding span of next agent response. The input current turn, dialogue history and one associated document; the output is a text span. The evaluation is based on token-level F1 and exact match score (Rajpurkar et al., 2018).

3.2 Subtask 2

This subtask is to generate the next agent utterance. The input is current turn, dialogue history and the document context; the output is utterance in natural language. The evaluation is based on SacreBLEU (Post, 2018). We also perform human evaluation on the top three submissions with highest SacreBLEU for determining the final rank.

Human evaluation We ask human annotators to rank a group of three utterances from the three submissions based on *relevance* and *fluency* given document context and dialogue history. *relevance* is used to measure how well the generated utterance is relevant to grounding span as a response to the previous dialogue turn(s). *fluency* indicates whether the generated utterance is grammatically correct and generally fluent in English. We randomly select 100 generated turns where the utterances are not all the same. We collect five judgements per group.

4 Baseline

Subtask 1 The baseline model for Subtask 1 is based on BERT-QA (Devlin et al., 2019). For each token, it computes the probabilities of start and end positions by a linear projection from the last hidden layers of the BERT model. Then it multiplies the scores of the start and end positions for estimating the probability of the corresponding span. As a baseline, we fine-tune BERT-base on Doc2Dial dataset where the input is dialogue query and the associated document context. The dialogue query is the concatenation of dialogue turns in reverse order.

Subtask 2 The task is formulated as an end-to-end text generation task. The baseline approach for Subtask 2 is based on sequence-to-sequence model BART by (Lewis et al., 2020). We fine-tune the pre-trained BART model (`bart-cnn-large`) on Doc2Dial dataset. The source input consists of current turn, dialogue history along with document title and content that are separated by special tokens. The target output is next agent utterance.

[1] `https://doc2dial.github.io/file/doc2dial_v1.0.1.zip`

5 Shared Task Submissions

We hosted the leaderboards [2] for Dev-Test and Test phase for the two subtasks on `eval.ai`. The Dev-Test and Test phase lasted three months and one week respectively. There are a total of 23 teams that participated Dev-Test phase. For final Test phrase, 11 teams submitted to the leaderboard of Subtask 1, and 9 teams submitted to the leaderboard of Subtask 2. Among the best-performing systems, some teams utilize additional data for augmentation for pre-training (e.g., CAiRE (Xu et al., 2021), SCIR-DT (Li et al., 2021)), some teams employ neural retrievers for obtaining most relevant document passages (e.g., RWTH (Daheim et al., 2021) and ER). For the first task, the best system achieved 67.1 Exact Match and 76.3 F1. For the second subtask, the best system achieved 41.1 sacrebleu and rank the best by human evaluation. Next, we provide a brief summary of the work by 8 teams as listed in Table 2, who submitted their technical system papers.

5.1 ER

ER[3] participates Subtask 1. It introduces a model that leverages the structure in grounding document and dialogue context. It applies a multi-passage reader model based on transformer-based encoder to encode each passage concatenated with dialogue context and document title. It optimizes both passage selection, start and end position selection with gold knowledge passage during training. The final submission is an ensemble of 12 models and achieves the best results for Subtask 1.

5.2 SCIR-DT

SCIR-DT (Li et al., 2021) participates Subtask 1. Their methods include data augmentation, model pretraining/fine-tuning, postprocessing, and ensemble. For data augmentation, they use back-translation and synonym substitution to obtain 5 times of document and dialogue data, which are then paired into 25 times data. They use the augmented data for pretraining BERT and RoBERTa with whole word masking technique and doc2dial data for fine-tuning BERT, RobERTa and ELECTRA. The ensemble method selects the most probably rank span based on the linear combination of ranking results per model and learn the hyperparameter for inference. The team ranks 2nd based on the average of normalized F1 and EM scores used for the final evaluation.

Team	Affiliation
CAiRE	The Hong Kong University of Science and Technology
ER	Anonymous
JARS	Carnegie Mellon University
KU_NLP	Konkuk University & Kangwon National University
RWTH	RWTH Aachen University
SB_NITK	National Institute of Technology Karnataka
Schlussstein	Carnegie Mellon University Bosch Research Pittsburgh
SCIR-DT	Harbin Institute of Technology

Table 2: Participating teams and affiliations.

5.3 KU_NLP

KU_NLP (Kim et al., 2021) participates both tasks. For Subtask 1, they adopt pretrained RoBERTa as backbone and predict dialogue act and span jointly. For Subtask 2, they include several tokens and embeddings based on document structure into input representation for BART. Instead of random order of the training instances, they propose to apply curriculum learning (Xu et al., 2020) based on the computed task difficulty level for each task respectively. The final submission on Subtask 2 is based on the span prediction by a single model. It achieves best SacreBLEU and human evaluation results.

5.4 RWTH

RWTH (Daheim et al., 2021) participates both tasks. For Subtask 1, it applies BERTQA with additional span-based specifics in their approach. First, they restrict start and end position only to the begin and end of sub-clauses since Doc2Dial dataset is based on preprocessed spans. In addition, they consider modeling the joint probability of a span inspired by Fajcik et al. (2020). The final submission is the ensemble of multiple models, where the probability of a span is obtained by marginalizing the joint probability of span and model over all models. For Subtask 2, they propose to cascade over all spans where they use top N (=5) spans as a approximation. The probability is computed jointly. The generation model is trained with cross-entropy using an n-best list obtained from the separately

[2] https://eval.ai/web/challenges/challenge-page/793/overview

[3] The submission is non-archival.

trained selection model.

5.5 CAiRE

CAiRE (Xu et al., 2021) participates both tasks. They utilize data augmentation methods and several training techniques. For the first task, it uses QA data such as MRQA shared task dataset (Fisch et al., 2019) and conversational QA data such as CoQA (Reddy et al., 2019) for pretraining RoBERTa with multi-task learning strategy and the models are fine-tuned on Doc2Dial dataset. For the second task, they pretrain BART on Wizard-of-Wikipedia dataset (Dinan et al., 2019). Then they fine-tune the model using the knowledge prediction results from the first task. The final submission is based on the ensemble of multiple models where the best span is determined by the majority vote by models.

5.6 SB_NITK

SB_NITK (Bachina et al., 2021) participates Subtask 1. They also adapt data augmentation approaches that utilize additional Question Answering dataset such as SQuAD 2.0 (Lee et al., 2020), Natural Questions (Kwiatkowski et al., 2019) and CORD-19 (Wang et al., 2020) for pretraining several models including RoBERTa, ALBERT and ELECTRA. Then they experiment with different combinations of ensemble models. The final submission is based on the ensemble of ensemble ALBERTA and RoBERTa using all three additional datasets.

5.7 JARS

JARS (Khosla et al., 2021) participates in Subtask 1. It also uses transformer-based QA models, for which it pretrains on different Question Answering datasets such as SQuAD, different subsets of MRQA-2019 training set along with conversational QA data such as CoQA and QuAC. The experiments suggest that conversational QA datasets are more helpful comparing to QA datasets. They compare three different ensemble methods and use the highest average probability score for span prediction based on multiple models.

5.8 Schlussstein

Schlussstein (Chen et al., 2021) submit to both subtasks. For Subtask 1, they pretrain BERT on datasets such as SQuAD and CoQA before fine-tuning on Doc2Dial. To incorporate longer document content in Doc2Dial dataset, they also experiment with longer document stride and observe performance improvement. For Subtask 2, it pretrains BART model on CoQA dataset before fine-tuning it on Doc2Dial dataset.

	Team	Exact_Match	F1
1	ER	67.1 (61.8)	76.3 (73.1)
2	SCIR-DT	63.9 (59.1)	75.6 (71.6)
3	RWTH	63.5 (58.3)	75.9 (73.2)
4	CAiRE	60.7 (-)	75.0 (-)
5	KU_NLP	58.7 (58.7)	73.4 (73.4)
6	SB_NITK	58.6 (-)	73.4 (-)
7	JARS	53.5 (52.6)	70.9 (67.4)
-	baseline	35.8 (35.8)	52.6 (52.6)

Table 3: Participating teams of Subtask 1. The rank is based on the average of normalized average of F1 and EM scores.

Rank	Team	SacreBLEU
1	KU_NLP	41.1 (41.1)
2	RWTH	40.4 (39.1)
3	CAiRE	37.7 (-)
4	SCIR-DT	30.7 (-)
-	baseline	17.6 (17.6)

Table 4: Participating teams and evaluation results on test set of Subtask 2.

6 Results

Subtask 1 We present the evaluation results on final Test phase of Subtask1 from 7 participating teams in Table 3. The submissions are ordered based on the average of normalized F1 and EM scores. All submissions of Test Phase outperform BERT-base baseline by large margin. The scores in parentheses are by single models. All other results except the ones by KU_NLP are based on various ensemble methods, which further improve the performances significantly in most cases.

Subtask 2 Table 4 presents the evaluation results on final test set of Subtask 2 from 4 participating teams. We performance human evaluations on the top three submissions based on SacreBLEU scores. We use three different ways to compute majority vote to get the aggregated results: (1) we consider the rank if it is agreed among at least three annotators; (2) we consider the rank if it is agreed among at least two annotators; (3) we also use the aggregation results provided by Appen platform, which takes consideration of annotator's historical performances.

7 Conclusion

We presented the results of 1st DialDoc 2021 Shared Task, which included two subtasks on document-grounded goal-oriented dialogue modeling. We received submissions from a total of 17 teams during entire phase for Subtask 1, and 9 teams for Subtask 2. All submissions during final Test phase outperformed baselines by a large margin for both subtasks. By organizing this shared task, we hope to invite researchers and practitioners to bring their individual perspectives on the subject, and to jointly advance the techniques toward building assistive agents to access document content for end users by conversing.

Acknowledgements

We would like to thank Luis Lastras, Sachindra Joshi, Siva Reddy and Siva Sankalp Patel for a lot of helpful discussions on organizing this shared task. We thank eval.ai for their help and support on hosting the leaderboards on their platform. Finally, we are thankful to IBM Research AI for sponsoring the shared task and competition.

References

Sony Bachina, Spandana Balumuri, and Sowmya Kamath S. 2021. Ensemble albert and roberta for span prediction in question answering. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Xi Chen, Faner Lin, Yeju Zhou, and Kaixin Ma. 2021. Building goal-oriented document-grounded dialogue systems. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang, and Doo Soon Kim. 2020. TutorialVQA: Question answering dataset for tutorial videos. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5450–5455, Marseille, France. European Language Resources Association.

Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2021. Cascaded span extraction and response generation for document-grounded dialog. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Martin Fajcik, Josef Jon, Martin Docekal, and Pavel Smrz. 2020. BUT-FIT at SemEval-2020 task 5: Automatic detection of counterfactual statements with deep pre-trained language representation models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 437–444, Barcelona (online). International Committee for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Sopan Khosla, Justin Lovelace, Ritam Dutt, and Adithya Pratapa. 2021. Team jars: Dialdoc subtask 1 - improved knowledge identification with supervised out-of-domain pretraining. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Boeun Kim, Dohaeng Lee, Yejin Lee, Harksoo Kim, Sihyung Kim, Jin-Xia Huang, and Oh-Woog Kwon. 2021. Document-grounded goal-oriented dialogue systems on pre-trained language model with diverse input representation. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. SQuAD2-CR: Semi-supervised annotation for cause and rationales for unanswerability in SQuAD 2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5425–5432, Marseille, France. European Language Resources Association.

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

Jiapeng Li, Mingda Li, Longxuan Ma, Weinan Zhangy, and Ting Liu. 2021. Technical report on shared task in dialdoc21. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A survey of document grounded dialogue systems (dgds). *arXiv preprint arXiv:2004.13818*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Yan Xu, Etsuko Ishii, Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, and Pascale Fung. 2021. Caire in dialdoc21: Data augmentation for information-seeking dialogue system. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering*. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.