Can You Unpack That? Learning to Rewrite Questions-in-Context

Ahmed Elgohary, Denis Peskov, Jordan Boyd-Graber*
Department of Computer Science, UMIACS, iSchool, Language Science Center
University of Maryland, College Park
{elgohary, dpeskov, jbg}@cs.umd.edu

Abstract

Question answering is an AI-complete problem, but existing datasets lack key elements of language understanding such as coreference and ellipsis resolution. We consider sequential question answering: multiple questions are asked one-by-one in a conversation between a questioner and an answerer. Answering these questions is only possible through understanding the conversation history. We introduce the task of question-in-context rewriting: given the context of a conversation's history, rewrite a context-dependent into a self-contained question with the same answer. We construct, CANARD, a dataset of 40,527 questions based on QuAC (Choi et al., 2018) and train Seq2Seq models for incorporating context into standalone questions.

1 Introduction

Question Answering (QA) is an AI complete problem (Webber, 1992), but existing QA datasets do not rise to the challenge: they lack key NLP problems like anaphora resolution, coreference disambiguation, and ellipsis resolution. The logic needed to answer these types of questions requires deeper NLP understanding that simulates the context in which humans naturally answer questions.

Neural techniques question answering have improved (Devlin et al., 2018) machine reading comprehension (Rajpurkar et al., 2016, MRC): computers can take a single question and extract answers from datasets like Wikipedia. However, QA models struggle to generalize when questions do not look like the standalone questions systems in training data: e.g., new genres, languages, or closely-related tasks (Yogatama et al., 2019).

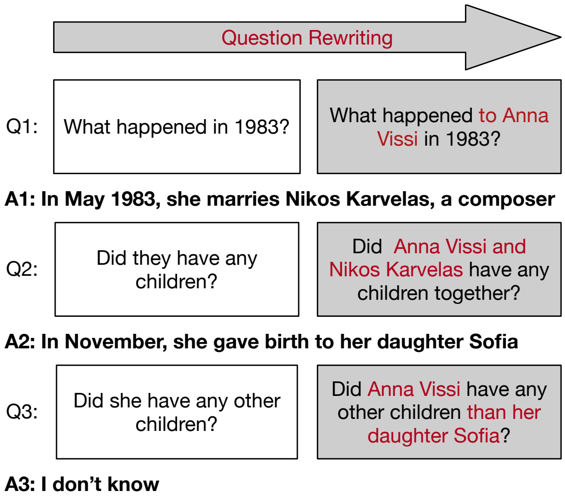
Question Rewriting
Q1:
What happened in 1983?
What happened to Anna Vissi in 1983?
A1: In May 1983, she marries Nikos Karvelas, a composer
Q2:
Did they have any children?
Did Anna Vissi and Nikos Karvelas have any children together?
A2: In November, she gave birth to her daughter Sofia
Q3:
Did she have any other children?
Did Anna Vissi have any other children than her daughter Sofia?
A3: I don't know

Figure 1: Question-in-context rewriting task. The input to each step is a question to rewrite given the dialog history which consists of the dialog utterances (questions and answers) produced before the given question is asked. The output is an equivalent, context-independent paraphrase of the input question.

Conversational question answering (Reddy et al., 2019, CQA) is a generalization that ask *multiple* questions in an information-seeking dialogs. Unlike MRC, CQA requires models to link questions together to resolve the conversational dependencies between them: each question needs to be understood in the conversation context. For example, the question *"What was he like in that episode?"* cannot be understood without knowing what *"he"* and *"that episode"* refer to, which can be resolved using the conversation context.

We reduce challenging, interconnected CQA examples to independent, stand-alone MRC to create CANARD—Context Abstraction: Necessary Additional Rewritten Discourse—a new dataset[1] that rewrites QuAC (Choi et al., 2018) questions. We crowdsource context-independent paraphrases of QuAC questions and use the paraphrases to train and evaluate question-in-context rewriting.

[*] Now at Google AI Zürich

[1] http://canard.qanta.org

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5918–5924,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

Characteristic	Ratio
Answer Not Referenced	0.98
Question Meaning Unchanged	0.95
Correct Coreferences	1.0
Grammatical English	1.0
Understandable w/o Context	0.90

Table 1: Manual inspection of 50 rewritten context-independent questions from CANARD suggests that the new questions have enough context to be independently understandable.

Section 2 formally defines the task of question de-contextualization. Section 3 constructs CANARD, a new dataset of question-in-context with corresponding context-independent paraphrases. Section 5 analyzes our rewrites (and the underlying methodology) to understand the linguistic phenomena that make CQA difficult. We build several baseline rewriting models and compare their BLEU scores to our human rewrites in Section 4.

2 Defining Question-In-Context Rewrites

We formally define the task of question-in-context rewriting (de-contextualization). Given a conversation topic t and a history H of $m-1$ turns, each turn k is a question q_i and an answer a_i; the task is to generate a rewrite q'_m for the next question q_m based on H. Since q_m is part of the conversation, its meaning often involves references to parts of its preceding history. A valid rewrite q'_m should be self-contained: a correct answer to q'_m by itself is a correct answer to q_m combined with the question's preceding history H.

Figure 1 shows the assumptions of CQA and how they are made explicit in rewrites. The first question omits the title of the page (Anna Vissi), the second question omits the answer to the first question (replacing both Anna Vissi and her husband with the pronoun "they"), and the last question adds a scalar implicature that must be resolved.

3 Dataset Construction

We elicit paraphrases from human crowdworkers to make previously context-dependent questions *unambiguously* answerable. Through this process, we resolve difficult coreference linkages and create a pair-wise mapping between ambiguous and context-enriched questions. We derive CANARD from QuAC (Choi et al., 2018), a sequential question answering dataset about specific Wikipedia sections. QuAC uses a pair of workers—a "student" and a "teacher"—to ask and respond to questions. The "student" asks questions about a topic based on only the title of the Wikpedia article and the title of the target section. The "teacher" has access to the full Wikipedia section and provides answers by selecting text that answers the question. With this methodology, QuAC gathers 98k questions across 13,594 conversations. We take their entire dev set and a sample of their train set and create a custom JavaScript task in Mechanical Turk that allows workers to rewrite these questions. JavaScript hints help train the users and provides automated, real-time feedback.

We provide workers with a comprehensive set of instructions and task examples. We ask them to rewrite the questions in natural sounding English while preserving the sentence structure of the original question. We discourage workers from introducing new words that are unmentioned in the previous utterances and ask them to copy phrases when appropriate from the original question. These instructions ensure that the rewrites only resolve conversation-dependent ambiguities. Thus, we encourage workers to create minimal edits; in Section 5.2, we take advantage of this to use BLEU for evaluating model-generated rewrites.

We display the questions in the conversation one at a time, since the rewrites should include only the previous utterance. After a rewrite to the question is submitted, the answer to the question is displayed. The next question is then displayed. This repeats until the end of the conversation. The full set of instructions and the data collection interface are provided in the appendix.

We apply quality control throughout our collection process. During the task, JavaScript checks automatically monitor and warn about common errors: submissions that are abnormally short (e.g., 'why'), rewrites that still have pronouns (e.g., 'he wrote this album'), or ambiguous words (e.g., 'this article', 'that'). Many QuAC questions ask about 'what/who else' or ask for 'other' or 'another' entity. For that class of questions, we ask workers to use a phrase such as 'other than', 'in addition to', 'aside from', 'besides', 'together with' or 'along with' with the appropriate context in their rewrite.

We gather and review our data in batches to screen potentially compromised data or low quality workers. A post-processing script flags suspicious

ORIGINAL: Was this an honest mistake by the media?
REWRITE: Was the claim of media regarding Leblanc's room come to true?
ORIGINAL: What was a single from their album?
REWRITE: What was a single from horslips' *album*?
ORIGINAL: Did they marry?
REWRITE: Did Hannah Arendt and Heidegger marry?

Table 2: Not all rewrites correctly encode the context required to answer a question. We take two failures to provide examples of the two common issues: Changed Meaning (top) and Needs Context (middle). We provide an example with no issues (bottom) for comparison.

rewrites and workers who take and abnormally long or short time. We flag about 15% of our data. *Every* flagged question is manually reviewed by one of the authors and an entire HIT is discarded if one is deemed inadequate. We reject 19.9% of submissions and the rest comprise CANARD. Additionally, we filter out under-performing workers based on these rejections from subsequent batches. To minimize risk, we limit the initial pool of workers to those that have completed 500 HITs with over 90% accuracy and offer competitive payment of $0.50 per HIT.

We verify the efficacy of our quality control through manual review. A random sample of fifty questions sampled from the final dataset is reviewed for desirable characteristics by a native English speaker in Table 1. Each of the positive traits occurs in 90% or more of the questions. Based on our sample, our edits retain grammaticality, leave the question meaning unchanged, and use pronouns unambiguously. There are rare occasions where workers use a part of the answer to the question being rewritten or where some of the context is left ambiguous. These infrequent mistakes should not affect our models. We provide examples of failures in Table 2.

We use the rewrites of QuAC's development set as our test set (5,571 question-in-context and corresponding rewrite pairs) and use a 10% sample of QuAC's training set rewrites as our development set (3,418); the rest are training data (31,538).

	Dev	Test
Copy	33.84	36.25
Pronoun Sub	47.72	47.44
Seq2Seq	51.37	49.67
Human Rewrites*	59.92	

Table 3: BLEU scores of the baseline models on development and test data. Seq2Seq improves up to four points over naive baselines but still well below human accuracy. Human accuracy (*) is computed from a small subset of the validation set.

4 Baselines

We compare three baseline models for the question-in-context rewriting task. In the **Copy** baseline, the rewrite q'_m is set to be the same as the input question q_m without making any changes.

We also try a **Pronoun Substitution** baseline in which the first pronoun in q_m is replaced with the topic entity of the conversation. We use the title of the corresponding Wikipedia article to the original QuAC conversation as the topic entity. Similar to the Copy baseline, the training data is not used in that baseline.

Unlike the previous baselines which do not use our rewrites as training data, the third baseline is a neural sequence-to-sequence (**Seq2Seq**) model with attention and a copy mechanism (Bahdanau et al., 2015; See et al., 2017). We construct the input sequence by concatenating all utterances in the history H, prepending them to q_m, and adding a special separator token between utterances. We use a bidirectional LSTM encoder-decoder model with shared the word embeddings between the encoder and the decoder.[2]

Since questions are written by humans, a human rewrites are the upper-bound for this task. However, annotators (especially crowdworkers) can be inconsistent or disagree. To estimate the human accuracy, we collect 100 pairs of rewritten questions; each pair has two rewrites of the same question (in its given context) by two different workers. We manually verify that all rewrites are valid and then use the pair of rewrites as a hypothesis and a reference.

Table 3 shows the BLEU scores produced by the baselines and humans over both the validation and the test sets.[3] Although a well-trained standard

[2] We initialize the embeddings with GloVE (Pennington et al., 2014) and train with a batch-size of 16 for 200000 steps. We use OpenNMT (Klein et al., 2018) implementation.

[3] We use multi-bleu-detok.perl (Sennrich et al., 2017)

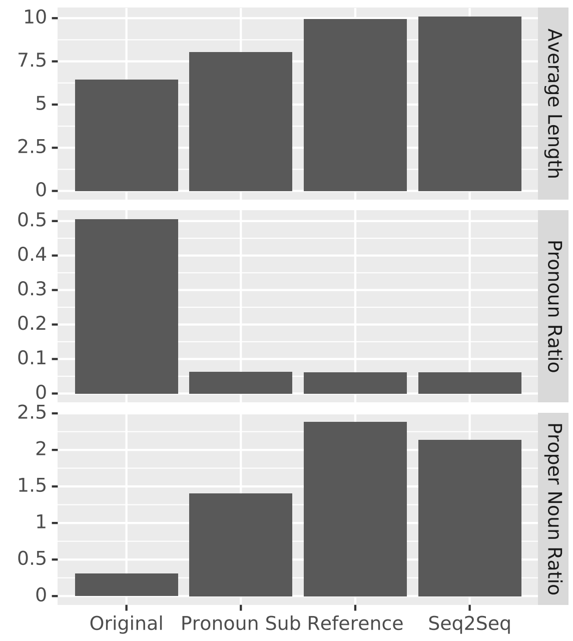

Figure 2: Human rewrites are longer, have fewer pronouns, and have more proper nouns than the original QuAC questions. Rewrites are longer and contain more proper nouns than our Pronoun Sub baseline and trained Seq2Seq model.

neural sequence-to-sequence improves 2–4 BLEU points over naive baselines, it is still 9 BLEU points below human-accuracy. We analyze sources of errors in the following section.

5 Dataset and Model Analysis

We analyze our dataset with automatic metrics after validating the reliability of our data (Section 3). We compare our dataset to the original QuAC questions and to automatically generated questions by our models. Then, we manually inspect the sources of rewriting errors in the seq2seq baseline.

5.1 Anaphora Resolution and Coreference

Our rewrites are longer, contain more nouns and less pronouns, and have more word types than the original data. Machine output lies in between the two human-generated corpora, but quality is difficult to assess. Figure 2 shows these statistics. We motivate our rewrites by exploring linguistic properties of our data. Anaphora resolution and coreference are two core NLP tasks applicable to this dataset, in addition to the downstream tasks evaluated in Section 4.

Pronouns occur in 53.9% of QuAC questions. Questions with pronouns are more likely to be ambiguous than those without any. Only 0.9% of these have pronouns that span more than one category (e.g., ‘she’ and ‘his’). Hence, pronouns within a single sentence are likely unambiguous. However, 75.0% of the aggregate history has pronouns and the percentage of mixed category pronouns increase to 27.8% of our data. Therefore, pronoun disambiguation potentially becomes a problem for a quarter of the original data. An example is provided in Table 4.

Label	Text
QUESTION	How long did he stay there?
REWRITE	How long did Cito Gaston stay at the Jays?
HISTORY	*Cito Gaston* **Q:** What did Gaston do after the world series? . . . **Q:** Where did he go in 2001? **A:** In 2002, he was hired by the Jays as special assistant to president and chief executive officer Paul Godfrey.

Table 4: An example that had over ten flagged proper nouns in the history. Rewriting requires resolving challenging coreferences.

Approximately one-third of the questions generated by our pronoun-replacement baseline are within 85% string similarity to our rewritten questions. That leaves two-thirds of our data that cannot be solved with pronoun resolution alone.

5.2 Model Analysis

By manually examining the predictions of the seq2seq model, we notice that the main source of errors is that the model tends to find a short path to completing the rewrites. That often results in *under-specified questions* as in Example 1 in Table 5, *question meaning change* as in Example 2 or *meaningless questions* as in Example 3.

Another source of errors is having related entities mentioned in the context as Example 4 in Table 5, where the model confused “Copa America” with “Argentina”. The model also struggles with listing multiple entities mentioned in different parts of the context. Example 5 in Table 5 show the output and the reference rewrites of the question *“Did she have any more works than those 3?”*, where two of the three entities—“United States of Banana”, “La Comedia” and “Asalto al tiempo”—are lost in the rewrite.

	Seq2Seq output	**Reference**
1	What did Chamberlain's men do?	What did Chamberlain's men do during the Battle of Gettysburg?
2	How many games did Ozzie Smith win?	How many games did the Cardinals win while Ozzie Smith played?
3	Did 108th get to the finals?	Did the US Women's Soccer Team get to the finals in the 1999 World Cup?
4	Did Gabriel Batistuta reside in any other countries, besides touring in the Copa America?	Besides Argentina, did Gabriel Batistuta reside in any other countries?
5	Did La Comedia have any more works than La Comedia 3?	Did Giannina Braschi have any more works than United States of Banana, La Comedia and Asalto al tiempo?

Table 5: Example erroneous rewrites generated by the Seq2Seq models and their corresponding reference rewrites. The dominant source of error is the model tendency to produce short rewrites (Examples 1–3). Related entities (Copa America and Argentina in Example 4) distract the model. The model struggles with listing multiple entities mentioned in different parts of the context (Example 5).

6 Related Work and Discussion

Recent work in CQA has used simple concatenation (Elgohary et al., 2018), sequential neural models (Huang et al., 2019), and transformers (Qu et al., 2019a) for modeling the interaction between the conversation history, the question and reference documents. Some of the components in those models, such as relevant history turn selection (Qu et al., 2019b), can be adopted in question rewriting models for our task. An interesting avenue for future work is to incorporate deeper context, either from other modalities (Das et al., 2017) or from other dialog comprehension tasks (Sun et al., 2019).

Parallel to our work, Rastogi et al. (2019) and Su et al. (2019) introduce utterance rewriting datasets for dialog state tracking. Rastogi et al. (2019) covers a narrow set of domains and the rewrites of Su et al. (2019) are based on Chinese dialog with two-turn fixed histories. In contrast, CANARD has histories of variable turn lengths, covers wider topics, and is based on CQA.

Training question rewriting using reinforcement learning with the task accuracy as reward signal is explored in retrieval-based QA (Liu et al., 2019) and in MRC (Buck et al., 2018). A natural question is whether reinforcement learning could learn to retain the necessary context to rewrite questions in CQA. However, our dataset could be used to pretrain a question rewriter that can further be refined using reinforcement learning.

More broadly, we hope CANARD can drive human-computer collaboration in QA (Feng and Boyd-Graber, 2019). While questions typically vary in difficulty (Sugawara et al., 2018), existing research either introduces new benchmarks of difficult (adversarial) stand-alone questions (Dua et al., 2019; Wallace et al., 2019, inter alia), or models that simplify hard questions through paraphrasing (Dong et al., 2017) or decomposition (Talmor and Berant, 2018). We aim at studying QA models that can ask for human assistance (feedback) when they struggle to answer a question.

The reading comprehension setup of CQA provides a controlled environment where the main source of difficulty is interpreting a question in its context. The interactive component of CQA also provides a natural mechanism for improving rewriting. When the computer cannot understand (rewrite) a question because of complicated context, missing world knowledge, or upstream errors (Peskov et al., 2019) in the course of a conversation, it should be able to ask its interlocutor, "can you unpack that?" This dataset helps start that conversation; the next steps are developing and evaluating models that efficiently decide when to ask for human assistance, and how to best use this assistance.

Acknowledgments

We thank Jo Shoemaker and Michelle Yuan for crowdsourcing feedback. This work was supported by NSF grant IIS-1822494. Boyd-Graber and Peskov were also supported by IIS-1748663. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *Proceedings of the International Conference on Learning Representations*.

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 Empirical Methods in Natural Language Processing*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Computer Vision and Pattern Recognition*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. Dataset and baselines for sequential open-domain question answering. In *Proceedings of Empirical Methods in Natural Language Processing*.

Shi Feng and Jordan Boyd-Graber. 2019. What AI can do for me: Evaluating machine learning interpretations in cooperative play. In *International Conference on Intelligent User Interfaces*.

Hsin-Yuan Huang, Eunsol Choi, and Wen tau Yih. 2019. FlowQA: Grasping flow in history for conversational machine comprehension. In *Proceedings of the International Conference on Learning Representations*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of Association for Machine Translation in the Americas*.

Ye Liu, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2019. Generative question refinement with deep reinforcement learning in retrieval-based QA system. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Denis Peskov, Joe Barrow, Pedro Rodriguez, Graham Neubig, and Jordan Boyd-Graber. 2019. Mitigating noisy inputs for question answering. In *Proceedings of the Annual Conference of the International Speech Communication Association*.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. BERT with history modeling for conversational question answering. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the ACM International Conference on Information and Knowledge Management*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Pushpendre Rastogi, Arpit Gupta, Tongfei Chen, and Lambert Mathias. 2019. Scaling multi-domain dialogue state tracking via query reformulation. In *Conference of the North American Chapter of the 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*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Association for Computational Linguistics*.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the Conference of the European Chapter of the Association for Computational Linguistics*.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance

ReWriter. In *Proceedings of the Association for Computational Linguistics*.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of Empirical Methods in Natural Language Processing*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. In *Transactions of the Association for Computational Linguistics*.

Bonnie Webber. 1992. Question answering. In Stuart C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, 2nd edition, pages 814–822. John Wiley & Sons, Inc., New York, NY.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.