ACL Anthology

Presentation on

Fluent Response Generation for Conversational Question Answering

Author(s): Ashutosh Baheti, Alan Ritter, Kevin Small

Presentation by:
Sabbir Ahmed Sibli, 20266027
Nuray Jannat, 20266028
M.Sc. in CSE
BRAC University

01

About the Research Paper

02

Fluent Conversational QA

- Definition.
- Case Study

03

Proposed Model

- Syntactic Transformation (ST).
- BERT-based Response Classifier.
- Initial Evaluation: BERT vs Baseline models.
- Sequence to Sequence (Seq2Seq) Dialogue Model.
- Final Evaluation: Seq2Seq vs Baseline models.

04

In the End

- Example Responses.
- Summary of the Contributions

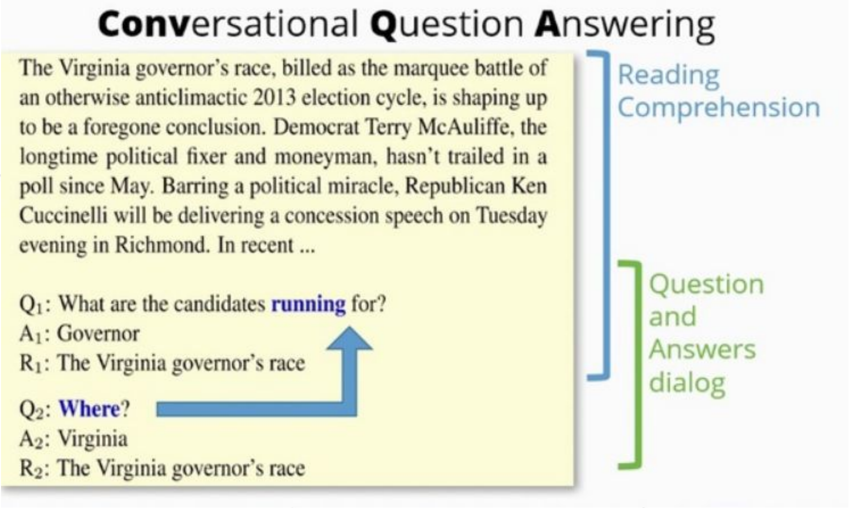
01 About the Research Paper

Paper Title: Fluent Response Generation for Conversational Question Answering

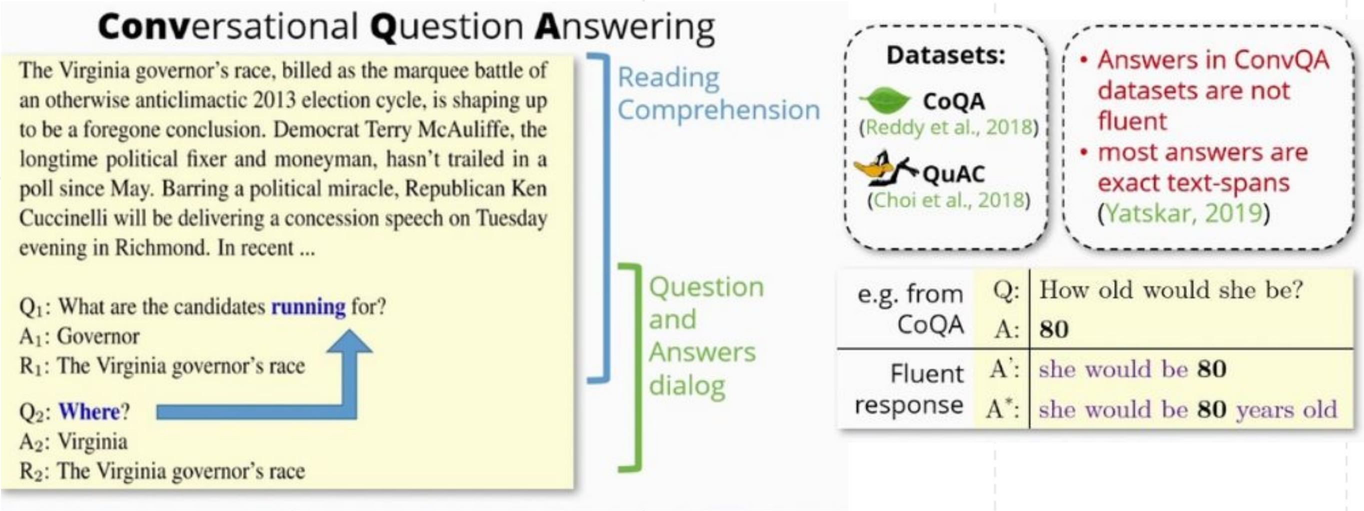
- **Author(s):**
  - *Ashutosh Baheti*, Computer Science & Engineering, Ohio State University.
  - *Alan Ritter*, Computer Science & Engineering, Ohio State University.
  - *Kevin Small*, Amazon Alexa

- **Conference:** Annual Meeting of the Association for Computational Linguistics. ACL Anthology
- **Year:** 2020.
- **Publisher:** arXiv preprint arXiv:2005.
- **Author Email(s):** {baheti.3, ritter.1492}@osu.edu, smakevin@amazon.com

02 Fluent Conversational QA: Definition



02 Fluent Conversational QA: Definition



02 Fluent Conversational QA: Definition

Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q_1: What are the candidates **running** for?
A_1: Governor
R_1: The Virginia governor's race

Q_2: **Where**?
A_2: Virginia
R_2: The Virginia governor's race

Reading Comprehension

Question and Answers dialog

Datasets:

- CoQA (Reddy et al., 2018)
- QuAC (Choi et al., 2018)

- Answers in ConvQA datasets are not fluent
- most answers are exact text-spans (Yatskar, 2019)

e.g. from CoQA	Q:	How old would she be?
	A:	**80**
Fluent response	A':	she would be **80**
	A*:	she would be **80** years old

Data augmentation + Neural Dialog models → Fluent Response Generation in ConvQA

Preview of our generation model

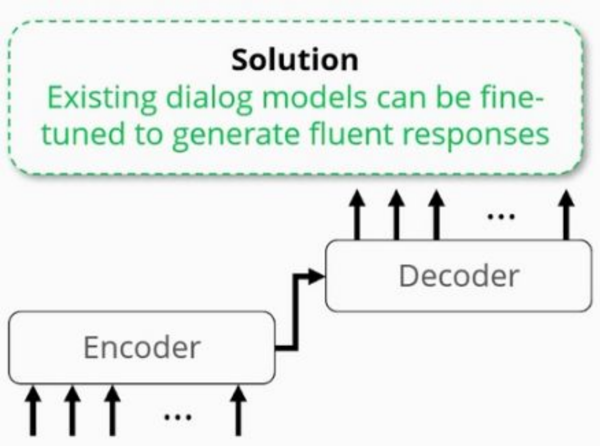
Q :	what revolt did he lead after that ?
CoQA:	**autumn harvest uprising**
Ours:	he led the **autumn harvest uprising**

02 Fluent Conversational QA: Case Study

Problem Statement: Given a question ***q*** and the answer span ***a***, have to generate an answer ***r*** which is right, fluent and grammatically correct.

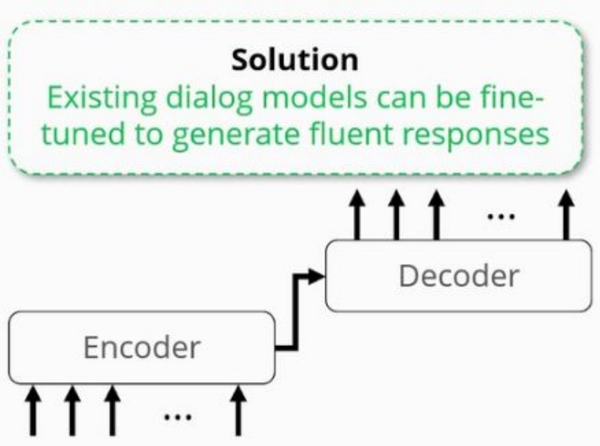
02 Fluent Conversational QA: Case Study

Problem Statement: Given a question ***q*** and the answer span ***a***, have to generate an answer ***r*** which is right, fluent and grammatically correct.



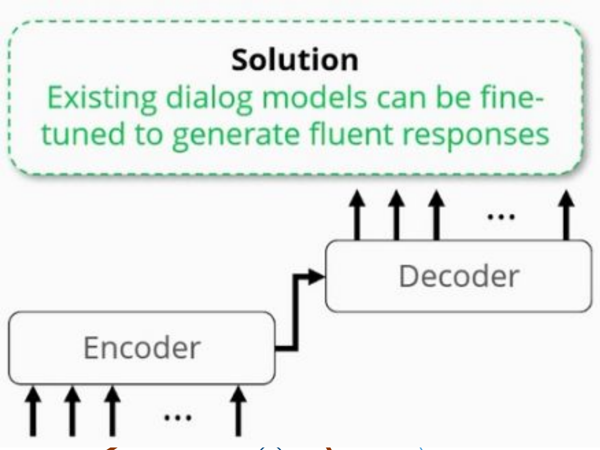
02 Fluent Conversational QA: Case Study

Problem Statement: Given a question ***q*** and the answer span ***a***, have to generate an answer ***r*** which is right, fluent and grammatically correct.



02 Fluent Conversational QA: Case Study

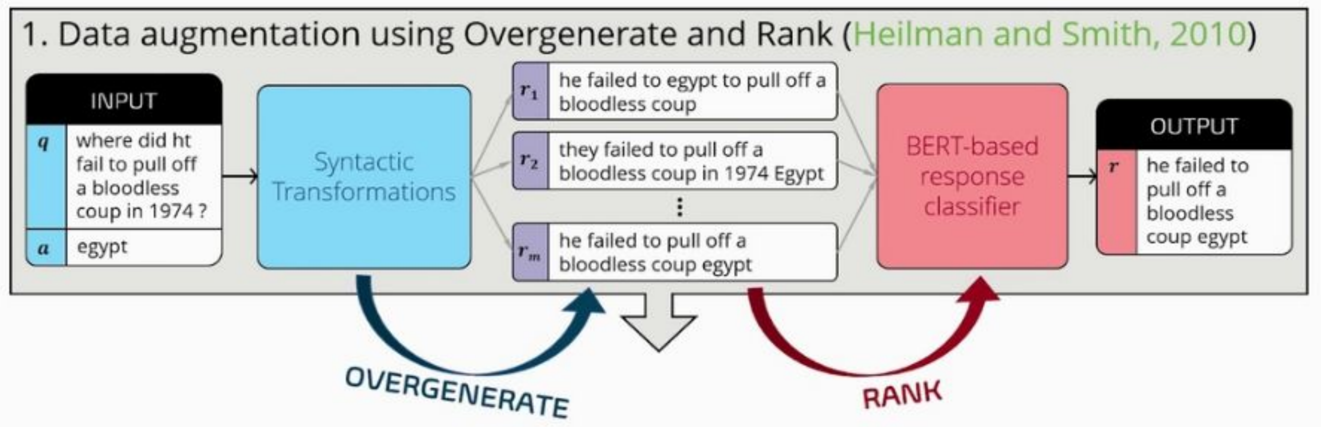
Problem Statement: Given a question ***q*** and the answer span ***a***, have to generate an answer ***r*** which is right, fluent and grammatically correct.



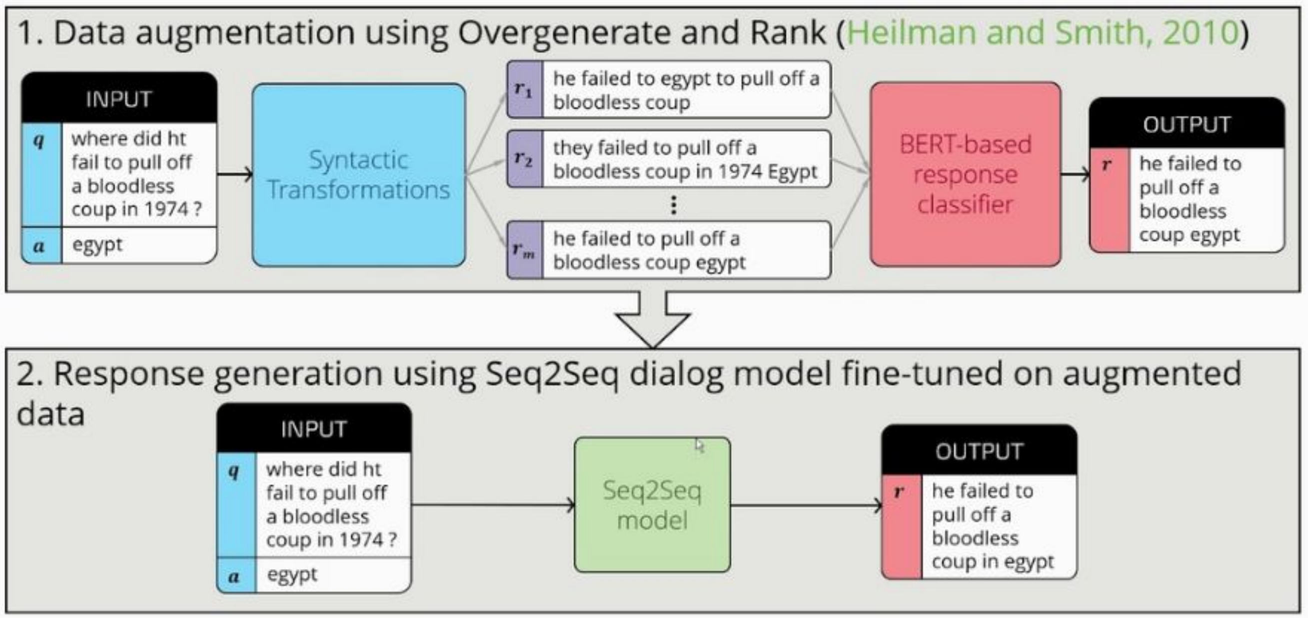
Crowdsourcing a fluent QA dataset is expensive

Transform existing QA dataset to support fluent answer-responses using **data augmentation**!

03 Proposed Model



03 Proposed Model



03 Syntactic Transformation

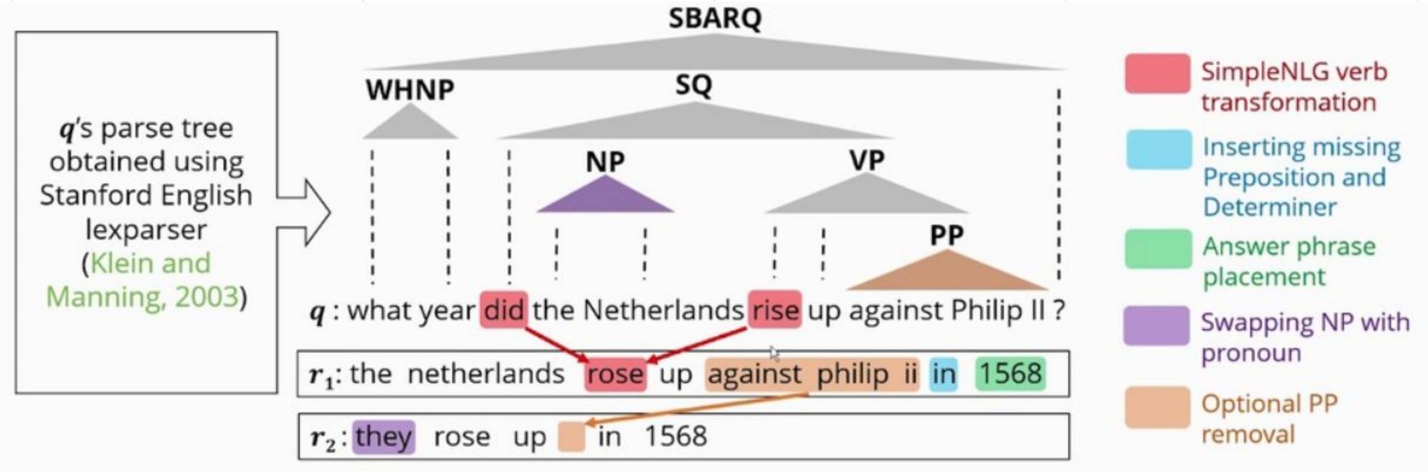
Multiple Syntactic Transformations (STs) on questions (***q***) builds parse tree to generate a huge list of candidate responses,

$$\boldsymbol{R}=\{r1, r2, \ldots, rm\}$$

03 Syntactic Transformation

Multiple Syntactic Transformations (STs) on questions (***q***) builds parse tree to generate a huge list of candidate responses,

R={r1, r2, … , rm}



For the dataset used in this paper, this 5 properties of STs can generate **m>5000** responses

03 BERT based Classifier

BERT based sentence pair classifier (***F***) used to predict the most suitable response (***r***) for a question (***q***) from the candidate responses (***R***).

03 BERT based Classifier

BERT based sentence pair classifier (***F***) used to predict the most suitable response (***r***) for a question (***q***) from the candidate responses (***R***).

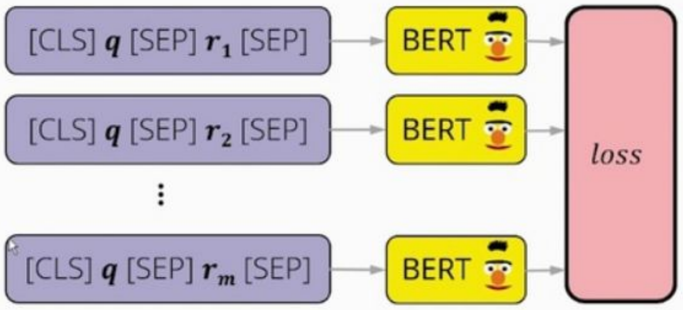
Model Trained using objective based error margins (***M***)

$$M_j(F) = F(q, a, r_1) - F(q, a, r_j)$$

$$Softmax\ loss = \log(1 + \sum_{j=2}^{m} e^{-M_j(F)})$$

Key idea:

Unsuitable responses should be ranked lower than the suitable ones



03 BERT based Classifier

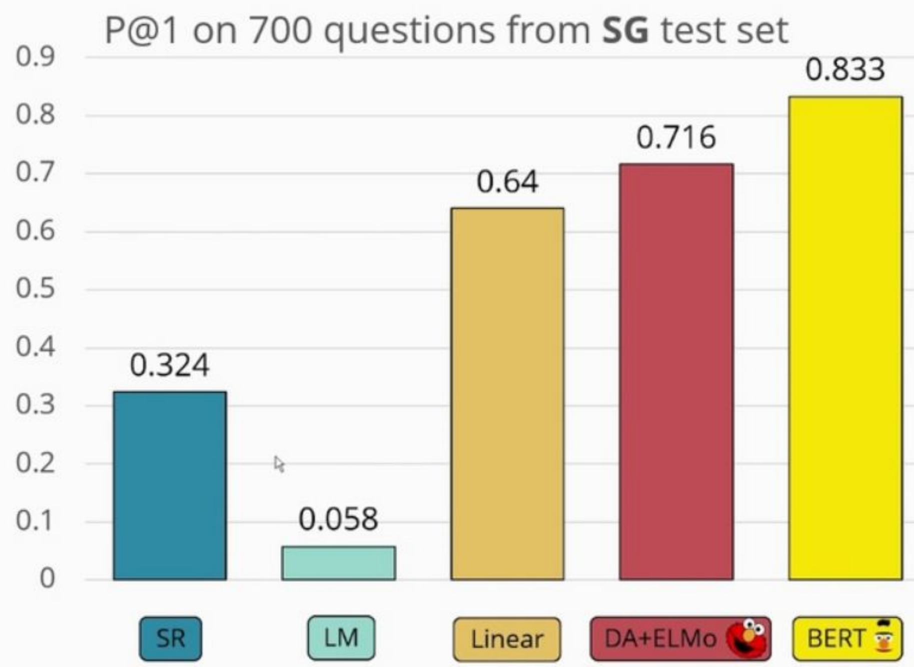
BERT based sentence pair classifier (***F***) used to predict the most suitable response (***r***) for a question (***q***) from the candidate responses (***R***).

Training data Acquisition

1. Generate candidate responses for $(\boldsymbol{q}, \boldsymbol{a})$ from SQuAD 2.0 (Rajpurkar et al., 2018)
2. crowd-annotate best responses for ***3000*** $(\boldsymbol{q}, \boldsymbol{a})$ to get SQuAD Gold (**SG**) dataset
3. Split **SG** into ***2000/300/700*** to get ***train/dev/test*** segments

03 BERT vs Other Baseline Models

Abbr.	Sentence Pair Classification Model
SR	Shortest Response
LM	Language Model
Linear	Linear classifier using features inspired from (Heilman and Smith, 2010) and (Wan et al., 2006)
DA + ELMo	Decomposable Attention (Parikh et al., 2016) with ELMo (Peters et al., 2018) embedding
BERT	$\text{BERT}_{\text{BASE}}$ uncased (Devlin et al., 2019) response classifier



03 Sequence to Sequence (Seq2Seq) Dialogue Model

Why S2S is **NEED** even after ST+BERT can generate expected output?

1. Parser fails to recognize questions ~20% of the time
2. All the candidate responses generated by STs can be incorrect!
3. Extending STs requires lot of manual efforts

03 Seq2Seq vs Baseline Models

Using STs+BERT on (***q,a***) pairs from SQuAD Dataset, Natural Questions Dataset and HarvestingQA dataset, almost **1 million** instances for **SNH** data has been achieved.

Seq2Seq Dialogue Model	Dataset Details
Pointer Generator Network (**PGN**)	14 millions question-only subset of OpenSubtitles Dataset
Generative Pretrained Transformer (**GPT-2**)	
Dialogue GPT-2 (**DGPT**)	147 millions Reddit conversation

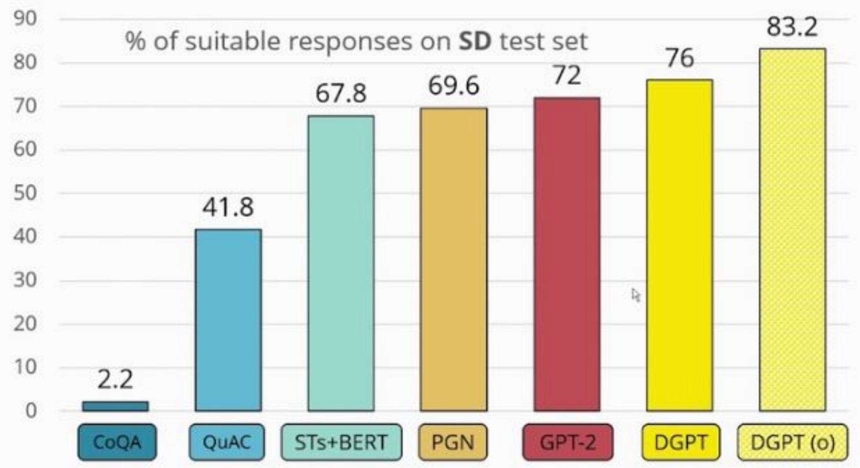
Baseline Models
BERT-based model trained on **CoQA** dataset
BiDAF model trained on **QuAC** dataset
Data augmentation used to create SNH data to train **STs+BERT** model

03 Seq2Seq vs Baseline Models

Evaluation on a sample of 500 test instances from SQuAD 2.0 Dataset, named SD test Data

Human annotators judge a response if it is suitable, grammatically correct, complete sentence with the fluent correct answers or not.

03 Seq2Seq vs Baseline Models



Most answers from **CoQA** and **QuAC** are either exact-answer spans or answer-spans with few surrounding words

~18% of **STs + BERT** responses are exact answer spans (parser failures)

All **Seq2Seq** models outperform **STs+BERT** baseline thanks to dialog pretraining

With oracle answer-spans **DGPT** can do even better!

04 In the End: Example Responses

	Question	**_Which sea was oil discovered in?_**
Answers	CoQA	North Sea
	QuAC	“It’s scotland’s oil” campaign of the scottish national party (snp)
	STs+BERT	Oil was discovered in north
	DGPT	It was discovered in the north sea

04 In the End: Summary of the Contributions

- Data Augmentation method: STs+BERT classifier can transform existing datasets to generate fluent responses.
- Dialogue GPT model outperforms the STs+BERT augmentation by exploiting dialogue pretraining.
- Proposed model in this research doesn’t require any QA passage to generate fluent responses.

“When you don’t understand, it’s sometimes easier to look like you do as you like.”

— Malcolm Forbes

REFERENCES

[1] 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.

[2] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau 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.

[3] Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

[4] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In Proc. ACL.

[5] 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.

[6] Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. Computational Linguistics, 31(1):25–70.

[7] 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.

[8] Ankur Parikh, Oscar Tackstr ¨ om, Dipanjan Das, and ¨ Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

[9] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[10] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog, 1(8).

[12] Jorg Tiedemann. 2009. News from opus-a collection ¨ of multilingual parallel corpora with tools and interfaces. In Recent advances in natural language processing, volume 5, pages 237–248.

[13] Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

[14] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.