Answering Naturally : Factoid to Full length Answer Generation

Vaishali Pal
LTRC, IIIT-H, Hyderabad
vaishali.pal@research.iiit.ac.in

Irshad Bhat
IMS, University of Stuggart
bhatid@ims.uni-stuttgart.de

Manish Shrivastava
LTRC, IIIT-H, Hyderbad
m.shrivastava@iiit.ac.in

Abstract

In recent years, the task of Question Answering over passages, also pitched as a reading comprehension, has evolved into a very active research area. A reading comprehension system extracts a span of text, comprising of named entities, dates, small phrases, etc., which serve as the answer to a given question. However, these spans of text would result in an unnatural reading experience in a conversational system. Usually, dialogue systems solve this issue by using template-based language generation. These systems, though adequate for a domain specific task, are too restrictive and predefined for a domain independent system. In order to present the user with a more conversational experience, we propose a pointer generator based full-length answer generator which can be used with most QA systems. Our system generates a full-length answer given a question and the extracted factoid/span answer without relying on the passage from where the answer was extracted. We also present a dataset of 315,000 question, factoid answer and full-length answer triples. We have evaluated our system using ROUGE-1,2,L and BLEU and achieved 74.05 BLEU score and 86.25 Rogue-L score.

1 Introduction

Factoid question answering (QA) is the task of extracting answers for a question from a given passage. These answers are usually short spans of text, such as named entities, dates, etc. Modern factoid QA systems which use machine-comprehension datasets, predict the answer span from relevant documents using encoder-decoder architectures with co-attention. Conversely, knowledge-base (KB) oriented QA systems retrieve relevant facts using structured queries or neural representation of the question. Formulating the retrieved factoid answer into a full-length natural sentence is, hence, a natural extension and post-processing step of any QA system.

System Input:
Question : When were the normans in normandy?
Factoid Answer : 10th and 11th centuries
System Output :
During the 10th and 11th centuries , the normans were in normandy.

Table 1: Full-length natural answer generation from the question and the factoid answer

A simple approach for this task might be to use hand-crafted rules to restructure the question into a declarative statement as described in (Jurafsky and Martin, 2018). However, such rule based approaches fail when the extracted answer span, contains words from the question or when there are multiple independent clauses and the system has to choose words specific to the question to formulate the answer. This leads to unnatural repetition of words in the full-length answer or grammatically incorrect sentence formulation.

On the other hand, neural-network based approaches in modern dialogue systems use end-to-end encoder-decoder architectures to convert an abstract dialogue action into natural language utterances. Such modern task-oriented dialogue systems usually learn to map dialogue histories to system response. Non-task oriented dialogue systems such as generative systems can formulate responses not present in the training data but lacks the capability to incorporate factual information without external knowledge bases.

Unlike conversational chat-bots designed to mimic human conversation without the need to be factually correct, or task-oriented dialogue systems which place the retrieved answer in a predefined template, our system automatically gener-

Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 1–9
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

ates accurate full-length answers, thereby, enhancing the system's usage in these situations. Table 1 shows a sample of our system input and output. Our system can be used in any such task-specific scenarios where natural answers are desired, without being restricted to a limited set of templates.

Our overall research contributions are listed as follows:

- We introduce a system which generates factually correct full-length answers from the questions and the factoid answers. Our system can be used as a post-processing plug-in to any QA system, be it a KB-based system or machine comprehension based system, thereby improving readability of the system output and promoting fluency and variation in the natural answer generation.
- We have also released a dataset comprising of tuples of questions, factoid answers and full-length answers which can be further augmented using any other QA datasets using the techniques we describe in section 3.1.

2 Related Work

There has been a lot of interest recently in QA and task-oriented dialogue systems. End-to-end memory networks (Sukhbaatar et al., 2015) use a language modelling architecture which learns query embeddings in addition to input and output memory representations from source sequences and predicts an answer. Rule based systems such as (Weston et al., 2015) sets up a variety of tasks for inferring and answering the question. (Bordes and Weston, 2016) improves on the memory networks and handles out-of-vocabulary (OOV) words by inserting special words into the vocabulary for each knowledge base entity types. These systems are dependent on templates or special heuristics to reproduce facts. We demonstrate through our baseline model that generating template-like sentences from factual input can be achieved with limited success.

Recent works on KB-based end-to-end QA systems such as (Yin et al., 2015; He et al., 2017a; Liu et al., 2018a) generate full-length answers with neural pointer networks(Gülçehre et al., 2016; Vinyals et al., 2015; He et al., 2017b) after retrieving facts from a knowledge base (KB). Dialogue systems such as (Liu et al., 2018b; Lian et al., 2019) extract information from knowledge bases to formulate a response. Systems such as (Fu and Feng, 2018) uses KB based key-value memory after extracting information from documents or external KBs. However, these systems are restricted to only information modeled by the KB or slot-value memory. Our system, is generic and can be used with any knowledge source, structured such as a knowledge base or free form such as machine-comprehension dataset. Since our system doesn't use any additional relational information as modelled in a KB, it is invariant to the type of dataset. The pointer generator network, introduced in (See et al., 2017), is a generative summarization model that can copy out-of-vocabulary (OOV) words from a source sequence. Our work is inspired from the ability of this network to accurately reproduce information from source.

To the best of our knowledge, there is no existing QA data-set which addresses the task directly. However, Knowledge-based QA dataset such as (Yin et al., 2015) creates a knowledge-base from Chinese websites and extracts question-answer pairs from Chinese communityQA webpage. The system built over this dataset, is able to generate natural answers to simple questions. The recently released CoQA dataset(Reddy et al., 2018) is an abstractive conversational question answering dataset through which the system generates free-form answers from the whole conversational history using the aforementioned pointer-generator network. While the CoQA challenge extracts free-form text from the passages, our system incorporates the structure of the question to give a full-length sentence as answer to the given query.

3 Data

Since there is no available dataset for the task, we used the standard machine comprehension datasets such as SQuAD (Rajpurkar et al., 2016) and HarvestingQA (Du and Cardie, 2018) to create auto-annotated data. This provide us with questions and factoid answers which we use as input to our system. For the ground-truth, we automatically extract full-length answers from the passages of these datasets by applying certain heuristics (explained in section 3.1). We extract ∼300,000 samples (question, factoid answer, full-length answer) from SQuAD and HarvestingQA. Additionally, we have manually annotated 15000 samples from SQuAD of which 2500 are used for development, 2500 for testing and we augment the rest

10000 with the auto-annotated data.

3.1 Automatic Data Generation

Creating datasets for any new task is a challenge since modern systems based on neural architectures requires a large amount of data to train. To make the data creation task scalable, most of our training data is automatically generated from SQuAD and HarvestingQA. For each question-answer pair, we automatically extract the target full-length answers from corresponding passages. We iterate over the sentences in the context passage that contain the factoid answer and select the one that has the highest BLEU score with the question, given $BLEUscore \geq 35\%$. Given the question-answer pair (Q, A) and the passage P, the full-length answer T is the sentence, S, in the passage:

$$\begin{aligned} T = \underset{S \in P}{\operatorname{argmax}}\, BLEU(Q, S) \\ iff\ A \in S\ \ \&\ \ BLEU(Q, S) \geq 35\% \end{aligned} \tag{1}$$

The target sentences having a low BLEU score(between $35\% - 50\%$) may not be completely aligned with the question but provide sufficient information to train the system to generate full-length sentences containing the factoid answer.[1] As the whole sentence is extracted from the corresponding passage, these samples may also contain additional information from the passage which is not related to the question.

Our method of automatically extracting samples from existing QA datasets is scalable and can be reproduced with any modern QA datasets to generate more samples to augment our auto-generated samples extracted from HarvestingQA and SQuAD. The table 2 shows some auto-generated samples from the dataset. Our auto-generated data samples follow a similar question distribution as SQuaD and is biased towards what" and "who" questions as shown in the trigram distribution of the questions in figure 1.

Question : what is the name of the term that is used in the united states ? **Factoid** : great plains **Target** : the term great plains is used in the united states to describe a sub-section of the even more vast interior plains physiographic division
Question : who is the only country among the united nations security council ? **Factoid** : germany **Target** : germany is the only country among the top five arms exporters that is not a permanent member of the united nations security council .
Question : what lake is now connected to the sea ? **Factoid** : lake voulismeni **Target** : lake voulismeni at the coast , at aghios nikolaos , was formerly a sweetwater lake but is now connected to the sea .
Question : what is a bus driving on this route ? **Factoid** : the capacity of the lane will be more and will be more and will increase when the traffic level increases **Target** : when there is a bus driving on this route , the capacity of the lane will be more and will increase when the traffic level increases .

Table 2: Automatically created dataset samples

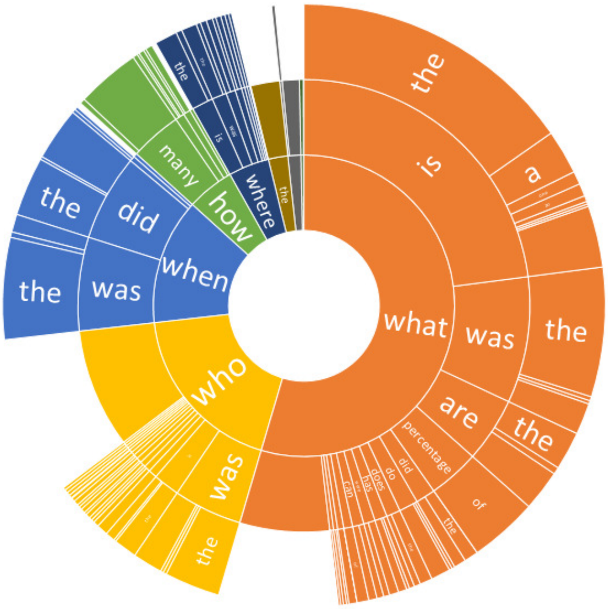

Figure 1: Question trigram distribution of automatically created dataset

3.2 Manual Data Generation

The auto-generated samples contain extra information in the ground-truth full-length sentences which are not aligned with the question or factoid answer. To refine our dataset to be more attuned to questions and also to capture the variability humans bring when generating new sentences, we manually annotated 15000 QA pairs, from the

[1] We found that samples with BLEU score of less than 35 were significantly noisy.

SQuAD dataset. We used multiple ways to answer the same question, such as in active and passive voice, to incorporate more variation to the target sentences. Apart from generating samples with the full-length answers well aligned with the question, we have also chosen complex samples from SQuAD which have long phrasal factoid answers to add more complexity to the data samples. These samples have sentential factoid answers containing more than one independent clause which are not present in the ground-truth full-length natural answer. The inclusion of such examples is to aid to the system to learn to only choose words which are required to form a syntactically correct answer and omit other synonymous or superfluous words. The table 3 shows some manual generated samples. The manual samples contains questions more evenly distributed than the auto-generated ones as shown in the figure 2 displaying the trigram distribution of questions.

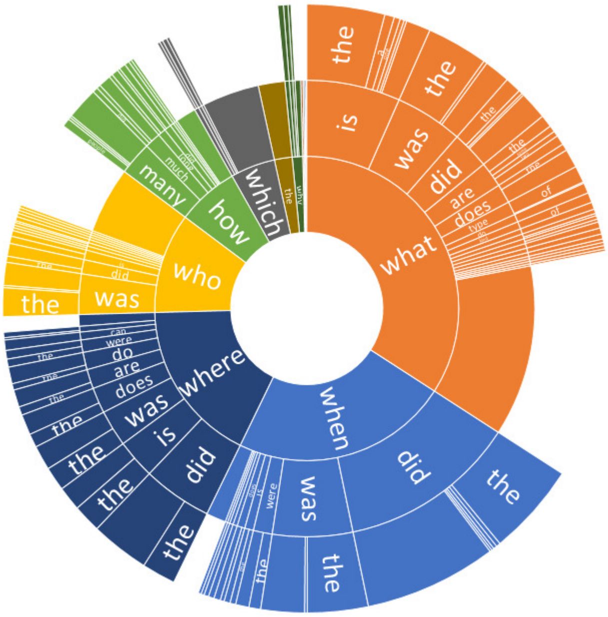

Figure 2: Question trigram distribution of manually created dataset

4 System Architecture

We framed the problem of generating full-length answer from the question and the factoid answer into a Neural Machine Translation (NMT) task using two approaches. We built a model based on the pointer-generator architecture described in (See et al., 2017) except we use two encoders on the source side to encode question and factoid answer separately as shown in Figure 3.

Question	How much more were her earnings that the year before?
Factoid	more than double her earnings
Target 1	Her earnings were more than double than that of the year before.
Target 2	She earned more than double her earnings than that of the year before.
Question	How many digital copies of her fifth album did Beyonc sell in six days?
Factoid 1	one million
Factoid 2	one million digital copies
Target	Beyonc sold one million digital copies of her fifth album in six days.
Question	How well did Kanye do in high school?
Factoid	A's and B's
Target	Kanye did well in high school by scoring A's and B's.
Question	What do scholars recognize about the life of the Buddha?
Factoid	Most accept that he lived, taught and founded a monastic order
Target	Most scholars recognize and accept that Buddha lived, taught and founded a monastic order.
Question	Where did english and scotch irish descent move to florida from?
Factoid	English descent and americans of scots-irish descent began moving into northern florida from the backwoods of georgia and south carolina
Target	English and Scotch Irish descent moved to Florida from the backwoods of Georgia and South Carolina.

Table 3: Manual dataset samples

Let the question be represented by words $Q = \{q_1, q_2, ..., q_n\}$. Let the factoid answer be represented by words $A = \{a_1, a_2, a_3, ..., a_m\}$.

We encode the question and answer sequence using two 3-layered bidirectional LSTMs which share weights. This produces two sequences of hidden states

$$h_Q^t = BILSTM(h_Q^{t-1}, q_t) \tag{2}$$

$$h_A^t = BILSTM(h_A^{t-1}, a_t) \tag{3}$$

We choose to encode the source sequences separately, since there is no syntactic connection between the question and the factoid answer. We

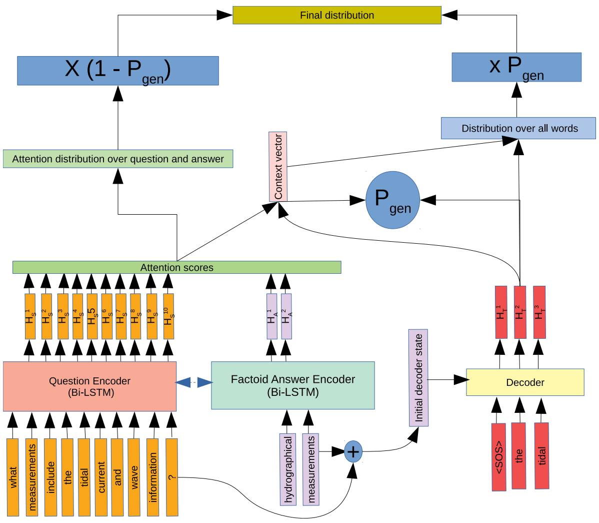

Figure 3: The 2 encoder pointer generator uses the question and factoid answer as input to generate a full-length answer in an end-to-end learning environment.

then stack together the encoded hidden states of the 2 encoders to produce a single list of source hidden states, $h_S = [h_Q; h_A]$. The decoder is initialized with the combined final states of the two encoders as

$$h_T^0 = h_Q^n + h_A^m \tag{4}$$

Following the global attention mechanism described in (Luong et al., 2015), context vector, C_t, is generated. For each decoder state, h_T^t, at time t, the alignment score, $a(h_T^t, h_S^i)$, with each encoder state, h_S^i, is calculated as follows:

$$a(h_T^t, h_S^i) = softmax(h_T^t W_a h_S^i) \tag{5}$$

The challenge to correctly reproduce factual information in the full-length answer led us to use copy attention from the pointer generator network as described in (See et al., 2017). The copy distribution, using an extended vocabulary comprising of source words, will capture the probability of replicating words from either the question or answer, whereas the global attention distribution has the ability to generate new words from the vocabulary. The final probability of predicting a word is as follows:

$$P(W_{final}) = p_g P_{gen} + (1 - p_g) P_{copy} \tag{6}$$

The parameter, p_g, is learned as

$$\sigma(W_c C^t + W_{h_t} h_T^t + W_x X^t) \tag{7}$$

where C^t is the context vector and X^t is the input to the decoder. We calculate the copy distribution, a distribution over the source words, $w = Q \cup A$:

$$P_{copy}(w) = \sum_{i:w_i = w} a(h_T^t, h_S^i) \tag{8}$$

The final probability of generating a word is as shown in equation 6. For out-of-vocabulary words which are present only in the source $w \in (Q \cup A)$ and $w \notin V$, only P_{copy} is used predict the word. These words are usually factual information from the question or answer, such as dates and named entities and hence needs to be copied exactly as it appears in the source sequences. Prepositions, conjunctions and other placeholders, such as at, $between$, in, which help in combining the question and answer sequences are usually in-vocab words not present in the source ($w \notin (Q \cup A)$ and $w \in V$), and are predicted with P_{gen}. For in-vocabulary words which are present in the source, $w \in (Q \cup A)$ and $w \in$ V, the final probability of predicting the word uses both the terms of equation 6.

5 Experiments

For all our experiments, we used a 6GB 1060TX Nvidia GPU. We trained the system on batch size of 32, dropout rate of 0.5, RNN size of 512 and decay steps 10000. Since, our dataset is small, we shared the vocabulary between source and target. We used pre-trained GloVe embeddings (300 dimension) to initialize both the encoder and decoder words. Since our manually created samples are less, we oversampled the manually annotated data 3 times to mitigate any bias introduced by the synthetic dataset. We have built our system over the OpenNMT-pytorch code base(Klein et al., 2017). We have tested our models independently on both the manual dataset and auto-created dataset. We have used 2500 samples of the manually annotated SQuAD data set and 3284 samples of the auto-generated dataset to evaluate the models' performance. These samples were selected randomly from the respective datasets. To evaluate the effectiveness of the manual data samples, we have compared the performance of our 2-encoder pointer-generator network trained on the auto-generated data and on the whole augmented dataset, containing both the manual and auto-generated data. For this comparison, training on the whole augmented data instead of only the

Model	Training Dataset	BLEU	ROGUE-1	ROGUE-2	ROGUE-L
Seq2Seq+Attention+Mask	Augmented	62.2	86.23	72.23	79.52
2 Encoder Pointer-Gen	Auto-only	67.5	87.94	77.85	82.77
2 Encoder Pointer-Gen	Augmented	**74.05**	**91.24**	**81.91**	**86.25**
Seq2Seq+Attention+Mask	Augmented	71.10	90.03	81.82	85.09
2 Encoder Pointer-Gen	Auto-only	73.63	91.50	85.02	87.56
2 Encoder Pointer-Gen	Augmented	**73.69**	**91.65**	**84.98**	**87.40**

Table 4: The top section displays BLEU and ROGUE scores for the models tested on the manually created test dataset. The bottom section displays the scores for the models tested on the auto-created test dataset. (All scores are in the range of 0-100)

Model	Training Dataset	BLEU	ROGUE-1	ROGUE-2	ROGUE-L
2 Encoder Pointer-Gen	Auto-only	71.54	92.64	82.31	90.06
2 Encoder Pointer-Gen	Augmented	**73.29**	**95.38**	**87.18**	**93.65**
2 Encoder Pointer-Gen	Auto-only	64.67	91.17	75.58	82.87
2 Encoder Pointer-Gen	Augmented	**75.41**	**93.46**	**82.29**	**87.50**

Table 5: The top section displays the scores for the models tested on the 500 randomly chosen NewsQA dataset. (All scores are in the range of 0-100). The bottom section displays BLEU and ROGUE scores for the models tested 900 randomly chosen Freebase test samples.

manual data is required due to the limited number of samples(15000) of the manual annotated data. We have compared our system with a Seq2Seq model with attention where only the question and full-length answer are considered as source and target to the model respectively. We mask the factoid answer in the target full-length answer with the string *a-n-s-w-e-r*. The mask, which acts as a placeholder to the factoid answer, is replaced with the actual factoid answer in a post-processing step. The masking in the data copes with the named entities and other OOV words in the dataset.

We have also performed cross-dataset evaluation on a knowledge base dataset(Freebase) and a machine comprehension dataset(NewsQA) to test the generalization capability of our system. We randomly selected 900 samples, comprising of question and object-names(factoid answers), from the test samples provided by SimpleQA(Golub and He, 2016) which were extracted from the KB dataset Freebase(Bollacker et al., 2008). We also randomly extract 500 test samples, questions and factoid answers, from the machine comprehension NewsQA(Trischler et al., 2017) dataset. The system predictions were compared with the manually annotated ground-truth full-length answers for these samples.

Model	Training Dataset	Acc
2-Enc Pointer-Gen	Synthetic-only	83.4
2-Enc Pointer-Gen	Augmented	**92.8**

Table 6: Accuracy Scores(in the range of 0-100) for the various models

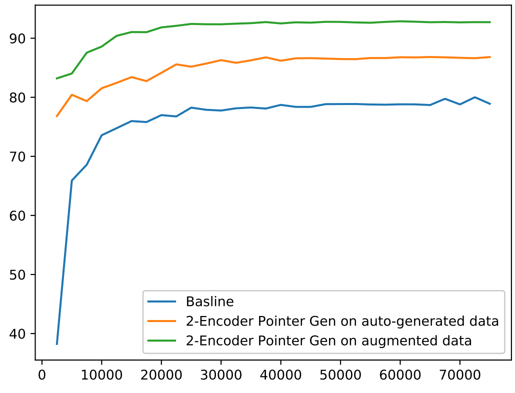

Figure 4: Validation Accuracy

6 Results

As shown in table 4, 5, 6 and 7, augmenting the manually annotated data with the auto-generated data for training leads to significant improvements for the 2-encoder pointer generator network. From our best assumption, this is not only due to cleaner samples in the manually annotated data which

Question : who was the eldest son of alfonso iii and what did he become king of? **Factoid Answer :** garca , became king of len **Target :** the eldest son of alfonso iii was garca and he become king of len. **Seq2Seq+Attention+Mask:** he became king of garca , became king of len. **Modified PointerGen :** the eldest son of alfonso iii was garca and he become king of len.
Question : where does the catalan word alfabia come from? **Factoid Answer :** of arabic origin **Target :** the catalan word alfabia is of arabic origin. **Seq2Seq+Attention+Mask:** the catalan word alfabia comes from of arabic origin . **Modified PointerGen :** the catalan word alfabia is of arabic origin .
Question : what job does debra byrd do on american idol? **Factoid Answer :** vocal coach **Target :** debra byrd is a vocal coach on american idol. **Seq2Seq+Attention+Mask:** amy byrd has vocal coach on american idol. **Modified PointerGen :** debra byrd is the vocal coach on american idol.
Question : when did the yuan dynasty start and end? **Factoid Answer :** 1271 to 1368 **Target :** the yuan dynasty ruled from 1271 to 1368 **Seq2Seq+Attention+Mask:** the yuan dynasty started and ended in 1271 to 1368 **Modified PointerGen :** the yuan dynasty started in 1271 to 1368

Table 7: Comparison of predictions of the Seq2Seq+Attention+Mask and Augmented Pointer generator systems. Example 1 depicts non-contiguous factoid answer which have to be interleaved in the full-length answer. Example 2 shows that the pointer generator is able to suppress conflicting preposition Example 3 depicts that masking is unable to handle named entities in the question where they are not masked. Masking is also unable to capture contextual information while formulating the natural sentence as depicted in Example 4

does not contain extra unnecessary information, but also samples with variations in the factoid and full-length groundtruth. The manual data also has long phrasal factoid answers from which the system has to learn to copy and generate words as needed. Table 7 shows that the pointer-generator system handles tense agreement and generation of new words. The Seq2Seq model suffers to capture contextual information, resolve anaphora, or reproduce factual information and handle out-of-vocabulary words. As shown in table 7, non-contiguous factoid answers are not interleaved in the full-length sentence predictions as expected. The pointer-generator network is able to handle these issues. The *BLEU* and *ROGUE* scores are better on the auto-generated test data as it lacks the variation and complexity in the full-length answers compared to the manually created dataset. The validation accuracy of the 2-encoder pointer generator network as shown in the figure on the development dataset also shows significant improvement from the start of the training, with the augmented dataset providing significant increase in accuracy as shown in figure 4. The performance of our models on a KB dataset such as SimpleQA and a machine comprehension dataset such as NewsQA is shown in the table 5. As observed from the BLEU and ROGUE scores, the augmented dataset improves performance across these datasets and provide better generalization capability to the system. Some of the failure cases of the system can be observed in the table 8.

7 Conclusion

In this work, we have introduced the task of generating full-length natural answers given the question and the factoid answer. We framed the problem into an NMT task using two different approaches. Our approach uses a 2-encoder pointer generator model, where factoid answers along with the questions are inputs to the system and the full-length answers for training and is better than the baseline model for both the *BLEU* and *ROGUE* scores. Additionally, as there were no datasets which directly address this task, we released a new dataset containing tuples of *questions, factoid answers, and full-length answers* of which 300,000 samples were automatically extracted and 15000 samples were manually annotated. Our automatic dataset creation approach is scalable and can be used over any other QA datasets to retrieve more samples. We have provided the additional manually annotated clean samples to introduce complexity and variation in

Question : what kind of metal is on handful of rain? **Factoid Answer :** heavy metal **Target :** on handful of rain is heavy metal . **Modified PointerGen :** heavy metal is on handful of rain.
Question : Name an actor. **Factoid Answer :** Collien Ulmen-Fernandes **Target :** collien ulmen-fernandes is an actor. **Modified PointerGen :** collien ulmen-fernandes .
Question : Will the 10 be punished? **Factoid Answer :** no one should **Target :** no one should be punished. **Modified PointerGen :** the 10 be punished no one should punished.
Question : in which country the construction of the mosque is **Factoid Answer :** turkey **Target :** the construction of the mosque is in turkey . **Modified PointerGen :** in turkey .

Table 8: Failure Cases. Example 1 is from the Freebase dataset where the system confuses between the subject and the object. Example 2 is from Freebase not present in the training and validation data. Example 3 is from NewsQA dataset where the system fails to understand the semantics. Example 4 id from NewsQA dataset where the system fails to generate the complete full-length answer

the training data. We have performed cross-dataset evaluation by testing on a KB dataset(Freebase) and a machine comprehension dataset(NewsQA) to test the generalization capability of our system.

8 Future Work

For a deep learning model to generalize well with greater accuracy, a larger dataset comprising of a bigger vocabulary and sample size is required. Due to the limited data provided, even though our system handles tense agreement, there are instances where it fails to predict the correct tense for the verb. We plan on adding more variation to the data by annotating additional QA and machine comprehension datasets. Additionally, there is no explicit co-reference resolution module in our model. Further work needs to be done using state of the art architectures which can handle such cases and improve results. Augmenting our full-length natural answer generation system with a question answering module or a knowledge-base will provide insights into how the system performs with noisy and incorrect factoid answers. This needs to be explored further.

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Association for Computational Linguistics (ACL)*.

Yao Fu and Yansong Feng. 2018. Natural answer generation with heterogeneous memory. 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 185–195, New Orleans, Louisiana. Association for Computational Linguistics.

David Golub and Xiaodong He. 2016. Character-level question answering with attention. *ArXiv*, abs/1604.00727.

Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *ACL (1)*. The Association for Computer Linguistics.

Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. 2017a. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 199–208.

Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. 2017b. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–208, Vancouver, Canada. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2018. *Speech and Language Processing*. Draft, Stanford University.

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

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *CoRR*, abs/1902.04911.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. 2018a. Curriculum learning for natural answer generation. In *IJCAI*, pages 4223–4229. ijcai.org.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018b. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1489–1498.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, pages 2440–2448, Cambridge, MA, USA. MIT Press.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Phillip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2692–2700. Curran Associates, Inc.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2015. Neural generative question answering. *CoRR*, abs/1512.01337.