BTP- 1 & 2 Final Presentation

By: Manas Jain(170040068)

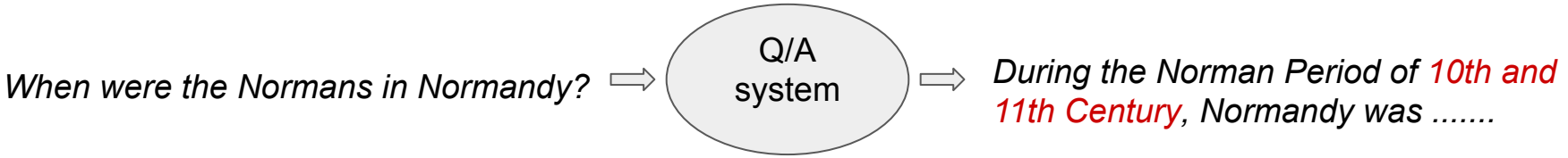
Guide: Prof. Pushpak Bhattacharyya
Prof. Jay Ghosh

Date - 19-06-2021

Natural Answer Generation : Factoid Answer to Full Length Answer

Manas Jain, Jay Ghosh, Pushpak Bhattacharyya

Motivation



- Applications like task-oriented conversational agents or chatbots often rely on QA systems to return factually correct responses to queries, but need to generate Natural Language Responses.
- Current QA systems usually return an **answer span** in the available context, or a Knowledge Base fact triplet (Subject, Predicate, Object).
- Using existing state-of-the art QA systems to generate full length natural responses is a natural extension of such systems.

Problem Statement

- Generate a response template (Natural answer) i.e generate a full length answer given a question and its factoid answer as input.
- Example :-

  Sample Input:
  - Question : *When were the normans in normandy?*
  - Factoid Answer : *10th and 11th centuries*

  Output:
  During the 10th and 11th centuries , the normans were in normandy.

  OR

  The normans were in normandy during the 10th and 11th centuries.

Solution Approaches Explored

1. Answering Naturally : Factoid to Full length Answer Generation (Pal et al 19) {Modified Pointer Generator approach}
2. Fine-tuned DialoGPT2 model
3. Rule based model based on parse tree of questions with GCM as a postprocessing step

Answering Naturally : Factoid to Full length Answer Generation (Pal et al 19)

Architecture

- Framed the problem of generating full-length answer from the question and the factoid answer into a Neural Machine Translation (NMT) task
- Model is based on the pointer-generator architecture except they used two encoders on the source side to encode question and factoid answer separately
- Form the natural full-length answer using copy words from source and generated words from vocabulary.

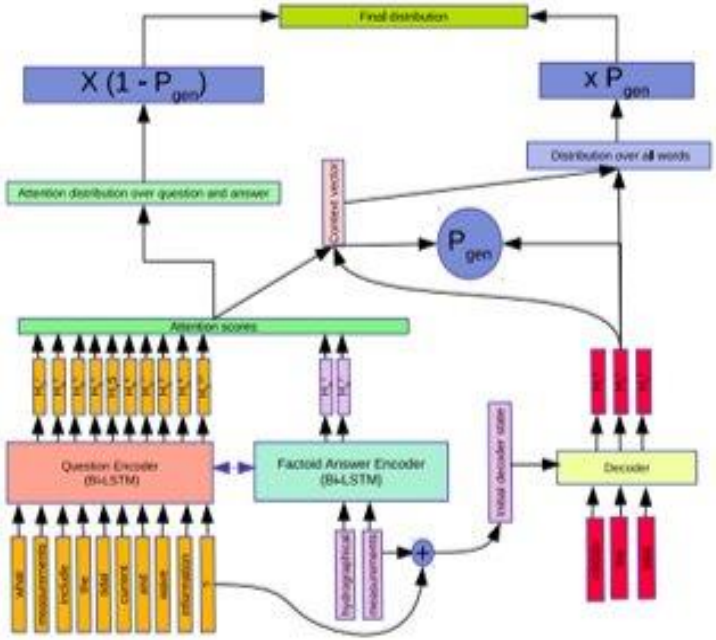


Figure 3: The 2 encoder pointer generator uses the question and factoid answer as input to generate a full-length answer in an end-to-end learning environment.

Architecture Details

- Embedding used - pre-trained GloVe (300 dimension) to initialize both encoder and decoder part
- 2 encoder - having 3 layer Bi-LSTM
- Decoder is initialized with combined final states of 2 encoder

$H_T^0 = H_Q^n + H_A^m$

$P(W_{final}) = p_{gen}P_{vocab} + (1 - p_{gen})P_{copy}$

$P_{copy}(w) = \Sigma_i a(h_T^t , h_S^i)$ $\rightarrow$ (used to predict OOV; here H_s is stack of 2 encoder hidden states) {Here a is alignment score calculated using softmax}

$P_{vocab} \rightarrow$ for $w \epsilon V$ and $w \notin \{Q \cup A\}$

Dataset details

- Training data was generated automatically from SQuAD and HarvestingQA
- The auto-dataset size is 300,000 samples.

Auto annotation technique

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 ≥ 35%. Given the question-answer pair (Q, A) and the passage P, the full-length answer T is the sentence, S, in the passage:

$$T = \text{argmax}_{S \in P} \text{BLEU}(Q, S)$$

iff $A \in S$ & $\text{BLEU}(Q, S) \geq 35\%$

- 15000 QA pairs manually annotated

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.

Model output and comparison

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 .*

Limitations/Failure cases

1. Incoherent sentence due to failure in reasoning
2. Outputs only the factoid answer
3. Outputs clausal answers
4. Failure to incorporate morphological variations

Examples:- Q :- *Is asia cruise male or female?* FA :- *female*

Target :- *asia cruise is female*

Modified PointerGen :- *female is asia cruise male or female*

Q :- *In which country construction of mosque is?* FA :- *turkey*

Target :- *the construction of mosque is in turkey*

Modified PointerGen :- *In Turkey*

Q :- *Which nationality is chin-hui tsao?* FA :- *taiwan*

Target :- *chin-hui tsao is taiwanese*

Modified PointerGen :- *chin-hui tsao is taiwan*

Some more failure examples

- QUES 80: ['what', 'kind', 'of', 'music', 'is', 'white', 'shoes', '&', 'the', 'couples', 'company', '</s>']
  ANS 80: ['jazz', '</s>']
  PRED 80: jazz is white shoes & the couples company
  GOLD 80: white shoes & the couples company is a type of jazz.
- QUES 2: ['What', 'is', 'a', 'disease', 'treated', 'by', 'virology', '?', '</s>']
  ANS 2: ['hepatitis', 'd', '</s>']
  PRED 2: What is a disease treated by virology d
  GOLD 2: hepatitis d is a disease treated by virology
- QUES 4: ['What', 'is', 'a', 'song', 'from', 'the', 'album', 'beyond', 'and', 'back', ':', 'the', 'x', 'anthology', '?', '</s>']
  ANS 4: ['los', 'angeles', '</s>']
  PRED 4: What is a song from the album beyond los angeles
  GOLD 4: los angeles is a song from the album beyond and back : the x anthology .

Fine-tuning of DialoGPT2 model

- DialoGPT (dialogue generative pre-trained transformer) is a tunable gigawordscale neural network model for generation of conversational responses, trained on Reddit data. Trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017
- Used Microsoft pre-trained DialoGPT2 small model from Hugging Face transformer library
- Ran various experiments changing data size and type of data (manual or auto). Finally fine-tuning on 13k manual data for 8 epochs gave the best results

Difference in input given vs normal input of DialoGPT

- Normally GPT models are used to make conversational chatbots where the input is the question asked and all the previous dialogs are kept as context and are passed as input to the model for training.
- Here for our task we concatenate the question with its extracted factoid answer and keep manually annotated answers as target in fine-tuning the model. For inference, question and factoid answer are concatenated and provided as input to the fine-tuned model to generate a response.

Model comparison

- Results on 420 newsQA test data

model	BLEU	ROUGE-1	ROUGE-L
DGPT finetuned on 13k manual data (8 epochs)	40.13	70.61	67.01
DGPT finetuned on 15k manual+auto (3 epochs)	33.77	59.27	53.68
Modified Pointer Generator	73.29	95.38	93.65

Model Comparison

- Results on 3200 test examples from SquAD data (manual)

model	BLEU	ROUGE-1	ROUGE-L
DGPT finetuned on 13k manual data (8 epochs)	33.23	66.51	60.35
Modified Pointer generator approach	74.05	91.24	86.25

Output examples

Q - what is going live on tuesday ?
FA - web-based on-demand television and movie service
Output - on tuesday, the web-based version of " net-based " television and film service
Target - the web-based on-demand television and movie service is going live on tuesday .

Q - what was the space station crew forced to take shelter from ?
FA - a piece of debris
Output - the space station was forced to flee from the space center and the crew was forced into a piece of material.
Target - the space station crew was forced to take shelter from a piece of debris .

Q - what is the album about ?
FA - dragging you out of the bedroom and onto the dance floor .
Output - the album about dragging you out from the bed and onto a dance floor is the " drag you out ".
Target - the album is about dragging you out of the bedroom and onto the dance floor .

Q - where did lewis partnership begin ?
FA - started as a single shop on oxford street in london , opened in 1864 by john
Output - lewis partnered with start as a sole shop onOxford street, opened early in 1866 by john.
Target - lewis partnership began as a single shop on oxford street in london , opened in 1864 .

Limitations of the finetuned DialoGPT2 model

- Problem of adding unwanted things in the final answers which doesn’t have any mention in the question and the factoid answer
- Repetition in the answer (Factoid answer not present in the final answer)
- Mismatch in the questions having some numerical data or year is mentioned
- The model has some errors copying the proper nouns as given in the questions. The final answer has that names but with changed spelling. (eg:- elizabeth -> elizabetha; alexander -> alexandrick)

Rule based approach based on question structure

- There are 2 versions which will be discussed in this section wherein the second version is an improvement over the first version.
- In the First version of our Rule based approach we just replaced the WH word present in the question with the factoid answer. In this method first we will find the position of the WH word present in the question then replace that word with given factoid answer to give a natural answer.
- Used AllenNLP to get the POS tags from constituency parse tree output of the question
- Iterated the question sentence, replaced words having POS tag as “WRB”, “WP” with the factoid answer to form the complete answer
- Question : What is the capital of India? Factoid answer : Delhi
- Rule Based Output v1 : Delhi is the capital of India

Rule based approach results

- Based on POS tags of wh questions like which,who,what etc taken from Allennlp
- Improvement in ROUGE, BLEU scores on 420 examples from Newsqa

model	BLEU	ROUGE-1	ROUGE-L
DGPT finetuned on 13k manual data (8 epochs)	40.13	70.61	67.01
Rule based approach v1	69.59	89.17	72.18
Modified Pointer Generator	73.29	95.38	93.65

Output examples

Q - what is going live on tuesday ?
FA - web-based on-demand television and movie service
Output - web-based on-demand television and movie service is going live on tuesday .
Target - the web-based on-demand television and movie service is going live on tuesday .

Q - what was the space station crew forced to take shelter from ?
FA - a piece of debris
Output - a piece of debris was the space station crew forced to take shelter from .
Target - the space station crew was forced to take shelter from a piece of debris .

Q - what is the album about ?
FA - dragging you out of the bedroom and onto the dance floor .
Output - dragging you out of the bedroom and onto the dance floor is the album about .
Target - the album is about dragging you out of the bedroom and onto the dance floor .

Q - where did lewis partnership begin ?
FA - started as a single shop on oxford street in london , opened in 1864 by john
Output - started as a single shop on oxford street in london , opened in 1864 by john did lewis partnership begin .
Target - lewis partnership began as a single shop on oxford street in london , opened in 1864 .

Update on rule based approach (AUX verb - Main verb position)

- To solve the problem of ordering of natural answer ie answer followed by question or question followed by answer. We have to look if the main verb and auxiliary verb are together then factoid answer is replaced with Wh part and if not then we have to add factoid answer in the end.
- Used dependency parse tree get AUX and VERB tag and check if they are together and added this condition to the existing rule based model
- If AUX and VERB tag are not together then we add factoid answer at the end of the question
- If question doesn’t have verb in it then we add all words after auxiliary word in the answer and add the factoid answer at end

Rule based v2 results

- ROUGE, BLEU scores on 420 examples from Newsqa

model	BLEU	ROUGE-1	ROUGE-L
DGPT finetuned on 13k manual data (8 epochs)	40.13	70.61	67.01
Rule based approach v2	63.51	90.35	83.33
Rule based approach v1	69.59	89.166	72.177
Modified Pointer Generator	73.29	95.38	93.65

Output examples

Q - what is going live on tuesday ?
FA - web-based on-demand television and movie service
Output - web-based on-demand television and movie service is going live on tuesday .
Target - the web-based on-demand television and movie service is going live on tuesday .

Q - what was the space station crew forced to take shelter from ?
FA - a piece of debris
Output - the space station crew was forced to take shelter from a piece of debris.
Target - the space station crew was forced to take shelter from a piece of debris .

Q - what is the album about ?
FA - dragging you out of the bedroom and onto the dance floor .
Output - the album about is dragging you out of the bedroom and onto the dance floor .
Target - the album is about dragging you out of the bedroom and onto the dance floor .

Q - where did lewis partnership begin ?
FA - started as a single shop on oxford street in london , opened in 1864 by john
Output - lewis partnership begin started as a single shop on oxford street in london , opened in 1864 by john.
Target - lewis partnership began as a single shop on oxford street in london , opened in 1864 .

Post Processing - Grammarly GECToR

- Used state of the art pre-trained transformer based Grammar Error Correction Model (GEC).
- We used GECToR (Omelianchuk et al. (2020)) GEC sequence tagging system that consists of three training stages: pretraining on synthetic data, fine-tuning on an errorful parallel corpus, and finally, fine-tuning on a combination of errorful and error-free parallel corpora. This model gives state of the art results on the task of Grammar Error Correction on CoNLL-2014 and BEA-2019 datasets.
- This model was available with 3 cutting edge transformer encoders namely BERT, RoBERTa and XLNET which we used as a post processing step in natural answer generation to improve our generated answers
- Input - Rule based algorithm output, Output - Grammar errors corrected answers

Qualitative Analysis (Factoid questions)

Q - in what year did san diego become part of mexico city ? FA - 1821
RB ans - san diego become part of mexico city did 1821.
GECTOR RoBERTa - San Diego became part of Mexico City in 1821.

Q - when were private prosecutors used in north carolina ?
RB ans - private prosecutors were used in north carolina 1975. FA - 1975
GECTOR RoBERTa - Private prosecutors were used in North Carolina in 1975.

Q - during what time period did the mauryan and the gupta era end ? FA - 200 bce – 500 ce
RB ans - the mauryan and the gupta era end did 200 bce – 500 ce.
GECTOR XLnet - The Mauryan and the gupta eras ended at 200 bce – 500 ce.
GECTOR RoBERTa - The Mauryan and the Gupta era ended in 200 bce – 500 ce.

Q - who did new jersey democrats promoted cleveland as ? FA - united states senate
RB ans - new jersey democrats promoted cleveland as united states senate.
GECTOR XLnet - New Jersey Democrats promoted Cleveland to the United States Senate.
GECTOR RoBERTa - New Jersey democrats promoted Cleveland to the United States senate..

Rule Based Approach for confirmatory Questions

- We found maximum confirmatory questions are of the form :- **AUX NP VP** so we developed a rule based approach using constituency parser and generated answers in the form **NP AUX VP**
- For indirect questions this approach fails as there is a phrase in the starting of the question.
- To make an indirect 'yes / no' question, we use 'if' and the word order of a normal positive sentence. Example :-
  - Direct question: Does David live in London? (Present simple with any verb except 'be')
  - Indirect question: Can you tell me if David lives in London?
  - Direct question: Did Amanda call John yesterday? (Past simple with any verb except 'be')
  - Indirect question: Can you tell me if Amanda called John yesterday?
  - Direct question: Was he late for the meeting? (Past simple with 'be')
  - Indirect question: Can you tell me if he was late for the meeting?

Qualitative Analysis (confirmation questions)

Q - Is my fridge support quick freeze feature?
RB ans - Yes, your fridge is support quick freeze feature.
RB+GECTOR - Yes, your fridge has a quick freeze feature.
GOLD - Yes, your fridge supports quick freeze feature.

Q - Is my fridge support quick freeze feature?
RB ans - No, your fridge is not support quick freeze feature.
RB+GECTOR - No, your fridge does not support the quick freeze feature.
GOLD - No, your fridge does not support quick freeze feature.

Q - Can you tell if fridge supports quick freeze feature?
RB ans - Yes, fridge does supports quick freeze feature.
RB+GECTOR - Yes, fridge does support a quick freeze feature.
GOLD - Yes, your fridge supports quick freeze feature

Q - Does the control panel has a Quick Freeze selector?
RB ans - No, the control panel does not has a Quick Freeze selector.
RB+GECTOR - No, the control panel does not have a Quick Freeze selector.
GOLD - The fridge has no Quick Freeze selector button.

Dataset creation

- By qualitative analysis of dataset we found various mistakes in existing dataset created in form of punctuation and not well written GOLD target.
- Also NLG output can have multiple outputs. Eg :-
  - Who is the ceo of google ? FA - sundar pichai
  - [Sundar pichai is the ceo of google., **the ceo of google is sundar pichai.]**
- So we improved the dataset by adding another GOLD answer wherever necessary and remove examples with meaningless question or factoid answer
- We created a high quality dataset - Factoid questions - 380 (NewsQA), 6768(SQuAD) ; confirmatory questions - 166 (with different types of indirect questions)

NewsQA (380)

model	BLEU	ROUGE-1	ROUGE--2	ROUGE-L
RB	79.1	96.1	85.5	93.1
RB + bert	77.6	94.4	85.4	92.4
RB + roberta	81.7	95.7	88.2	93.6
RB + xlnet	80.3	94.8	87.0	92.9
MPG	84.9	95.7	89.4	93.9

SQuAD dataset - (6768)

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
MPG	75.8	94.4	87.4	91.6
RBV2	74.8	95.3	83.1	90.3
RBV2+BERT	71.5	93.9	82.4	89.5
RBV2+RoBERTa	72.1	94.0	83.1	89.8
RBV2+XLNET	71.2	93.6	82.3	89.4

Confirmatory questions dataset (166)

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
RB	70.2	87.3	75	84.8
RB+BERT	62.7	85.5	71.6	83.4
RB+RoBERTa	66.6	84.5	73	84.2
RB+XLNET	67.5	86.6	74.0	84.6

Qualitative comparison of 3 approaches

Ques - where was the bus going ? **Factoid Answer** - phoenix , arizona

MPG - the bus going was at phoenix , arizona .

RBV2 [ours] - the bus was going phoenix , arizona .

RBV2+RoBERTa [ours] - The bus was going **to** Phoenix , Arizona .

Fine-tuned DialoGPT [ours] - the bus was going to phoenix, anrizona .

Limitations

- This approach just works on copy pasting things and so if factoid answer is not a fact or is a clausal answer and this approach will fail. For eg last example of previous slide output answer had both began and started in it which is not right
- Since the approach works on the question structure so if question is not properly well formed or incomplete then the answers will not be correct

  Question : where did lewis partnership begin? **Factoid answer** : started as a single shop on oxford street in london, opened in 1864 by john.
  RBV2 + GCM output : lewis partnership begin started as a single shop on oxford street in london, opened in 1864 by john.
  Target answer : lewis partnership begin started as a single shop on oxford street in london, opened in 1864 by john

Summary

- Explored hybrid neural approaches using abstractive & extractive techniques simultaneously and rule based systems using constituency and dependency parse of the question and DialoGPT model to solve task of generating full length answer from factoid answer.
- Evaluated the models based on BLEU, ROUGE scores and manual check on some examples
- Compared the explored model output with each other and analysed merits and demerits of the approaches implemented
- Submitted a short paper - **Natural Answer Generation: From factoid answer to full-length answer using Grammar Correction** in **CIKM-2021**

Conclusion

- Worked on the task of generating full-length natural answers given the question and the factoid answer.
- Explored neural as well as rule based approaches on a new dataset containing tuples of questions, factoid answers, and full-length answers.
- Using a pointer-generator Seq2Seq model leads to reasonable generation of full length natural answer but the sentence structure is not proper, DailoGPT model solves the fluency part but copy mechanism doesn't work well in this approach.
- Proposed a Rule based approach approach using syntactic parser to solve the task of natural answer generation. Using GCM on top of our rule based approach improves the quality of the answers and comes very close to the state of the art supervised approaches. **RB + RoBERTa works best for our task**

Future work

- Rule based approach to be made more accurate and can handle all corner cases
- Evaluation metric used was not appropriate for our task so explore some more evaluation metrics
- Incorporation of world knowledge in the form of coreference resolution, syntactic dependency or other linguistic modules or meta-data, can be Explored and used to improve the existing approaches.
- Think upon some hybrid model using both the modified pointer generator and rule based approach.

THANK YOU

Appendix

Indirect questions (lite phrase before the question)

- Indirect questions start with a phrase like 'could you tell me...' or 'do you know...' etc. Used when people ask for any help from somebody in a polite way
  - Direct question: Where is the bank?
  - Indirect question: Could you tell me where the bank is?
- To make an indirect 'yes / no' question, we use 'if' and the word order of a normal positive sentence. Example :-
  - Direct question: Does David live in London? (Present simple with any verb except 'be')
  - Indirect question: Can you tell me if David lives in London?
  - Direct question: Did Amanda call John yesterday? (Past simple with any verb except 'be')
  - Indirect question: Can you tell me if Amanda called John yesterday?
  - Direct question: Was he late for the meeting? (Past simple with 'be')
  - Indirect question: Can you tell me if he was late for the meeting? (Ref :- https://www.perfect-english-grammar.com/indirect-questions.html)

User Interface

Question Answering Testing App

Please enter your question here

What is the sum of 3 and 4?

Please Enter Category of the Question

Factoid

Please enter input answer here

7

Response:

Output Answer = the sum of 3 and 4 is 7.

Related Work

Transformer based Natural Language Generation for Question Answering (Akermi et. al. INLG 2020) {Approach}

Question: When did princess Diana die?

1. Question parsing and answer extraction using the system proposed in Rojas Barahona et al. (2019): short answer = {August 31, 1997}
2. Chunking the question into text fragments using the UDPipe based dependency analysis: Q={When, did die, princess Diana}
3. Removing question marker fragment (when) and updating the verb tense and form using a rule-based approach that we have defined: Q={died, princess Diana}
4. Adding the short answer: Q={died; princess Diana; August 31, 1997}
5. Generating the set of possible answer structures S: S={died princess Diana August 31, 1997; . August 31, 1997 died princess Diana; . princess Diana died August 31, 1997; . . . . }
6. Evaluating the different answer structures using a LM: princess Diana died August 31, 1997
7. Generating possible missing word for structure∗ with BERT: Princess Diana died [missing word] August 31, 1997 (missing word = on) Answer: Princess Diana died on August 31, 1997.

Transformer based Natural Language Generation for Question Answering (link) {INLG2020}

1. This paper builds upon the fact that we have already extracted the short answer to a given question and assumes that a user-friendly answer can consist in rephrasing the question words along with the short answer. This approach is composed of two fundamental phases: The dependency analysis of the input question and the answer generation using Transformer models.
2. Using the text fragments set Q, we proceed with a permutation based generation of all possible answer structures that can form the sentence answering the question asked.
3. These structures will be evaluated by a Language Model (LM) based on Transformer models which will extract the most probable sequence of text fragments that can account for the answer to be sent to the user

Transformer based Natural Language Generation for Question Answering (link) {Approach}

- Once the best structure is identified, we initiate the generation process of possible missing words
- In this paper, authors experiment the assumption that one word could be missing and that it is located before the short answer within the identified structure, as it could be the case for a missing article (the, a, etc.) or a preposition (in, at, etc.) i.e missing word position is fixed
- To predict this missing word, we use BERT as the generation model (GM) for its ability to capture bidirectionally the context of a given word within a sentence. In case when BERT returns a non-alphabetic character sequence, we assume that the optimal structure, as predicted by the LM, does not need to be completed by an additional word