FALL 2020 CS 395T

TEXAS
The University of Texas at Austin

Question Generation

CS 395T: Topics in Natural Language Processing

Rohan Nair & Rishab Goel, The University of Texas at Austin

Question Generation

- The goal of Question Generation is to generate a valid and fluent question according to a given passage and the target answer
  - Multiple valid target answers for a passage
  - Multiple valid questions for a passage
- Question Generation can be used in many scenarios, such as automatic tutoring systems, improving the performance of Question Answering models and enabling chatbots to lead a conversation.

Question Generation Strategies

- Heuristically generate questions with rules
  - Select target answer
  - Generate questions using wh* words
  - Apply syntactic transformation + type theory to get question

- Seq2Seq models
  - Select target answer + context
  - Conditionally generate question based on target answer and context

Seq2Seq Question Generation

- Du et al. 2017
  - Use SQuAD as training set
  - Use bidirectional LSTM with attention as model
  - Show effectiveness of using Seq2Seq architecture to generate questions
    - Much higher automatic evaluation metric performance over heuristic question generation models

2 Case Studies Utilizing QG

- Unsupervised Question Answering by Cloze Translation

- Asking and Answering Questions to Evaluate the Factual Consistency of Summaries

TEXAS
The University of Texas at Austin

Unsupervised Question Answering by Cloze Translation

Patrick Lewis, Ludovic Denoyer, Sebastian Riedel (ACL 2019)

CS 395T: Topics in Natural Language Processing

Rishab Goel, The University of Texas at Austin

Cloze Intro

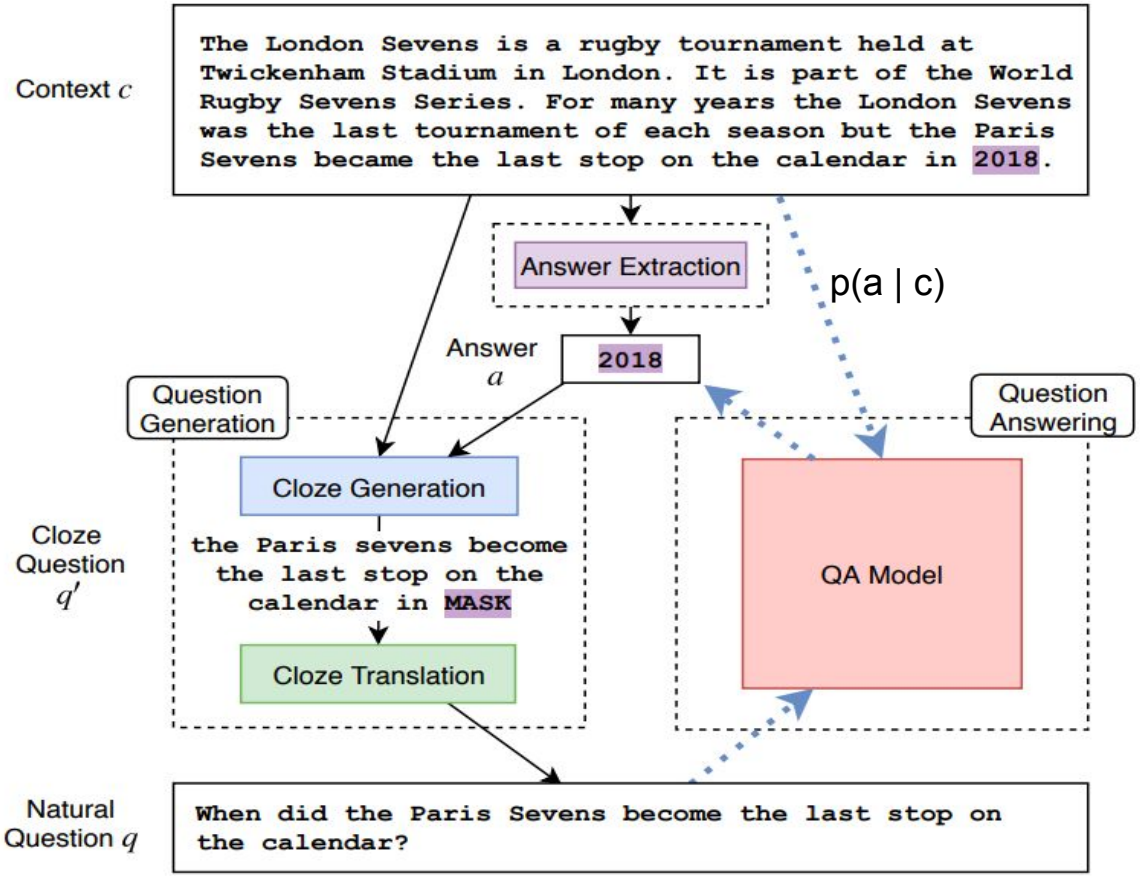
- Cloze task refers to language where some words are removed from speech and participant is asked to fill in the blank
- EX: Today, I went to the _____ and bought some milk and eggs
- This can be formulated as a question for QA by requiring the answer to the blank: Where did they go to buy milk and eggs?
- Previous analysis like LAMA (Petroni, et al.) found that LMs like BERT contain enough knowledge on their own to answer cloze questions by being able to guess masked answers

Cloze Question Answering

- Attempts to solve the problem of generating questions for training question answering models - both completely unsupervised and partially supervised
- Focus is extractive question answering - where you are given a passage of text and answer is within the text
- The method consists of three main steps -
  - identifying text that contains an answer
  - identify candidate answers and generate fill in the blank cloze questions
  - translating the question to a natural type of question using a Seq2Seq model

Generation Steps

- The goal is to create p(q, a, c) which is a generator that generates question q after generating a the answer based on context c
- p(q, a, c) = p(c)p(a|c)p(q|a, c) is the formula used to determine the generator - first context (p(c)) is found and then an answer based on the context is found (p(a|c)) and then a question is generated off of this - p(q|a, c)
- The challenge of the paper is how to generate all three of these components



p(c)

p(a | c)

p(q|a, c)

Figure 1 from paper

Context/Answer and Cloze Generation

- Given a corpus c, they use two different components in order to generate answers from the context - from a paragraph they used noun phrases and named entity recognition in order to identify answers
- Once an answer has been defined - the answer within the context is masked and then the subclause or sentence around the blank is regarded as a cloze question
- Now the task is to translate a cloze question to a natural question

Cloze Translation

- Given a cloze question, the paper has four methods in order to translate to a natural question
- Prevalent among these methods is the use of a wh* word (who, what, when, where, why) in order to form the natural question and in order to select a word a heuristic is used where the answer is categorized and used to determine a word
- Identity mapping where the answer is replaced with a wh* word, noisy cloze where a wh* word is prepended and then the sentence is perturbed, rule based where a syntactic transformation is used, and Seq2Seq

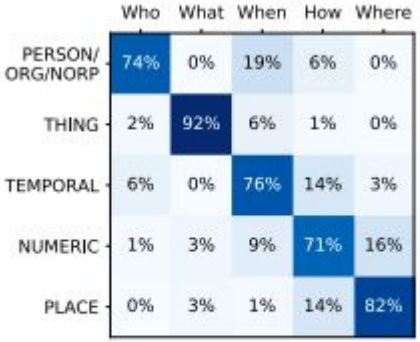


Figure 4: Wh* words generated by the UNMT model for cloze questions with different answer types.

#	Cloze Question	Answer	Generated Question
1	they joined with PERSON/NORP/ORG to defeat him	Rom	Who did they join with to defeat him?
2	the NUMERIC on Orchard Street remained open until 2009	second	How much longer did Orchard Street remain open until 2009?
3	making it the third largest football ground in PLACE	Portugal	Where is it making the third football ground?
4	he speaks THING, English, and German	Spanish	What are we , English , and German?
5	Arriving in the colony early in TEMPORAL	1883	When are you in the colony early?
6	The average household size was NUMERIC	2.30	How much does a Environmental Engineering Technician II in Suffolk , CA make?
7	WALA would be sold to the Des Moines-based PERSON/NORP/ORG for $86 million	Meredith Corp	Who would buy the WALA Des Moines-based for $86 million?

Experiments

- There are two methods in order to generate QA models
  - Train or finetune LM on data (BERT, BiDAF)
  - Use the posterior of the model in order to calculate p(a|c, q)
- The two methods above are evaluated via Exact Match and F1 on the SQuAD dataset after training on the generated data
- The best approach attains 54.7 F1 on the SQuAD test set and 44.2 on EM - the ensemble achieves better results than the single result alone

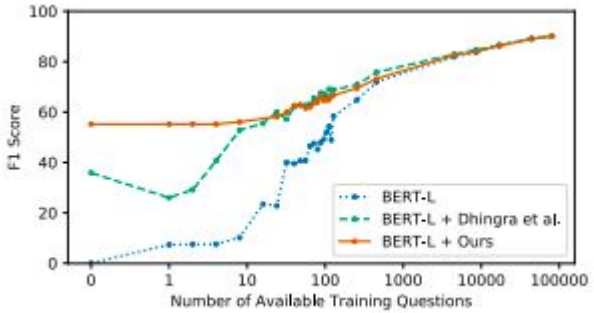
Unsupervised Models	EM	F1
BERT-Large Unsup. QA (ens.)	**47.3**	**56.4**
BERT-Large Unsup. QA (single)	44.2	54.7
BiDAF+SA (Dhingra et al., 2018)	3.2†	6.8†
BiDAF+SA (Dhingra et al., 2018)‡	10.0*	15.0*
BERT-Large (Dhingra et al., 2018)‡	28.4*	35.8*
Baselines	**EM**	**F1**
Sliding window (Rajpurkar et al., 2016)	**13.0**	**20.0**
Context-only (Kaushik and Lipton, 2018)	10.9	14.8
Random (Rajpurkar et al., 2016)	1.3	4.3
Fully Supervised Models	**EM**	**F1**
BERT-Large (Devlin et al., 2018)	**84.1**	**90.9**
BiDAF+SA (Clark and Gardner, 2017)	72.1	81.1
Log. Reg. + FE (Rajpurkar et al., 2016)	40.4	51.0

Ablation Studies

- Training on data far outperforms trying to use the posterior method - due to linguistic pretraining and BERT outperforms BIDAF
- NER instead of noun phrases improves the F1 score on BERT around 9 points
- Subclauses instead of whole sentences for the cloze translation is also better for improving the F1 score
- Shorter questions perform better on SQuAD - in this way adding noise to perform the cloze translation also improves F1 Score

Error Analysis

- The BERT model is shown to be capable of performing even when the answer isn't recognized by NER - showing it can generalize the task itself
- Without the WH* heuristic, the unsupervised NMT cloze translation model is still capable of generating the word some time although it struggles with certain categories
- They also show that the F1 score gets progressively better the more training examples are provided



Related Work

- Unsupervised NMT (Conneau et al., 2017; Lample et al., 2017, 2018; Artetxe et al., 2018) used to do QA tasks
- Question generation tried previously with
  - symbolic approaches
  - pipelines of templates and syntax rules
  - neural models that use SQuAD to generate more questions
- Use semi-supervised generation to improve model accuracy (Yang et al. (2017)
- QA datasets used to create inference datasets (Demszky et al.)

Conclusion and Discussion

- They find that overall they can surpass simple supervised models as shown in the data table and most unsupervised models for QA on the SQuAD dataset
- However the questions in the SQuAD dataset are relatively simple but it is impressive to do this unsupervised
- They may rely too much on heuristics in order to create questions
  - Reliant on NER
  - Reliant on WH* heuristic
  - Use Pretrained BERT

Follow Up Papers

- (Fabbri et al) also generates questions using context and sees improvement on SQuAD over this paper
- (Li et al) also improves on question generation
  - Retrieves QA pairs similar to this
  - Uses LMs to refine answers
  - Improves quality of dataset

Discussion Questions

- Do they rely too much on resources/heuristics like Named Entity Recognition in order to form questions - what else can they use instead?
- Should they implement more complex questions? In this one there is no “multi-hop” required - they just translate simple fill in the blanks. What is a mechanism they can use to do this?
- Are there any other approaches to creating various modules for cloze translation that you would’ve liked to see (such as LMs other than Seq2Seq)?
- Should they have used any other dataset to evaluate the final models other than SQuAD?

FALL 2020 CS 395T

TEXAS
The University of Texas at Austin

Asking and Answering Questions to Evaluate the Factual Consistency of Summaries

Alex Wang, Kyunghyun Cho, Mike Lewis (ACL 2020)

CS 395T: Topics in Natural Language Processing

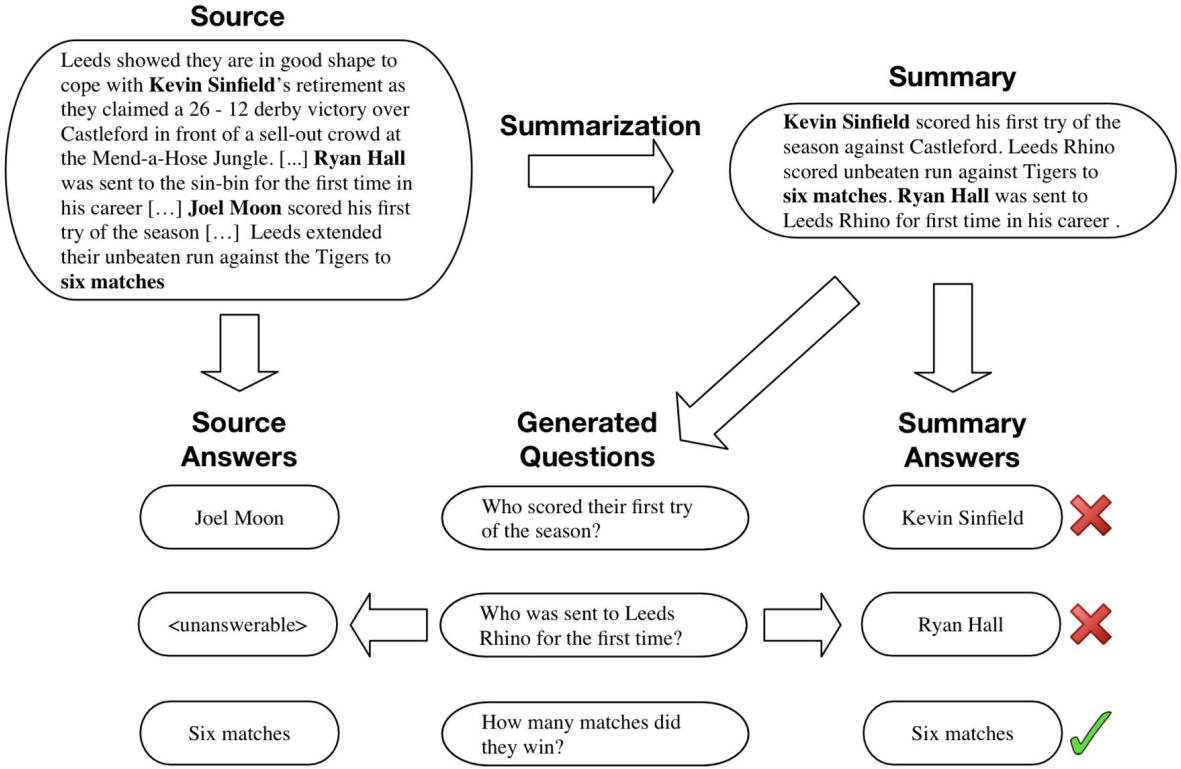
Rohan Nair, The University of Texas at Austin

QAGS Overview

- Goal: Can we utilize strong QA models for fact checking?
- QAGS:
  - Output summary Y for document X
  - Generate question Q from Y using $P(Q|Y)$
  - Use QA model to get answer A distribution from Q,Y and Q, X
    - Two distributions: $P(A|X,Q)$, $P(A|Y,Q)$
  - Use D to measure divergence between $P(A|X,Q)$, $P(A|Y,Q)$

- QAGS = $E_{Q\sim P(Q|Y)}[D(P(A|Q,X), P(A|Q,Y))]$

QAGS Overview



Related Work: Fact Checking Summaries

- NLI
  - Logical consistency between two statements
    - Sentence level consistency
  - MNLI (Williams et al. 2018) & SNLI (Bowman et al. 2015) common datasets
  - Pretrain model on NLI task, apply model to downstream fact checking task
    - Sentence level entailment checking (Kryściński et al. 2019)
    - Ranking factual summaries task (Falke et al. 2019)
- FEVER (Thorne et al. 2018)
  - Verification against textual sources
  - Claims classified as Supported, Refuted, Not Enough Info

Related Work: Fact Checking Summaries

- FactCC (Kryściński et al. 2019)
  - Train FactCC model on weakly supervised task
    - Heuristically create factual inconsistencies to train on
  - FactCC outperforms MNLI/FEVER based classifiers in manually annotated test dataset

Implementation Details

- Modeling P(Q|Y)
  - Seq2Seq Question Generator model is BART
  - Input is context + target answer
    - NER and noun phrases used to select target answers
  - Heuristically filter out poor questions (length < 3, duplicates, etc.)
- The QA Model
  - Use extractive QA
    - Limitation over abstractive QA, which could find paraphrases of similar answers
  - Use BERT variants as the QA model

Implementation Details

- Evaluate outputs from 2 abstractive summarization models
  - Bottom up Summarization (Gehrmann et al. 2018) trained on CNN/DM
  - BART trained on XSUM
- Scoring function D
  - Use token-level F1 as scoring function

QAGS Experiment

- Goal: Measure how well the QAGS metric matches up with human evaluation
- Create a human annotated score for each abstractive summary
- Baseline “Factuality” Metrics:
  - ROUGE (Recall), BLEU (Precision), BERTScore, METEOR
- Proposed Factuality Metric:
  - QAGS
- Measure Pearson correlation between human factuality scores and metrics

Annotation Process

- Goal: Determine is summary is factually consistent with source document
- Annotation Process:
  - Annotators given one summary sentence at a time + source document
    - determine if sentence is factually consistent (binary label)
  - Each sentence annotated 3 times, majority label is true label
    - Krippendorff's alpha of .51/.34 for CNN+DM/XSUM
  - Factuality score of summary is average factuality of sentences

Experimental Results

- QAGS characterizes factuality better than other common summarization metrics
- Increasing the number of questions leads to higher QAGS correlation with human factuality annotations

Metric	CNN/DM	XSUM
ROUGE-1	28.74	13.22
ROUGE-2	17.72	8.95
ROUGE-L	24.09	8.86
METEOR	26.65	10.03
BLEU-1	29.68	11.76
BLEU-2	25.65	11.68
BLEU-3	23.96	8.41
BLEU-4	21.45	5.64
BERTScore	27.63	2.51
QAGS	**54.53**	**17.49**

# Questions	CNN/DM	XSUM
5	41.61	15.63
10	41.17	15.49
20	54.53	17.49
50	57.94	17.74

Results

QA model	SQuAD (F1)	CNN/DM (Pear.)	XSUM (Pear.)
`bert-base`	75.95	55.20	20.71
`bert-large`	81.57	54.53	17.49
`bert-large-wwm`	84.36	51.36	18.07

- Higher quality QA models do not necessarily lead to a better QAGS correlation with human annotation
  - Weaker QA does better for CNN/DM, no clear trend for XSUM
- Lower perplexity leads to highers QAGS correlation with human factuality annotation for CNN/DM, but not for XSUM
- QAGS achieves SOTA on summary ranking task (Falke et al. 2019) too
  - Outperforms models using NLI pretrained tasks, without needing an NLI dataset

NewsQA (ppl.)	CNN/DM (Pear.)	XSUM (Pear.)
5.48	54.53	17.49
9.50	50.09	19.93
18.56	47.92	16.38

Model/Metric	% Correct (↑)
Random	50.0%
BERT NLI	64.1%
ESIM	67.6%
FactCC	70.0%
QAGS	**72.1%**

Qualitative Analysis

- Manually inspect 400 samples
  - Look at Question Generated, Predicted Answer, and Answer Similarity
  - Generally high quality questions (understandable and on topic)
    - 8.75% Nonsensical, 3.00% Unanswerable
  - Inspection of predicted answer from answerable (well formed) questions incorrectly answered
    - 1.75% incorrect answers from summaries
    - 32.5% incorrect answers from documents
      - QA for long documents seems to be lacking
  - F1 Scoring function generally seems to hold up
    - 8.00% answer is correct in both summary and document QA but F1 score marks it as incorrect

QAGS Success Case

Article: On Friday, 28-year-old Usman Khan stabbed reportedly several people at Fishmongers' Hall in London with a large knife, then fled up London Bridge. Members of the public confronted him; one man sprayed Khan with a fire extinguisher, others struck him with their fists and took his knife, and another, a Polish chef named ukasz, harried him with a five-foot narwhal tusk. [. . .]
Summary : On Friday afternoon , a man named Faisal Khan entered a Cambridge University building and started attacking people with a knife and a fire extinguisher .
Question 1: What did the attacker have ?
Article answer: a large knife **Summary answer:** a knife and a fire extinguisher
Question 2: When did the attack take place ?
Article answer: Friday **Summary answer:** Friday afternoon
Question 3: What is the attacker's name ?
Article answer: Usman Khan **Summary answer:** Faisal Khan
Question 4: Where did the attack take place ?
Article answer: Fishmongers' Hall **Summary answer:** Cambridge University building

QAGS Failure Case

Article: In findings published on Wednesday in the journal PLOS ONE, an international team of scientists report ancient Egyptians captured sacred ibises (Threskiornis aethiopicus) from the wild for use in ritual sacrifice rather than domesticating the birds. [...] The team collected DNA samples from mummified birds collected from six separate catacombs including sites at Abydos, Saqqara, and Tuna el-Gebel with permission from the Egyptian Ministry of State for Antiquity, and several museums offered to send tissue samples from the mummified ibises in their collections. [...]
Summary : Archaeologists have used DNA samples from ancient ibis birds to determine whether the birds were domesticated or sacrificed in ancient Egypt
Question 1: Archaeologists have used what to determine whether the birds were domesticated ?
Article Answer: hatchery structures **Summary Answer**: DNA samples
Question 2: Who used DNA samples to determine whether the birds were domesticated ?
Article Answer: [NO ANSWER] **Summary Answer:** Archaeologists
Question 3: What are archeologists using to determine whether the birds were domesticated ?
Article Answer: DNA samples **Summary Answer:** DNA samples
Question 4: Where were the birds found?
Article Answer: six separate catacombs **Summary Answer:** ancient Egypt

Table 6: Example questions and answers generated when computing QAGS. The questions are overwhelmingly fluent and relevant. The answers indicate which tokens in the summary are factually consistent or inconsistent.

Highly Related Work

- FEQA (Durmus et al. 2020)
  - Concurrently developed with QAGS
  - Propose abstractiveness and faithfulness scoring functions
    - Increasing abstractiveness leads to decreased faithfulness
    - Abstractiveness scoring function defined heuristically
    - Faithfulness scoring function uses summary for question generation and QA over source
  - Analyze correlation between abstractiveness and faithfulness scores to human evaluation metrics

FEQA Overview

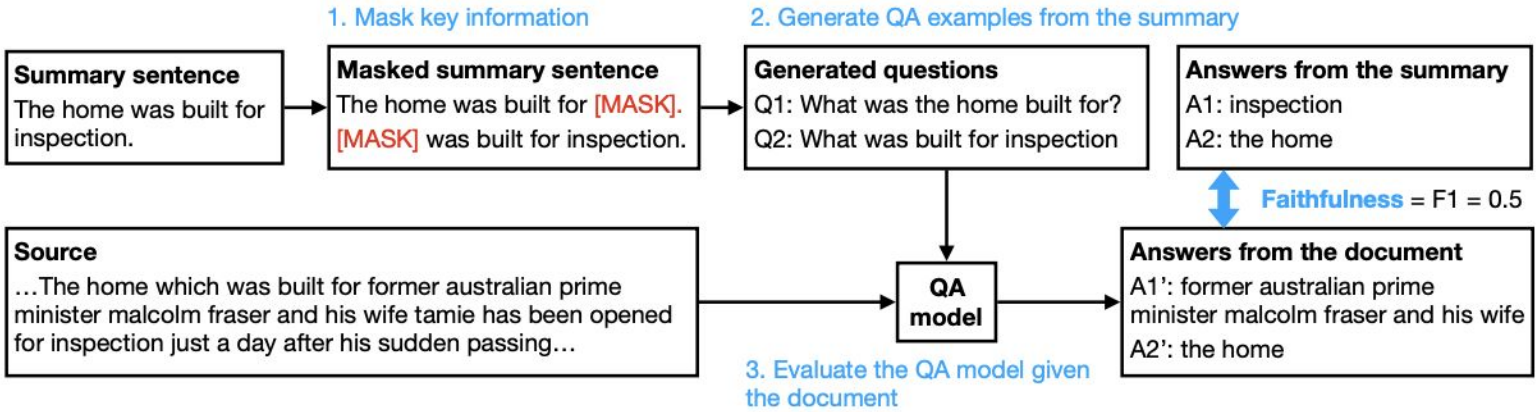


Figure 2: Overview of FEQA. Given a summary sentence and its corresponding source document, we first mask important text spans (e.g. noun phrases, entities) in the summary. Then, we consider each span as the "gold" answer and generate its corresponding question using a learned model. Lastly, a QA model finds answers to these questions in the documents; its performance (e.g. F1 score) against the "gold" answers from the summary is taken as the faithfulness score.

FEQA vs QAGS

- Differences between FEQA and QAGS
  - Generate target answers using a masking process
    - Both still use BART
  - No QA over summary
    - Compare masked out words to QA over document

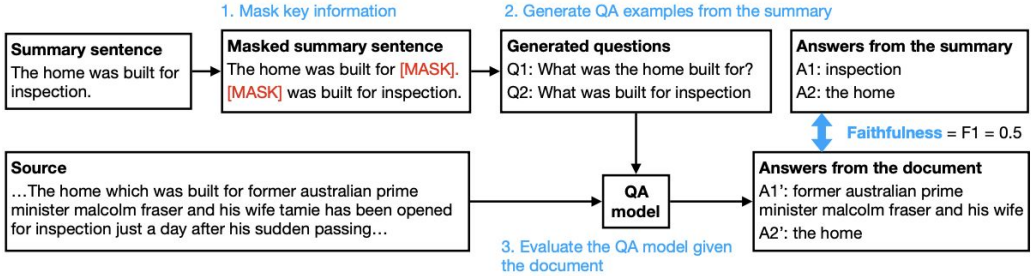
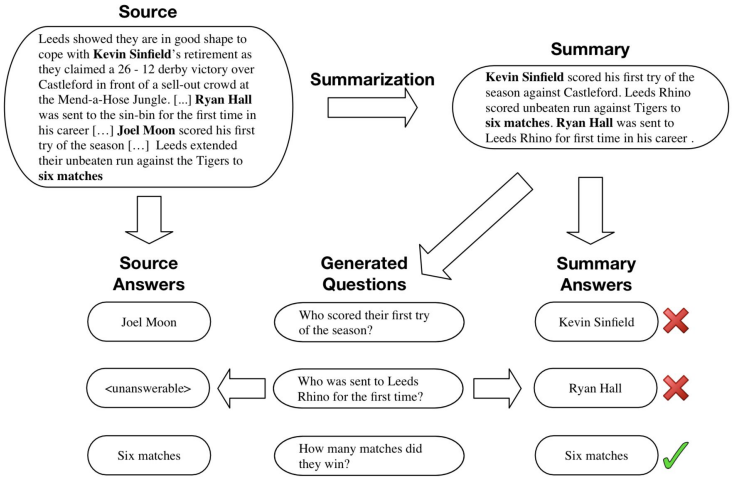


Figure 2: Overview of FEQA. Given a summary sentence and its corresponding source document, we first mask important text spans (e.g. noun phrases, entities) in the summary. Then, we consider each span as the "gold" answer and generate its corresponding question using a learned model. Lastly, a QA model finds answers to these questions in the documents; its performance (e.g. F1 score) against the "gold" answers from the summary is taken as the faithfulness score.



Summary: Metrics

- Fabbri et al. 2020:
  - Evaluate 12 summarization metrics
  - Benchmark 23 summarization models
  - Assemble abstractive summarization dataset
  - Release human judgements of summaries
- Contribution: Assembled toolkit of summarization metrics

Metric	Coherence	Consistency	Fluency	Relevance
ROUGE-1	0.2011	0.1811	0.1496	**0.3565**
ROUGE-2	0.1528	0.1583	0.0996	0.2685
ROUGE-3	0.1635	0.1587	0.0907	0.2611
ROUGE-4	0.1516	0.1522	0.0942	0.2313
ROUGE-L	0.1564	0.1578	0.1382	0.3347
ROUGE-su*	0.1897	0.1678	0.1360	0.3291
ROUGE-w	0.1525	0.1648	0.1209	0.3283
ROUGE-we-1	0.2020	0.1832	0.1513	**0.3546**
ROUGE-we-2	0.1525	0.1319	0.0882	0.2895
ROUGE-we-3	0.1270	0.1053	0.0567	0.2634
S^3-pyr	0.1667	0.1624	0.0813	**0.3469**
S^3-resp	0.1616	0.1609	0.0822	0.3227
BertScore-p	0.1449	0.1500	0.2056	0.1959
BertScore-r	0.1737	0.2082	0.1662	**0.3503**
BertScore-f	0.1854	0.2030	0.2162	0.3192
MoverScore	0.2115	0.1899	0.2005	0.3114
SMS	0.1797	0.1794	0.1701	0.2750
SummaQA^	0.0835	0.0802	-0.0298	0.2626
BLEU	**0.2212**	0.1750	0.1374	**0.3561**
CHRF	0.2009	0.2110	0.1716	0.2593
CIDEr	0.1586	0.1832	0.1311	0.3237
METEOR	0.0290	0.0336	0.0714	-0.0055
Length^	0.1623	0.1655	0.1036	0.3310
Novel unigram^	0.1108	**-0.3195**	**-0.2238**	-0.1043
Novel bi-gram^	0.0030	**-0.4417**	**-0.3231**	-0.1701
Novel tri-gram^	-0.0655	**-0.4660**	**-0.3499**	-0.1959
Repeated unigram^	**-0.2445**	-0.1309	-0.2130	-0.0396
Repeated bi-gram^	**-0.3205**	-0.1539	-0.2261	-0.1733
Repeated tri-gram^	**-0.2475**	-0.0801	-0.1619	-0.1264
Stats-coverage^	0.0402	**0.4613**	**0.3420**	0.2026
Stats-compression^	0.0506	-0.0604	0.0236	-0.2020
Stats-density^	**0.2775**	**0.2941**	**0.2488**	0.2596

Table 2: Pearson correlation coefficients of expert annotations along four quality dimensions with automatic metrics using 11 reference summaries per example. ^ denotes metrics which use the source document. The five most-correlated metrics in each column are bolded.

Discussion of Limitations

- QAGS only useful with other metrics
  - No measure of readability, variability, etc.
- Only pertains to abstractive summarization
- Standardized QA/QG models needed to standardize metric
  - May struggle with domain shifts
  - Requires good QA/QG models within target domain
- Payments to annotators are per summary
  - Impact quality of human annotations?