Question Generation

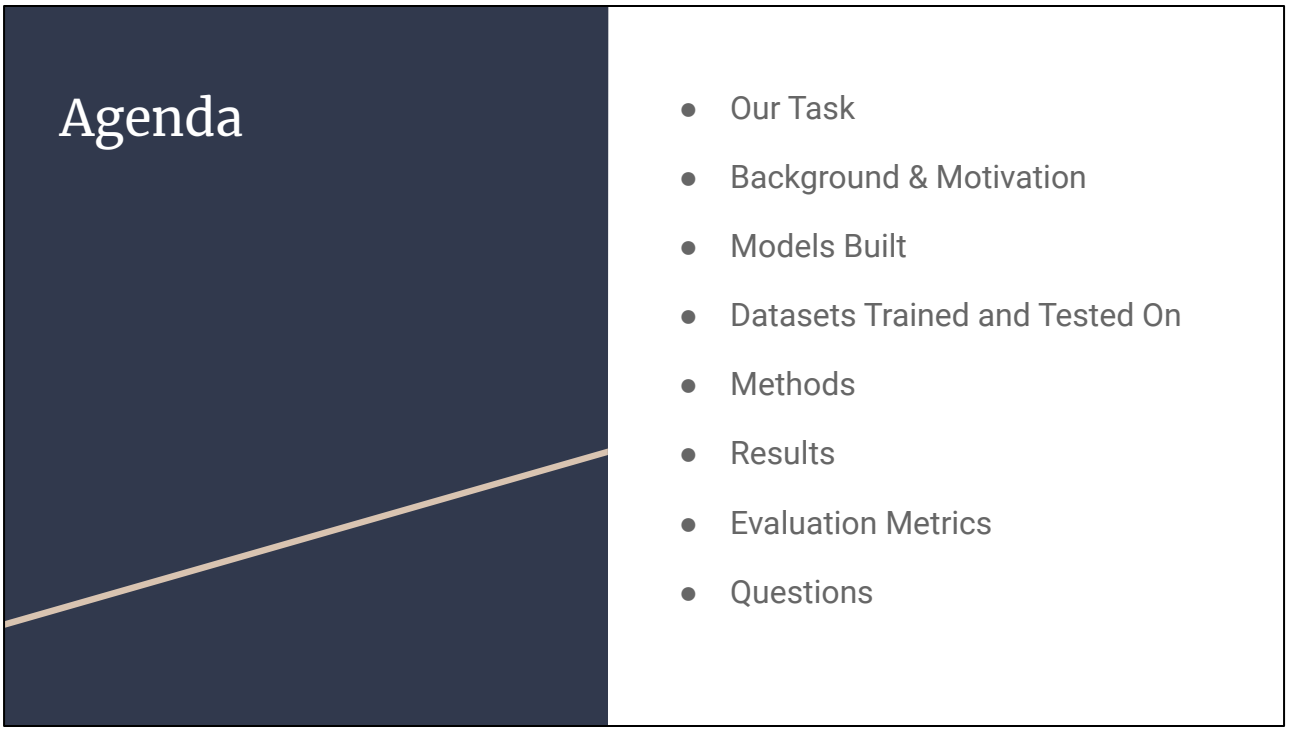
With Different Datasets, Training Strategies, and Evaluation Metrics

Bronte Baer, Jean-Luc Jackson, Richard Robbins

BRONTE

Our group is myself, Jean-Luc, and Richard. And we did Question Generation with different datasets, training strategies, and evaluation metrics.

DUE: DECEMBER 5, 2022
DESCRIPTION LINK



BRONTE

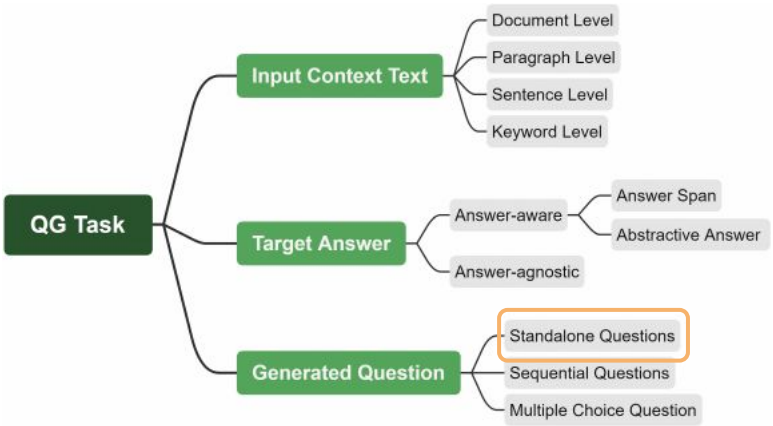
Here’s the agenda of what we’ll cover in our presentation today.

Question Generation

BRONTE

We were interested in text generation tasks, and we wanted to choose something that hasn't been commonly studied in the NLP field. The challenge of question answering was appealing, so we decided to flip the task on its head and do question generation.

Background & Motivation

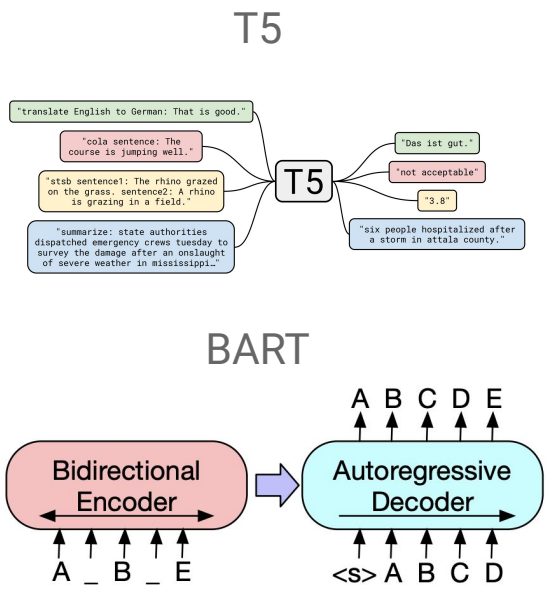


BRONTE

While the task is somewhat novel, its use isn't completely unprecedented. There have been others who have worked on solving the problem of question generation. Typically, the motivation for their work has been for education purposes, such as generating questions for reading comprehension materials. Question generation can also be used for chatbot integrations, such as starting a conversation or requesting feedback.

Image source: https://dl.acm.org/doi/fullHtml/10.1145/3468889

Models



BRONTE

After researching the question generation task, we decided to build and compare two model architectures: T5 and BART.

We chose T5 and BART for our models because each is an encoder-decoder architecture well-suited to sequence-to-sequence tasks like question generation.

T5 image source:
https://huggingface.co/mrm8488/t5-base-finetuned-question-generation-ap

BART image source: https://paperswithcode.com/method/bart

Datasets

- NQ
- QuAC
- SQuAD
- TriviaQA
- Blended

	NQ	QuAC	SQuAD	TriviaQA
training	74,218	69,109	87,599	77,085
validation	2,356	5,868	10,570	9,835

RICH

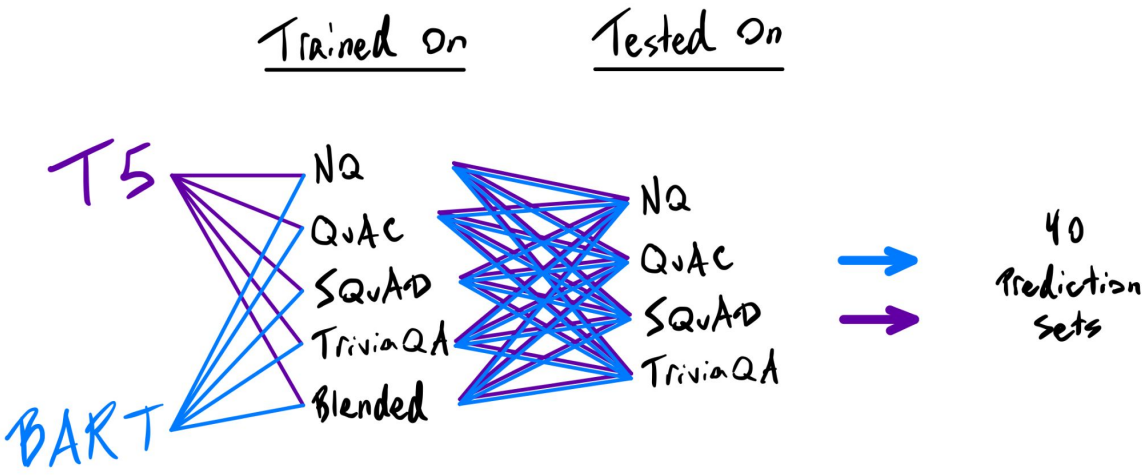
To train and test our models, we used four datasets that were created for question answering.

We did not find any datasets created specifically for question generation.

When it comes to question answering and question generation, people typically start with SQuAD. We did so too. The others are also well known and have different signature characteristics.

We wanted to see what would happen if we blended our four data sets into a single dataset, so when we talk about our blended dataset, that's what we mean.

Methods



RICH

Our baseline model is T5 fine-tuned for question generation using SQuAD.

We built three more models by training on each of our other source datasets.

We built a fifth model used our blended dataset, which we also randomly shuffled.

We repeated all of this by replacing T5 with BART.

We then tested each of the ten models against the validation data from the four source datasets yielding forty prediction sets that formed the basis for our report.

While we aren't going to dwell on it here, we also built and tested some models on a blended dataset that wasn't shuffled. (They under-performed)

Results

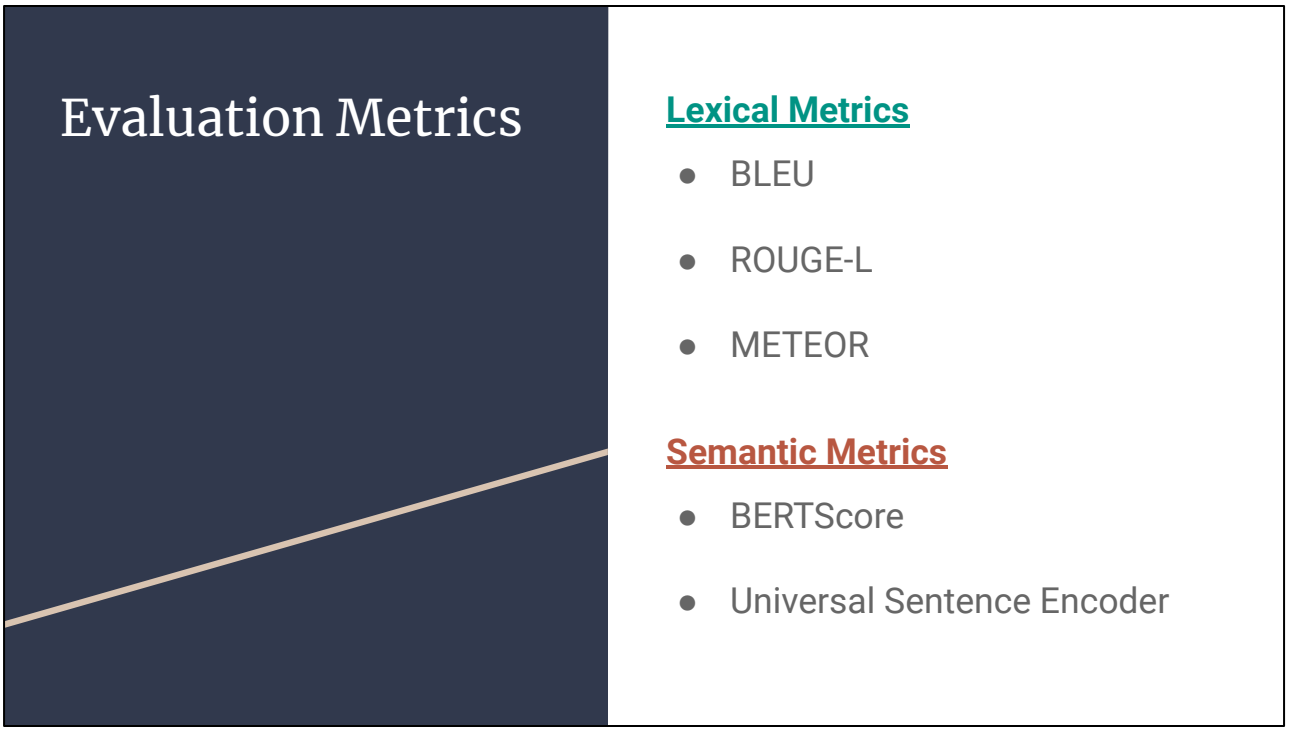
1. T5 and BART can be trained for question generation
2. Performance differences between them are slight
3. semantic metrics > lexical semantic + lexical is best
4. Training datasets impact performance; blending helps

Training Dataset	BLEU	ROUGE-L	METEOR	BERTScore	USE
Shuffled Blended	10.05	39.17	38.43	83.29	58.57
SQuAD	6.45	33.60	32.10	81.20	52.81

RICH

Before we talk about metrics and show some examples of questions our models generated, we share four primary observations.

- Both T5 and BART can be trained for question generation.
- There were no meaningful differences between the two.
- Metrics based on semantic similarity are better than metrics based on lexical similarity, but the best evaluations came when we used both kinds of metrics in harmony.
- Not surprisingly, the choice of dataset matters (which we explored in a number of ways). Our best all around performance came from models trained on the shuffled blended dataset.



JL

So how did we define success when our models generated questions?

We chose metrics that fall into two categories:

First is the **lexical group**, whose metrics you'll recognize from earlier weeks in this class. These metrics calculate similarity based the number of words in common.

Semantic metrics, on the other hand, use more modern tools like embeddings to capture conceptual meaning between sentences. We used:

- **BERTScore** because it utilizes BERT's contextual embeddings to measure the similarity of the words used in two sentences
- **Universal Sentence Encoder** is slightly different from BERTScore because it encodes a sentence into a single learned embedding. We can then take the cosine similarity between sentence embeddings.

Generated Questions

High Semantic & Low Lexical Scores

Target	Prediction
who is beauty and the beast written by	Who wrote La Belle et la Bête?
what is the first basic process in the light reaction of photosynthesis	How does photosynthesis begin?
What was the nationality of composer Frederic Chopin?	where did frédéric chopin come from
How large are Cytoplasmic ribosomes?	what is the size of a chloroplast ribosome
From which common English tree are cricket stumps traditionally made?	what kind of wood is used for stumps in cricket

High Semantic & High Lexical Scores

Target	Prediction
What parts of plants have chloroplasts?	Chloroplasts are found in which part of a plant?
what was the name of the other HD channel Virgin media could carry in the future?	What HD channel did Virgin Media have an option to carry in the future?
Approximately how many adherents does the United Methodist Church have across the world?	How many members are there in the United Methodist Church?
Singer Dan Reynolds is the frontman for which US rock band?	Dan Reynolds is the lead singer with which band?
Carbon dioxide in solid form is called what?	What is the frozen solid form of carbon dioxide known as?
Who's home runs record did Mark McGwire break in the 1998 season?	In 1998, Big Mac McGwire broke the Major League Baseball home run record by hitting 70 home runs. Who did he beat?

JL

Taking a look at some of the questions our models generated, we can observe some of the results that Rich mentioned before.

First off, these are coherent questions that maintain truthfulness to the target:

On the left, we see "What is the first basic…" vs "How does photosynthesis…"

These questions share the same essence but achieve it with different words.

Hence the high semantic scores but low lexical scores.

Next, on the right side we see questions with high semantic and high lexical scores.

Questions like "Singer Dan…" vs "Dan Reynolds is…"

And "Who's home runs" vs "in 1998…"

These pairs of questions align conceptually and have similar words in common.

They read fluently and actually make sense.

To summarize, semantic scores were generally the best indicators of high quality questions.

Lexical scores can be considered in tandem to maximize question quality and to control the language overlap between target and predicted questions.

Playing Around – Goldilocks . . .

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	the forest	Where did the three bears live?
T5	SQuAD	the forest	Where did the three bears go for a walk?
BART	Shuffled Blended	the forest	Where did Goldilocks live?
BART	SQuAD	the forest	Where did the three bears live?
T5	Shuffled Blended	it broke	What happened to the little chair?
T5	SQuAD	it broke	What happened to the little chair?
BART	Shuffled Blended	it broke	What happened to the little chair?
BART	SQuAD	it broke	What happened to the little chair?
T5	Shuffled Blended	it was too hot	What was the problem with the porridge?
T5	SQuAD	it was too hot	What was the problem with the porridge?
BART	Shuffled Blended	it was too hot	Why did Goldilocks want to eat porridge?
BART	SQuAD	it was too hot	Why did Goldilocks go for a walk in the forest?
T5	Shuffled Blended	tiny	What was the size of the baby bear?
T5	SQuAD	tiny	What was the size of the little bed that Goldilocks climbed into?
BART	Shuffled Blended	tiny	What was the size of Goldilocks's bed?
BART	SQuAD	tiny	How big was the little bed?

RICH

For fun, we tested our models on simple Children's stories including "Mary Had A Little Lamb" and lesser known not-quite classics including "Fossie Had A Little Lamb", "Irving Had A Little Lamb" and "Irving Had A Labradoodle Named Fossie"
(read the paper to see how little word changes impacted predictions and some examples of hallucination where our models confused Mary from the story with Mary from the bible . . .)

We gave it an AP news story recounting Cal's recent win over Stanford in the 125th edition of the Big Game.

Here we show some results when we gave some of our models the story of Goldilocks. Remember, we had no particular target question in mind – it just had to be a reasonable question given the story.

The questions in black are spot on. The ones in red are not so good. We don't know where Goldilocks lived, do we?

And I don't think Goldilocks went for a walk in the forest because it was hot.

The models we show generated good questions for "tiny" – and they have different meanings. Several things are described in the story as being "tiny".

There’s a lot more we could do with this project and a lot more we could talk about, but we ran out of weeks in the class and we are running out of time now. . .

Thank you

BRONTE

Questions?

BRONTE

Model Comparison

Training Dataset	Base Model	BLEU	ROUGE-L	METEOR	BERTScore	USE
Shuffled Blended	T5	10.05	39.17	38.43	83.29	58.57
Shuffled Blended	BART	9.87	38.87	38.60	83.15	58.40
SQuAD	BART	6.58	33.42	32.22	81.20	52.88
SQuAD	T5	6.45	33.60	32.10	81.20	52.81
TriviaQA	BART	4.47	29.44	30.90	79.61	50.75
NQ	BART	2.50	29.35	22.86	76.85	50.25
NQ	T5	2.35	29.80	23.00	76.92	50.10
TriviaQA	T5	3.59	26.75	29.13	79.01	48.49
QuAC	BART	2.47	26.01	22.00	78.02	42.92
QuAC	T5	0.50	9.17	9.78	70.55	17.93

-