“What makes a question inquisitive?”

A study on type-controlled inquisitive question generation

Lingyu Gao[1], Debanjan Ghosh[2], Kevin Gimpel[1]

1 TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO, 2 EDUCATIONAL TESTING SERVICE

11th Joint Conference on Lexical and Computational Semantics (*SEM), NAACL 2022

What is inquisitive question generation?

Source Sentence:

Santa Fe Pacific directors are expected to review the plan at a meeting today, according to people familiar with the transaction.

Informative

- What are Santa Fe Pacific directors expected to review?

What is inquisitive question generation?

Source Sentence:

Santa Fe Pacific directors are expected to review the plan at a meeting today, according to people familiar with the transaction.

Informative

- What are Santa Fe Pacific directors expected to review?

Inquisitive

- Why are they reviewing the plan?
- What will the review entail?

Motivation

- Automatically generating **inquisitive questions** controlled with question type
    - Seeking high level understanding of text
    - Closer to **human reader**'s natural thoughts
    - Curiosity-driven
    - For Educators: Obtain diverse questions for a specific source text
    - For Student: Build reasoning skills by practicing

Outline

- Research Questions
- Data
- Method
- Evaluations
- Conclusion

Outline

- **Research Questions**
- Data
- Method
- Evaluations
- Conclusion

Research Questions

- How to generate **diverse inquisitive** questions?

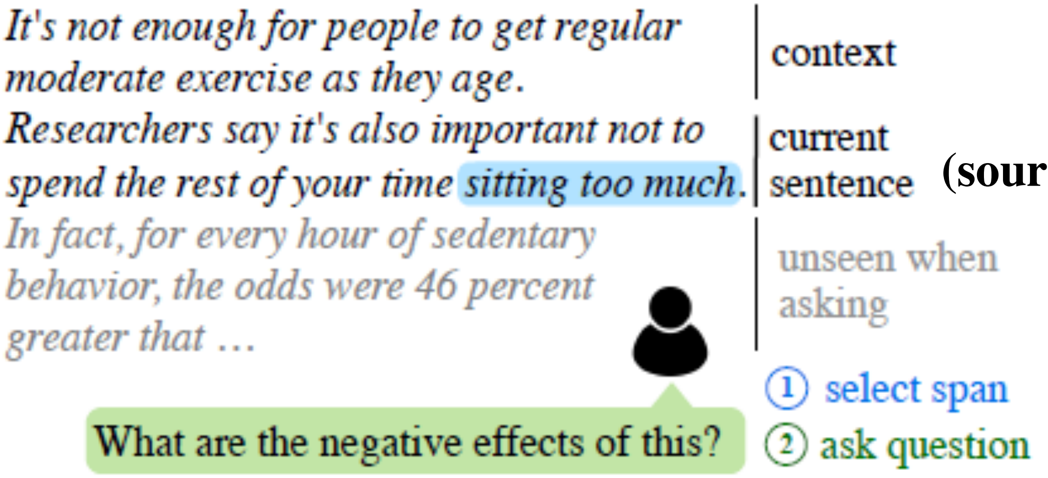
- How to evaluate the **quality** of the generated questions?

- How to select the **single high-quality** question or to rank them?

Outline

- Research Questions
- **Data**
- Method
- Evaluations
- Conclusion

Data



It's not enough for people to get regular moderate exercise as they age.

Researchers say it's also important not to spend the rest of your time sitting too much.

In fact, for every hour of sedentary behavior, the odds were 46 percent greater that …

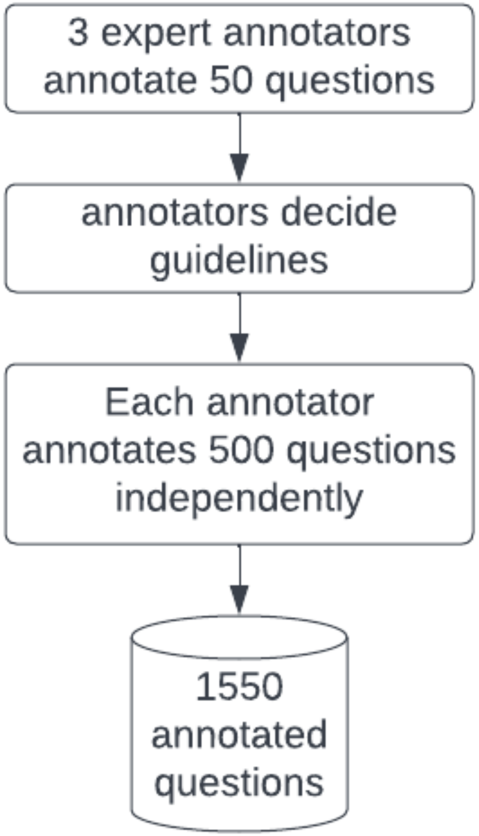
What are the negative effects of this?

(source sentence)

Train	Dev	Test
15897	1984	1885

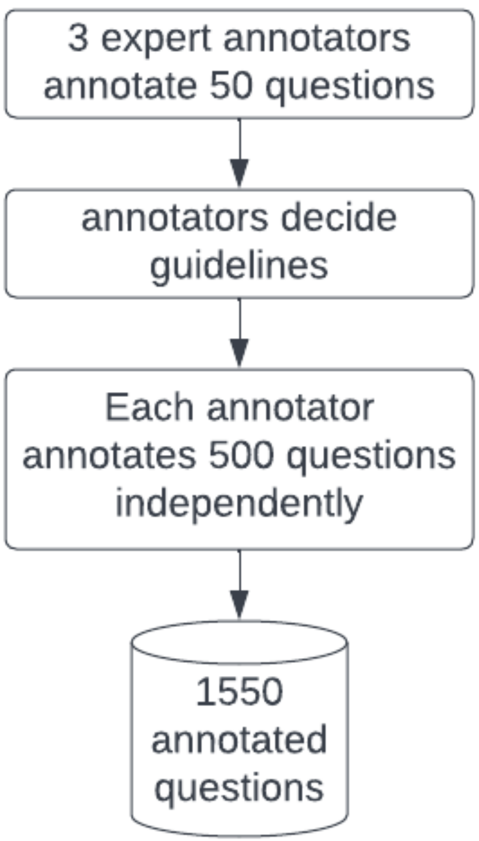
Wei-Jen Ko, et al. Inquisitive question generation for high level text comprehension. EMNLP 2020.

Data: Annotation of question types


3 expert annotators annotate 50 questions
annotators decide guidelines
Each annotator annotates 500 questions independently
1550 annotated questions

- Annotation inspired by the rhetorical structural theory (RST) on discourse types.
  - Explanation
  - Background
  - Elaboration
  - Instantiation
  - Definition
  - Forward
  - Other (e.g., inference questions)

Data: Annotation of question types

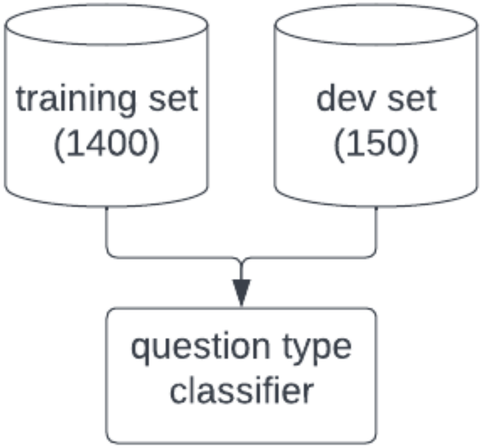


3 expert annotators annotate 50 questions

annotators decide guidelines

Each annotator annotates 500 questions independently

1550 annotated questions



training set (1400)

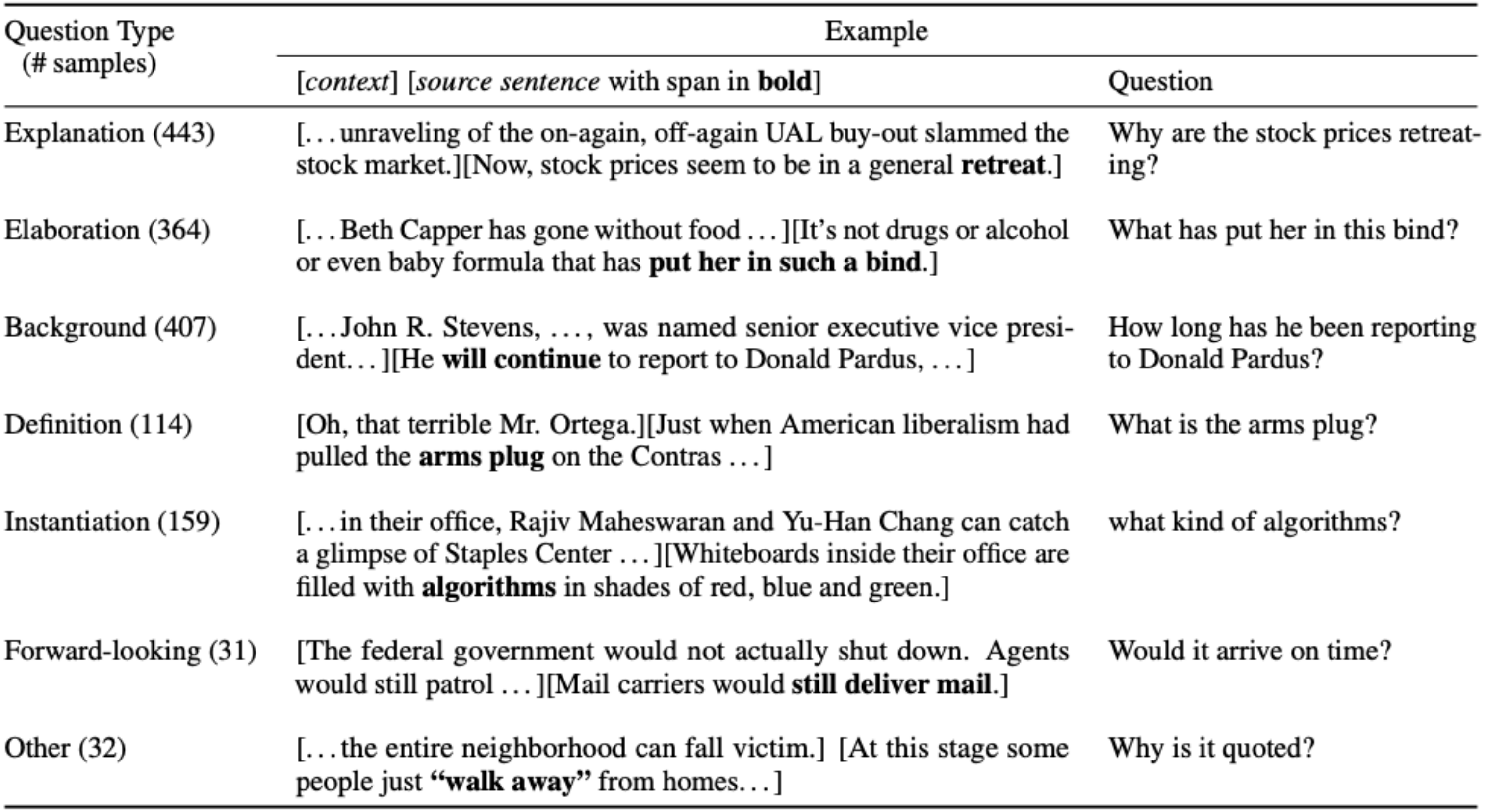
dev set (150)

question type classifier

- Question type classifier:
  - Input: concatenate context, source, span, question
  - RoBERTa: dev acc: 73.3%

- Generate question types for all the remaining data

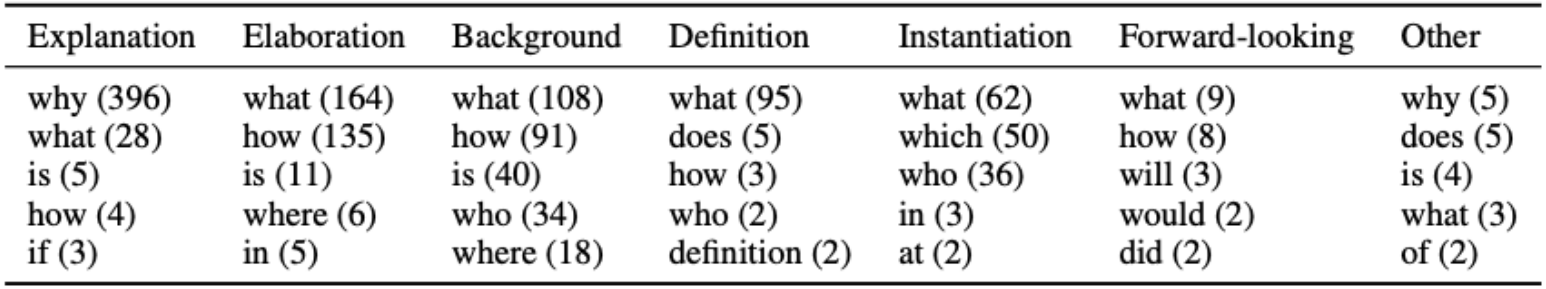
Data: question types



Question Type (# samples)	Example	
	[*context*] [*source sentence* with span in **bold**]	Question
Explanation (443)	[. . . unraveling of the on-again, off-again UAL buy-out slammed the stock market.][Now, stock prices seem to be in a general **retreat**.]	Why are the stock prices retreating?
Elaboration (364)	[. . . Beth Capper has gone without food . . .][It's not drugs or alcohol or even baby formula that has **put her in such a bind**.]	What has put her in this bind?
Background (407)	[. . . John R. Stevens, . . . , was named senior executive vice president. . .][He **will continue** to report to Donald Pardus, . . .]	How long has he been reporting to Donald Pardus?
Definition (114)	[Oh, that terrible Mr. Ortega.][Just when American liberalism had pulled the **arms plug** on the Contras . . .]	What is the arms plug?
Instantiation (159)	[. . . in their office, Rajiv Maheswaran and Yu-Han Chang can catch a glimpse of Staples Center . . .][Whiteboards inside their office are filled with **algorithms** in shades of red, blue and green.]	what kind of algorithms?
Forward-looking (31)	[The federal government would not actually shut down. Agents would still patrol . . .][Mail carriers would **still deliver mail**.]	Would it arrive on time?
Other (32)	[. . . the entire neighborhood can fall victim.] [At this stage some people just **"walk away"** from homes. . .]	Why is it quoted?

Data: question types

- Can we use a dedicated WH question for a single question type? (Zhou et al. 2019)
  - *Not, really…*



Explanation	Elaboration	Background	Definition	Instantiation	Forward-looking	Other
why (396)	what (164)	what (108)	what (95)	what (62)	what (9)	why (5)
what (28)	how (135)	how (91)	does (5)	which (50)	how (8)	does (5)
is (5)	is (11)	is (40)	how (3)	who (36)	will (3)	is (4)
how (4)	where (6)	who (34)	who (2)	in (3)	would (2)	what (3)
if (3)	in (5)	where (18)	definition (2)	at (2)	did (2)	of (2)

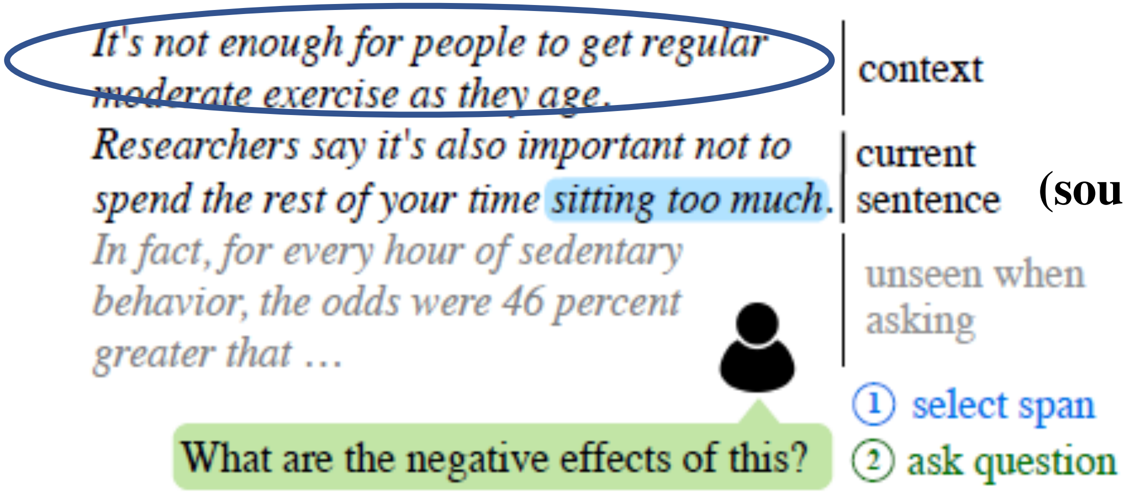
- WH question words cannot fully express the semantic content of questions

Outline

- Research Questions
- Data
- **Method**
- Evaluations
- Conclusion

Method

- State of the art (Ko et al., 2020):
  - Language model (input: context, source, span, gold questions) using GPT-2 transformers



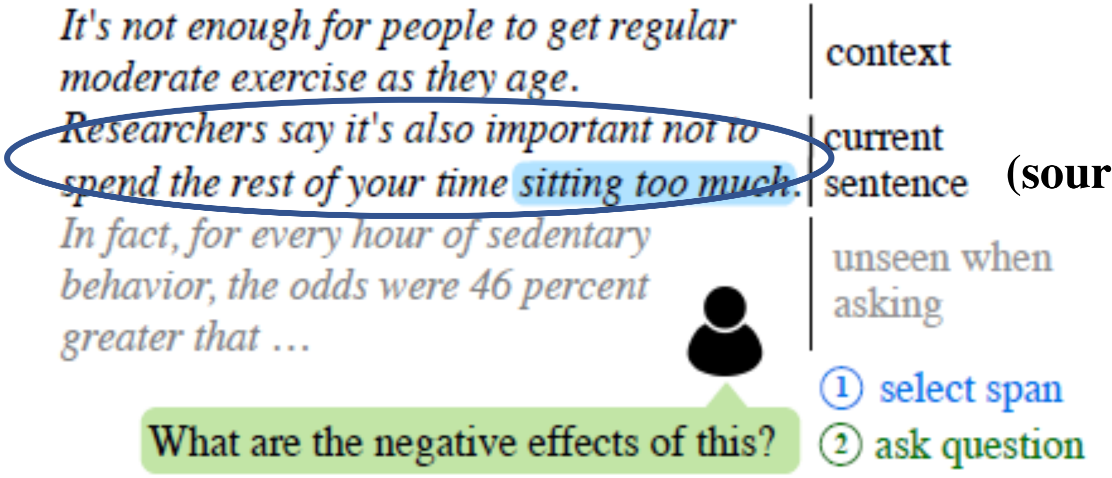
It's not enough for people to get regular moderate exercise as they age. Researchers say it's also important not to spend the rest of your time sitting too much. In fact, for every hour of sedentary behavior, the odds were 46 percent greater that …

What are the negative effects of this?

(source sentence)

Method

- State of the art (Ko et al., 2020):
  - Language model (input: context, source, span, gold questions) using GPT-2 transformers



It's not enough for people to get regular moderate exercise as they age.

Researchers say it's also important not to spend the rest of your time sitting too much.

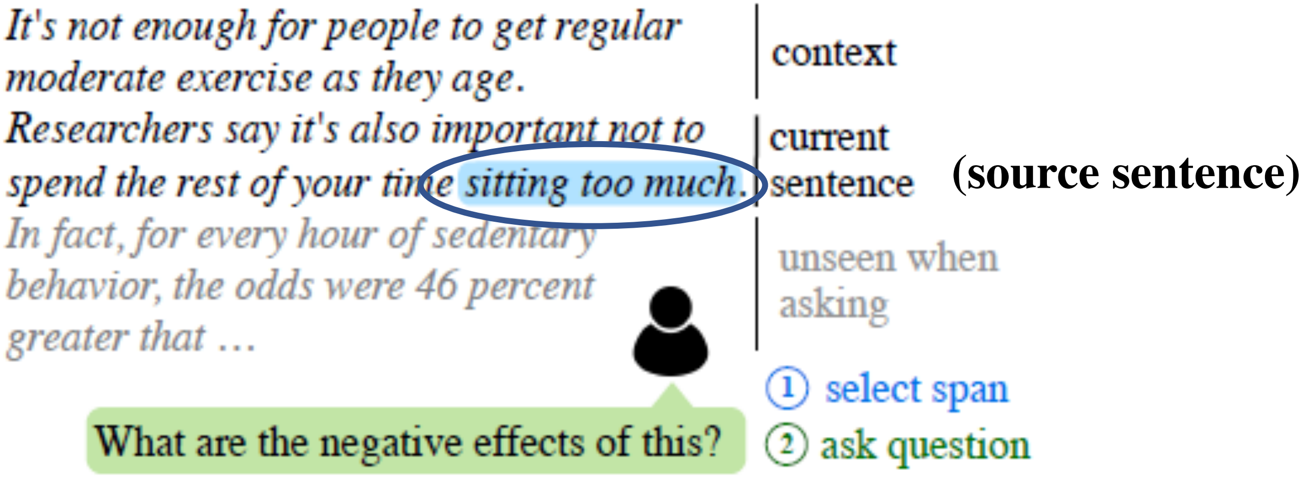
In fact, for every hour of sedentary behavior, the odds were 46 percent greater that …

What are the negative effects of this?

(source sentence)

Method

- State of the art (Ko et al., 2020):
  - Language model (input: context, source, span, gold questions) using GPT-2 transformers



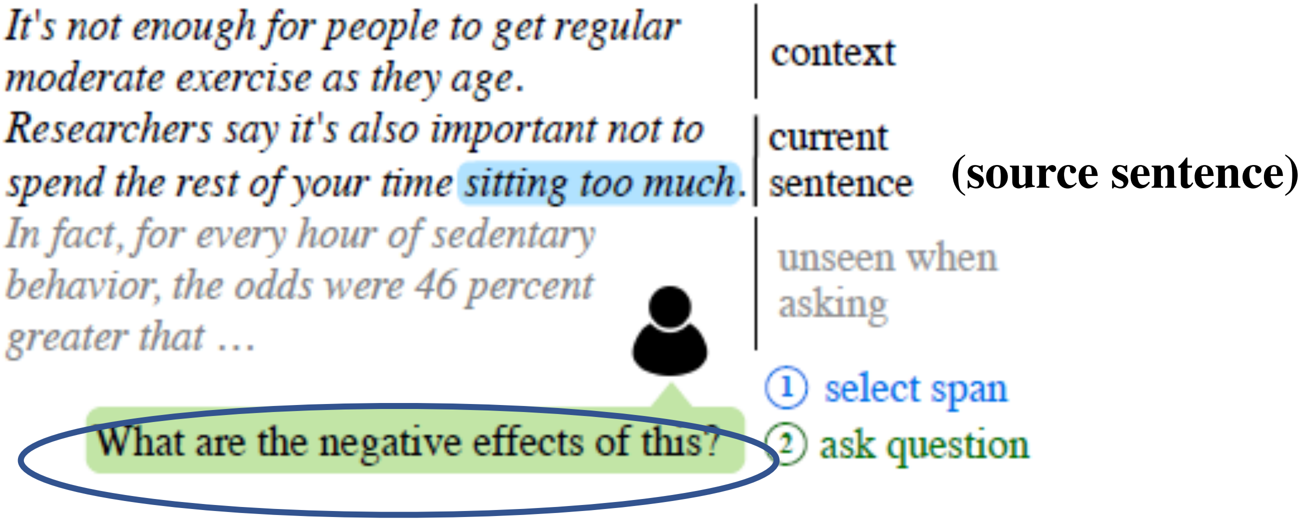
It's not enough for people to get regular moderate exercise as they age. Researchers say it's also important not to spend the rest of your time sitting too much. In fact, for every hour of sedentary behavior, the odds were 46 percent greater that …

What are the negative effects of this?

(source sentence)

Method

- State of the art (Ko et al., 2020):
  - Language model (input: context, source, span, gold questions) using GPT-2 transformers



It's not enough for people to get regular moderate exercise as they age. Researchers say it's also important not to spend the rest of your time sitting too much. In fact, for every hour of sedentary behavior, the odds were 46 percent greater that …

What are the negative effects of this?

Method

- State of the art (Ko et al., 2020):
  - Language model (input: context, source, span, gold questions) using GPT-2 transformers
- Our model:
  - Seq2seq using BART (bidirectional encoder + auto-regressive decoder; Lewis et al. 2020)
  - Methods:
    1. Conditional Generation (adding control codes)
    2. Automatic Question Type Selection

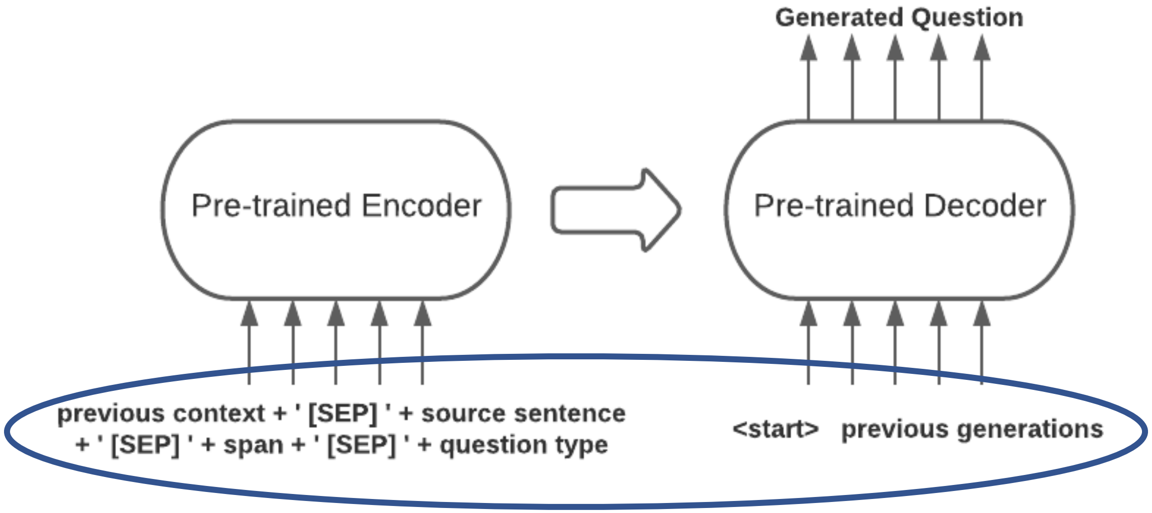
Method

- **Conditional Generation (adding control codes)**
- Automatic Question Type Selection

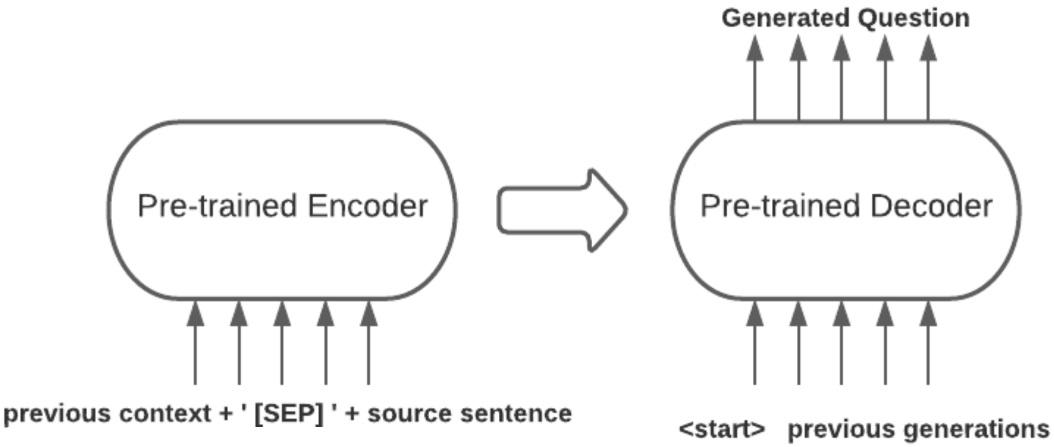
Method: conditional generation

- Conditional generation by adding control codes to source sentence (Syed et al. 2021)

$$p_{lm}(y|x) \rightarrow p_{lm}(y|x, c)$$


Generated Question
Pre-trained Encoder
Pre-trained Decoder
previous context + ' [SEP] ' + source sentence + ' [SEP] ' + span + ' [SEP] ' + question type
<start> previous generations

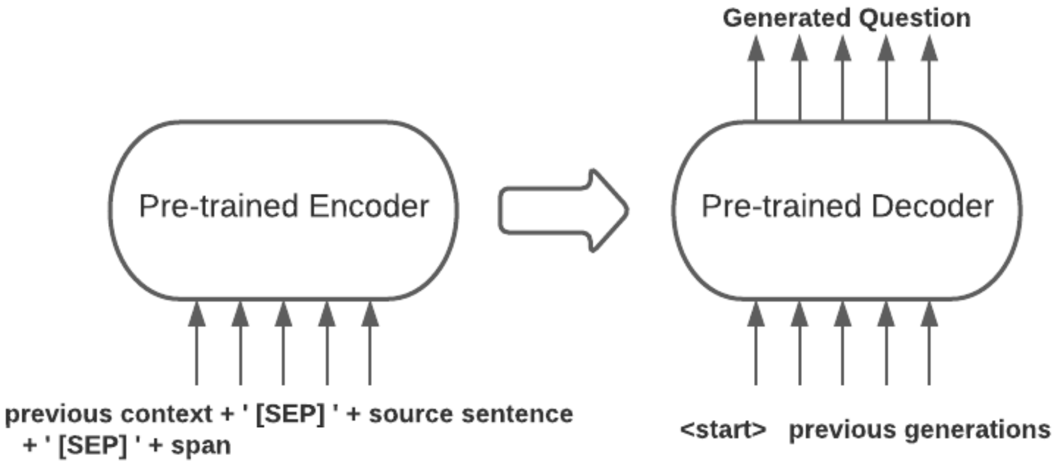
Method: examples



BASE (context + source):

People start their own businesses for many reasons. But a chance to fill out sales - tax records is rarely one of them. [SEP] Red tape is the bugaboo of small business.

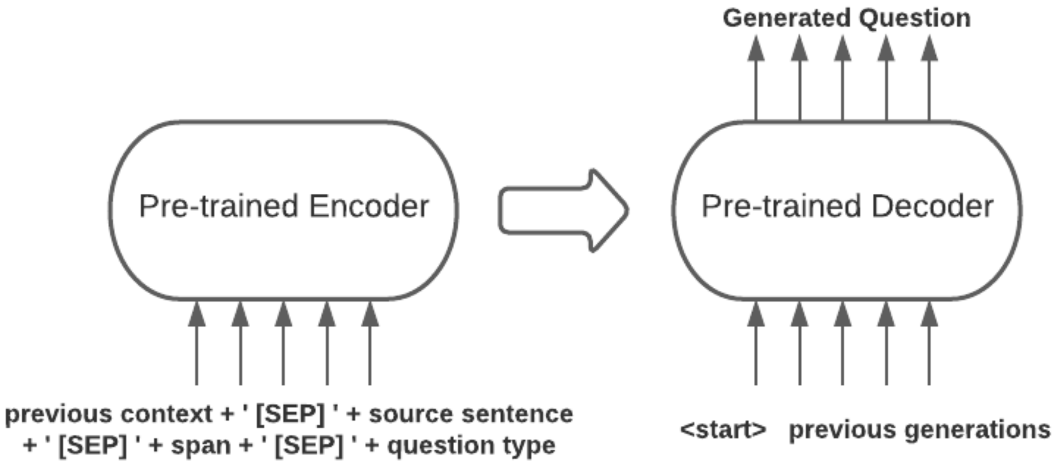
Method: examples


Generated Question
Pre-trained Encoder
Pre-trained Decoder
previous context + ' [SEP] ' + source sentence + ' [SEP] ' + span
<start>   previous generations

SPAN (context + source + span):

People start their own businesses for many reasons. But a chance to fill out sales - tax records is rarely one of them. [SEP] Red tape is the bugaboo of small business. [SEP] bugaboo

Method: examples


Generated Question
Pre-trained Encoder
Pre-trained Decoder
previous context + ' [SEP] ' + source sentence + ' [SEP] ' + span + ' [SEP] ' + question type
<start>
previous generations

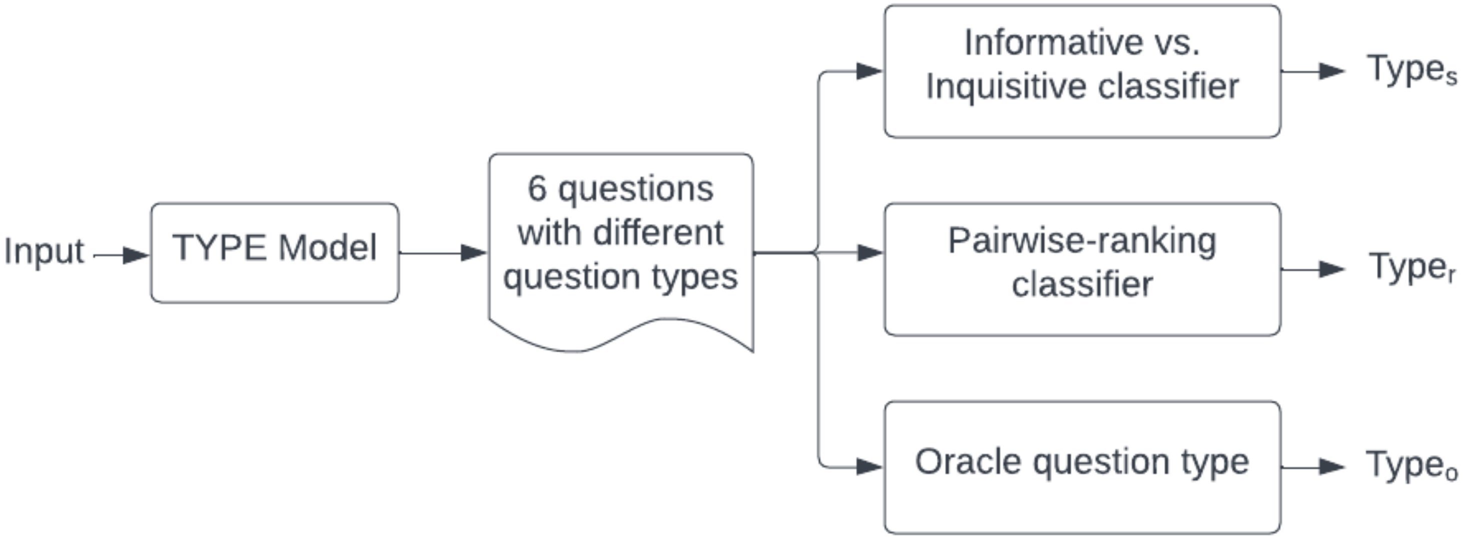
TYPE (context + source + span + question type):

People start their own businesses for many reasons. But a chance to fill out sales - tax records is rarely one of them. [SEP] Red tape is the bugaboo of small business. [SEP] bugaboo [SEP] Definition

Method

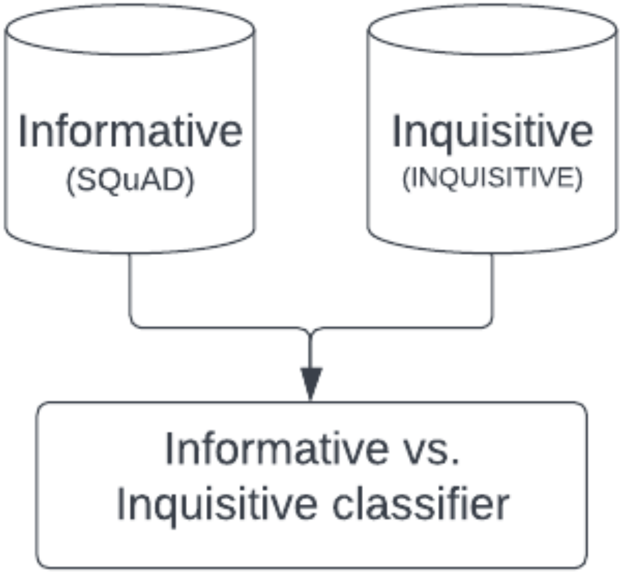
- Conditional Generation
- **Automatic Question Type Selection**

Method: Automatic Question Type Selection


Input
TYPE Model
6 questions with different question types
Informative vs. Inquisitive classifier
Type_s
Pairwise-ranking classifier
Type_r
Oracle question type
Type_o

“Other” question type is removed because it includes too many subtypes.

Informative vs. inquisitive classifier

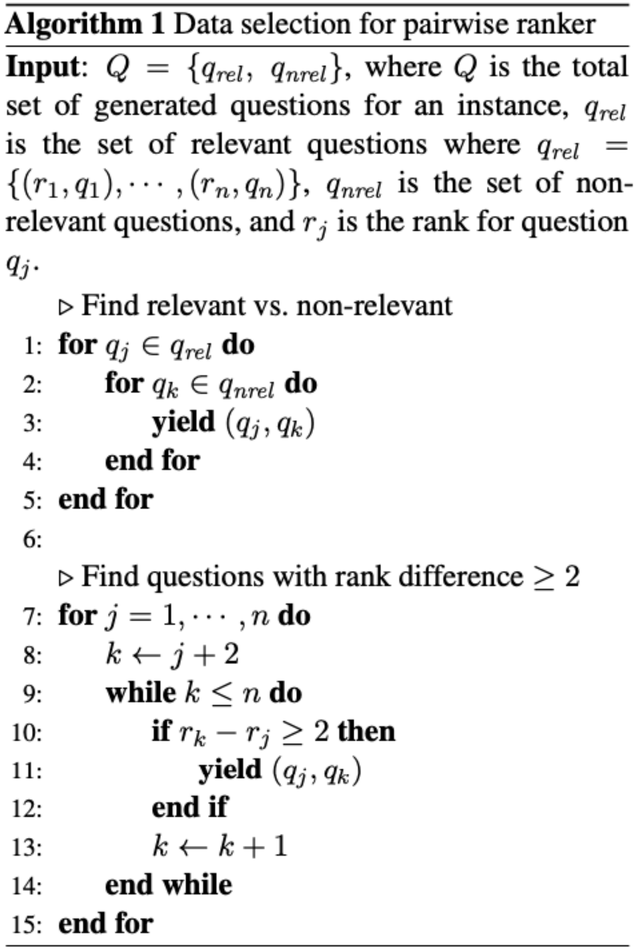


- Binary question classifier:
  - Input: question

- Dataset (fully balanced):
  - Training set: 16,000
  - Dev set: 3000
  - Test set: 3000

- *We did not use the context or source text here because the model was highly influenced by the text type (wiki vs. news data)*

Pairwise-ranking classifier



Algorithm 1 Data selection for pairwise ranker

Input: $Q = \{q_{rel}, q_{nrel}\}$, where Q is the total set of generated questions for an instance, q_{rel} is the set of relevant questions where $q_{rel} = \{(r_1, q_1), \cdots, (r_n, q_n)\}$, q_{nrel} is the set of non-relevant questions, and r_j is the rank for question q_j.

```
    ▷ Find relevant vs. non-relevant
for $q_j \in q_{rel}$ do
    for $q_k \in q_{nrel}$ do
        yield $(q_j, q_k)$
    end for
end for

    ▷ Find questions with rank difference $\geq 2$
for $j = 1, \cdots, n$ do
    $k \leftarrow j + 2$
    while $k \leq n$ do
        if $r_k - r_j \geq 2$ then
            yield $(q_j, q_k)$
        end if
        $k \leftarrow k + 1$
    end while
end for
```

- Manual Annotation (300 test instances)
  - Select at least 3 questions as the best with ranks

- Pairwise-ranking classifier
  - Input:
    source + [SEP] + q1 + [SEP] + q2
  - Winning question: the one is selected the most number of times
  - In case of a tie check the classifier' score

Outline

- Research Questions
- Data
- Method
- **Evaluations**
- Conclusion and Future Work

Evaluations

- Automatic Metrics

- Human Evaluations

Evaluations

- Automatic Metrics
  - BLEU, METEOR, ROUGE-L
  - BERTScore
  - Perplexity under GPT2-XL
  - Entropy (averaged over questions) of the question type classifier
  - Specific metrics measure the overlap of text between generation and source text (Ko et al. 2020)

Evaluations

- Automatic Metrics
  - Train-n: overlap with questions in the training set

$$\text{Train-}n = \frac{\text{Count}(w_{i:n+i} \in Q_G \cap Q_T)}{\text{Count}(w_{i:n+i} \in Q_G)}$$

  - Article-n: overlap with the current sentence or the previous context in the same article

$$\text{Article-}n = \frac{\text{Count}(w_{i:n+i} \in Q_G \cap (S_{Sent} \cup S_{Context}))}{\text{Count}(w_{i:n+i} \in Q_G)}$$

  - Span: overlap with the annotated span

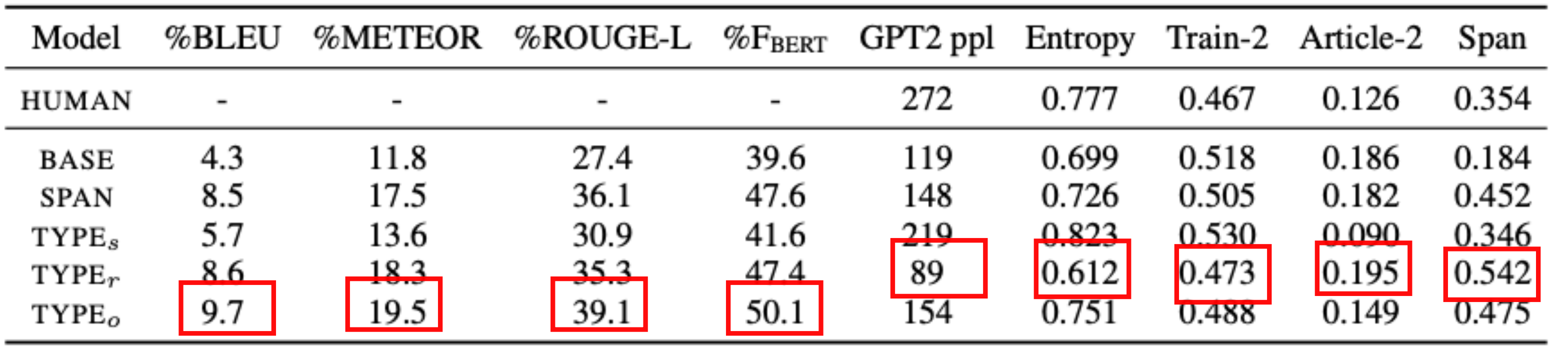
$$\text{Span} = \frac{\text{Count}(w_{i:n+i} \in S_{Span} \cap Q_G)}{\text{Count}(w_{i:n+i} \in S_{Span})}$$

Evaluation: Results

TYPE$_s$: classifier output

TYPE$_r$: pairwise ranker output

TYPE$_o$: oracle



Model	%BLEU	%METEOR	%ROUGE-L	%F$_{BERT}$	GPT2 ppl	Entropy	Train-2	Article-2	Span
HUMAN	-	-	-	-	272	0.777	0.467	0.126	0.354
BASE	4.3	11.8	27.4	39.6	119	0.699	0.518	0.186	0.184
SPAN	8.5	17.5	36.1	47.6	148	0.726	0.505	0.182	0.452
TYPE$_s$	5.7	13.6	30.9	41.6	219	0.823	0.530	0.090	0.346
TYPE$_r$	8.6	18.3	35.3	47.4	89	0.612	0.473	0.195	0.542
TYPE$_o$	9.7	19.5	39.1	50.1	154	0.751	0.488	0.149	0.475

- TYPE$_o$ is best for BLEU, METEOR, ROUGE-L and BERTScore
- TYPE$_r$ has the lowest GPT2 perplexity and Entropy
- TYPE$_r$ has the lowest Train-2, highest Article-2 and Span scores
- SPAN is a very competitive method (undoubtedly!)

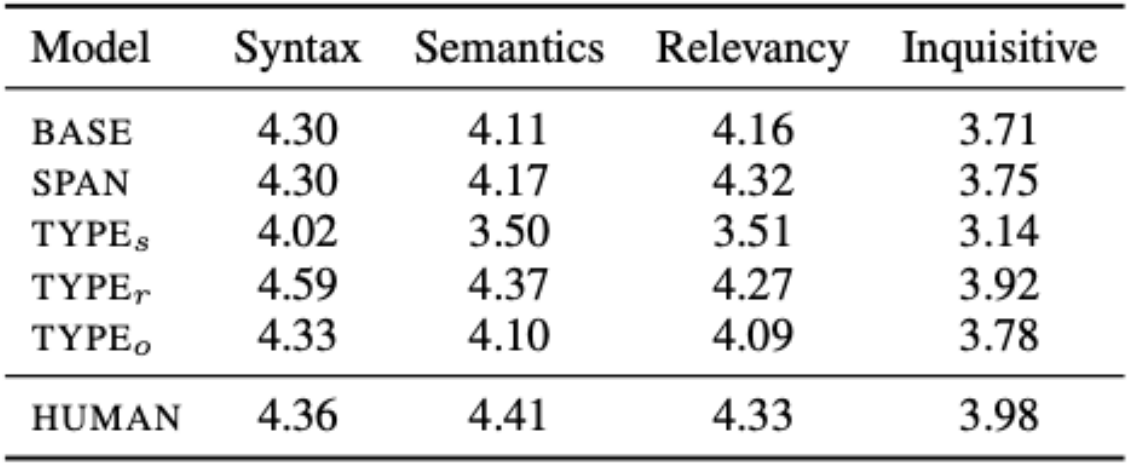
Evaluations

- Human Evaluations
  - Large scale Mturk evolution over 500 questions/each type with 3 Turkers.
    - Syntax
      - Grammatically correct?
    - Semantic
      - Meaningful or not? Are there hallucinations?
    - Relevancy
      - How relevant is the inquisitive question to the source?
    - Inquisitive
      - Asking deeper info such as background information?

Human Evaluation

- Annotator manually annotated 500 test instances

- Scoring between 1, 3 or 5 [1 is lowest, 5 is highest]



Model	Syntax	Semantics	Relevancy	Inquisitive
BASE	4.30	4.11	4.16	3.71
SPAN	4.30	4.17	4.32	3.75
TYPE$_s$	4.02	3.50	3.51	3.14
TYPE$_r$	4.59	4.37	4.27	3.92
TYPE$_o$	4.33	4.10	4.09	3.78
HUMAN	4.36	4.41	4.33	3.98

- Scores are above 4, often
- TYPE$_r$ and SPAN are competitive
- TYPE$_s$ is low given they often select "definition" and "instantiation" questions

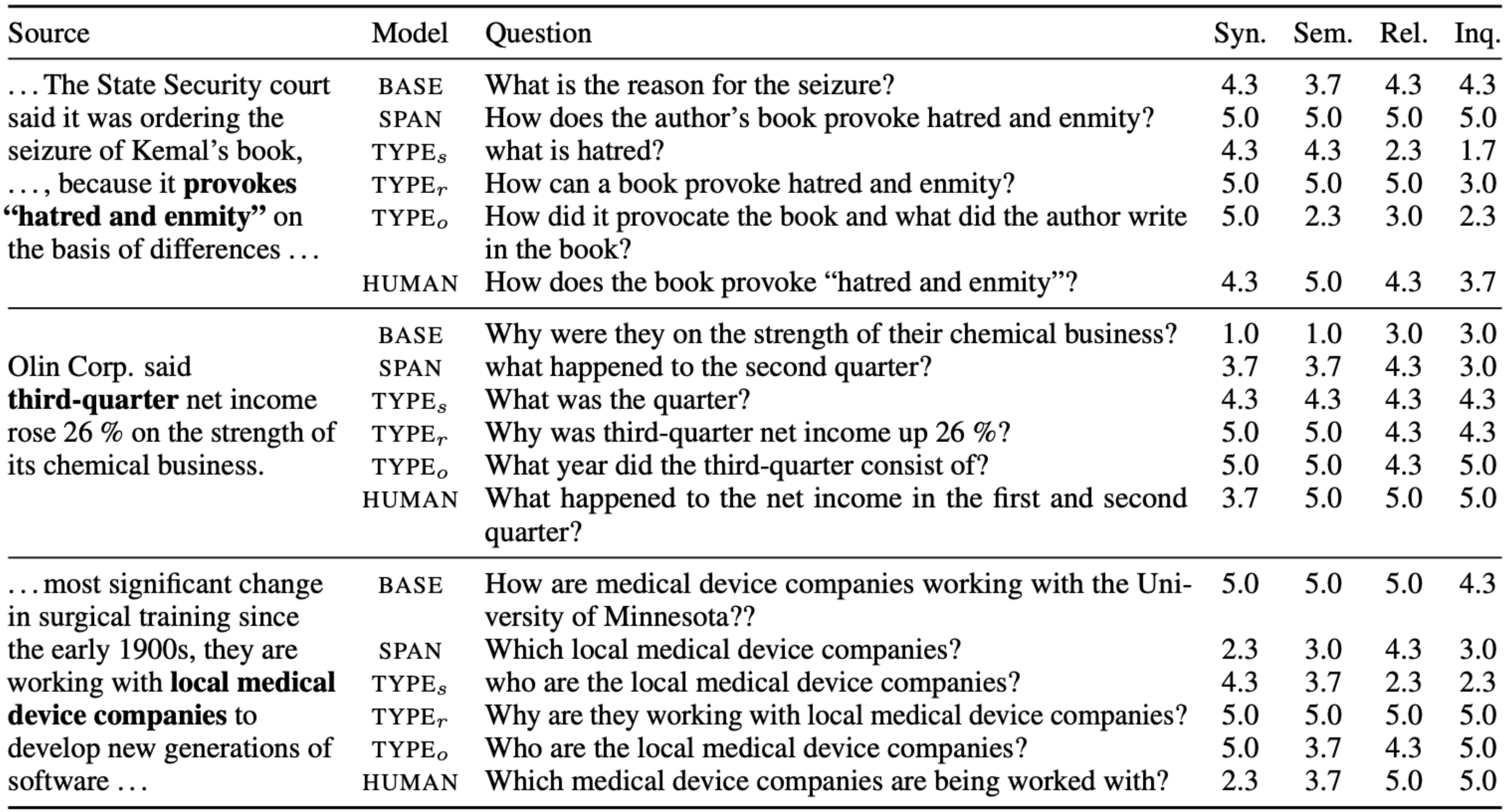
Human Evaluation

- *Wait, how does the Syntax score for TYPE$_r$ is better than HUMAN?*

- Analyzed all the questions from TYPE$_r$ and HUMAN where the majority of annotators rated 1 for syntax, and we found out there are 14 and 33 such questions, respectively (explains the high perplexity for HUMAN, too)

is it the aha?
how much has inflation?
nativity happens for buddha?
When he decide?
how much has inflation

Example Outputs



Source	Model	Question	Syn.	Sem.	Rel.	Inq.
…The State Security court said it was ordering the seizure of Kemal's book, …, because it **provokes "hatred and enmity"** on the basis of differences …	BASE	What is the reason for the seizure?	4.3	3.7	4.3	4.3
	SPAN	How does the author's book provoke hatred and enmity?	5.0	5.0	5.0	5.0
	TYPE$_s$	what is hatred?	4.3	4.3	2.3	1.7
	TYPE$_r$	How can a book provoke hatred and enmity?	5.0	5.0	5.0	3.0
	TYPE$_o$	How did it provocate the book and what did the author write in the book?	5.0	2.3	3.0	2.3
	HUMAN	How does the book provoke "hatred and enmity"?	4.3	5.0	4.3	3.7
Olin Corp. said **third-quarter** net income rose 26 % on the strength of its chemical business.	BASE	Why were they on the strength of their chemical business?	1.0	1.0	3.0	3.0
	SPAN	what happened to the second quarter?	3.7	3.7	4.3	3.0
	TYPE$_s$	What was the quarter?	4.3	4.3	4.3	4.3
	TYPE$_r$	Why was third-quarter net income up 26 %?	5.0	5.0	4.3	4.3
	TYPE$_o$	What year did the third-quarter consist of?	5.0	5.0	4.3	5.0
	HUMAN	What happened to the net income in the first and second quarter?	3.7	5.0	5.0	5.0
…most significant change in surgical training since the early 1900s, they are working with **local medical device companies** to develop new generations of software …	BASE	How are medical device companies working with the University of Minnesota??	5.0	5.0	5.0	4.3
	SPAN	Which local medical device companies?	2.3	3.0	4.3	3.0
	TYPE$_s$	who are the local medical device companies?	4.3	3.7	2.3	2.3
	TYPE$_r$	Why are they working with local medical device companies?	5.0	5.0	5.0	5.0
	TYPE$_o$	Who are the local medical device companies?	5.0	3.7	4.3	5.0
	HUMAN	Which medical device companies are being worked with?	2.3	3.7	5.0	5.0

Outline

- Research Questions
- Data
- Method
- Evaluation
- **Conclusion**

Conclusion

- We proposed a type-controlled framework that generates inquisitive questions
- We annotated a set of question types related to curiosity driven questions and demonstrated that our framework can generate a variety of questions from a single input
- We developed an effective method (TYPE_r) to select a single question using a pairwise ranker trained on a small set of ranking annotations
- Our generations show high novelty. Questions generated from TYPE_r rival human-written questions on all four aspects of quality based on human evaluation

Thank you!

Acknowledgements:

Yuan Wang and Hillary Molloy
Mengxuan Zhao and Rob Bobkoskie
Swapna Somasundaran