RANLP 2021: On Generating Fact-Infused Question Variations

Sujatha Das Gollapalli

Joint Work with: Arthur Deschamps and See-Kiong Ng

Institute of Data Science
National University of Singapore

Overview of the talk

- Motivation What is a fact-infused variation for a given question?
- New Dataset FIRS (**F**act-**I**nfused **R**ewrites of **S**QuAD questions)
- Our Model for Fact-Infused Question Generation
- Results

Background: Question Generation and Paraphrasing

- **Reverse of the Question Answering Task**
  - Given an input passage, (optional) answer focus segment from the passage, generate a natural language question
  - Uses: Tutoring in Educational Applications, Dialog Systems and Chatbots, improving robustness in QA systems
  - By far, QA and QG models are the most studied in extractive settings/factoid questions among other types

- **Paraphrasing**
  - What is the best way to reduce cholesterol without medicines? versus

    How can I reduce my cholesterol naturally?

    - Community QA such as forums/Quora

Fact-Infused Variations

- Limitation: Existing paraphrase datasets: syntactic variations of questions
  - Question: Who killed **Abraham Lincoln**? Answer: *John Wilkes Booth*

    Syntactic variations: Who shot Abraham Lincoln to death?

    Who assassinated Abraham Lincoln?
- How about adding more details to questions?
  - Add facts pertaining to entities related to the question
  - Two possibilities:
    - Who killed the American president Abraham Lincoln?
    - Which stage actor killed Abraham Lincoln?
- Question expansion, improves answerability, testing robustness of QA systems, educational scenarios: increasing difficulty-level or requiring reasoning

Where can we find entity facts?

FIRS dataset: Fact-Infused Rewrites of SQuAD questions

Collecting candidate question-answer pairs from SQuAD with Named-Entity mentions in the question or the answer

Question: In what year did IBM get its name? (ORGANIZATION)

Answer: 1924 (DATE)

Look up publicly-available resources/knowledge-bases for “tangible” entities

We chose Google Knowledge Graph API based on initial lookup experiments

Query-> result tuples (name, type, description, detailed-description)

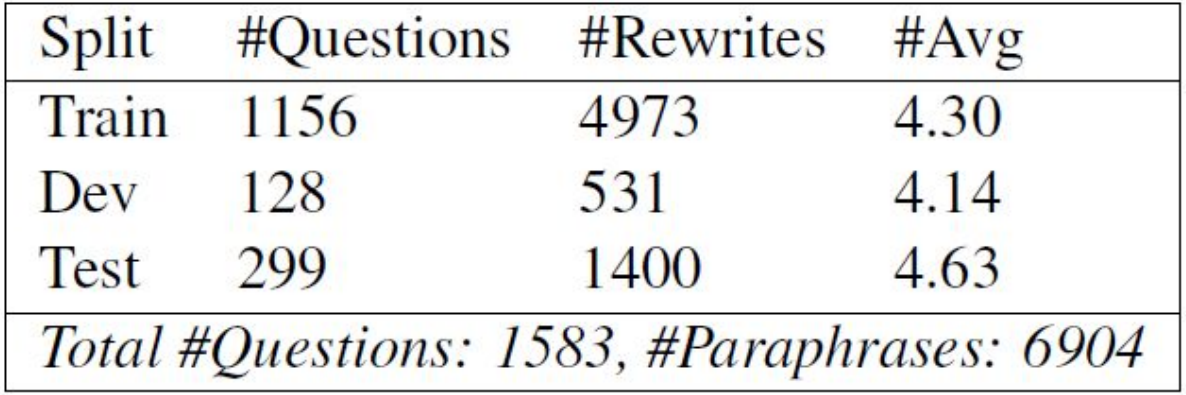
Similarity filters based on the SQuAD passage and description fields, type match filter to choose the best matching entity description

FIRS creation with Amazon Mechanical Turk

(Provide original SQuAD passage, question, and answer

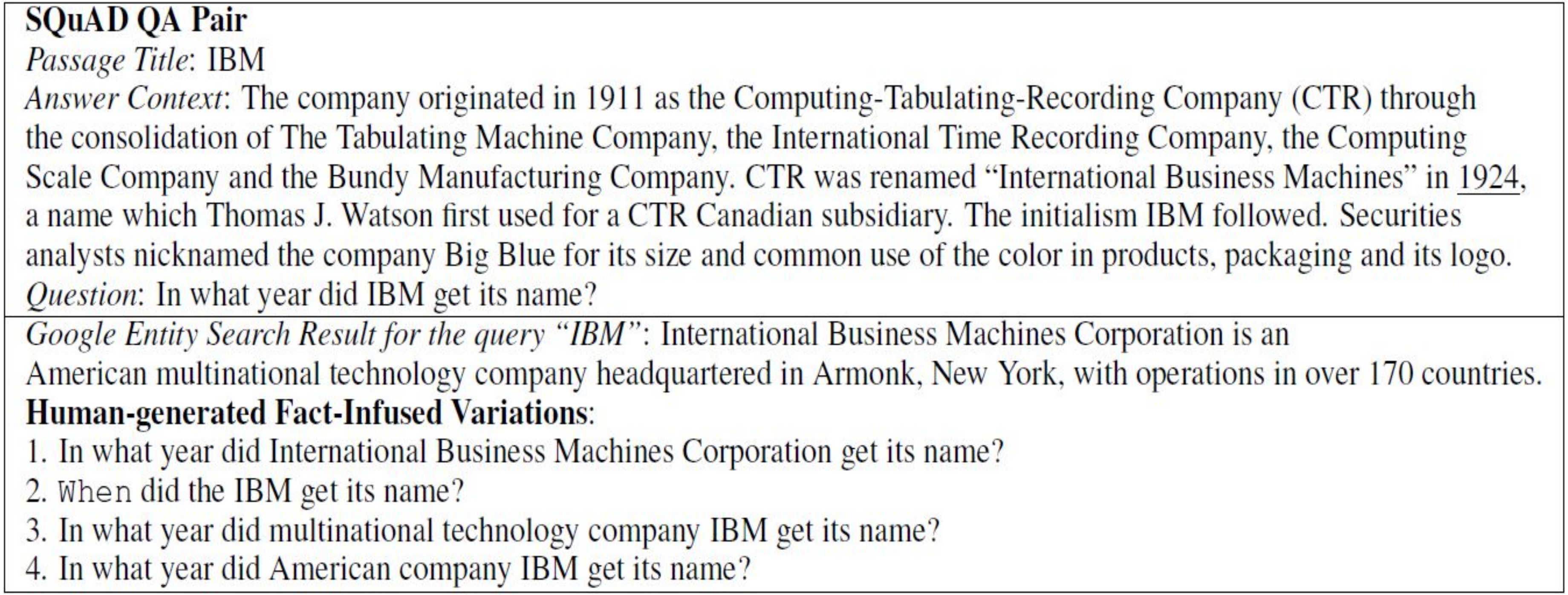
+ Entity descriptions from GKG via the Entity Search API)

Ask crowdworkers to produce questions that have the same intent as the original but also incorporating a fact from the entity description



Split	#Questions	#Rewrites	#Avg
Train	1156	4973	4.30
Dev	128	531	4.14
Test	299	1400	4.63
Total #Questions: 1583, #Paraphrases: 6904			

Fact-infused Variations: Example from our dataset



SQuAD QA Pair

Passage Title: IBM

Answer Context: The company originated in 1911 as the Computing-Tabulating-Recording Company (CTR) through the consolidation of The Tabulating Machine Company, the International Time Recording Company, the Computing Scale Company and the Bundy Manufacturing Company. CTR was renamed "International Business Machines" in 1924, a name which Thomas J. Watson first used for a CTR Canadian subsidiary. The initialism IBM followed. Securities analysts nicknamed the company Big Blue for its size and common use of the color in products, packaging and its logo.

Question: In what year did IBM get its name?

Google Entity Search Result for the query "IBM": International Business Machines Corporation is an American multinational technology company headquartered in Armonk, New York, with operations in over 170 countries.

Human-generated Fact-Infused Variations:

1. In what year did International Business Machines Corporation get its name?
2. When did the IBM get its name?
3. In what year did multinational technology company IBM get its name?
4. In what year did American company IBM get its name?

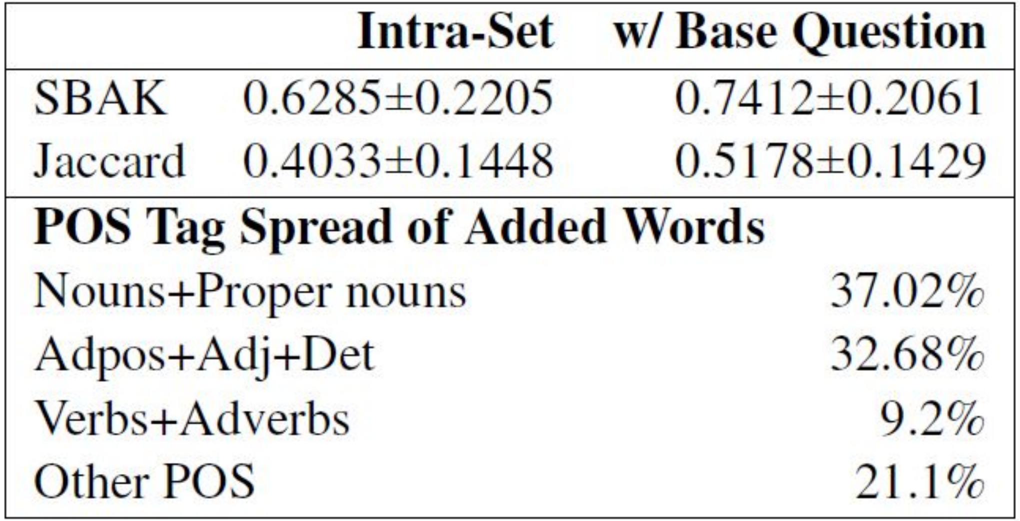
Properties of FIRS

Simple **B**igram **A**pproximate **K**ernel based on dependency trees of the sentences

Higher value indicates higher semantic similarity

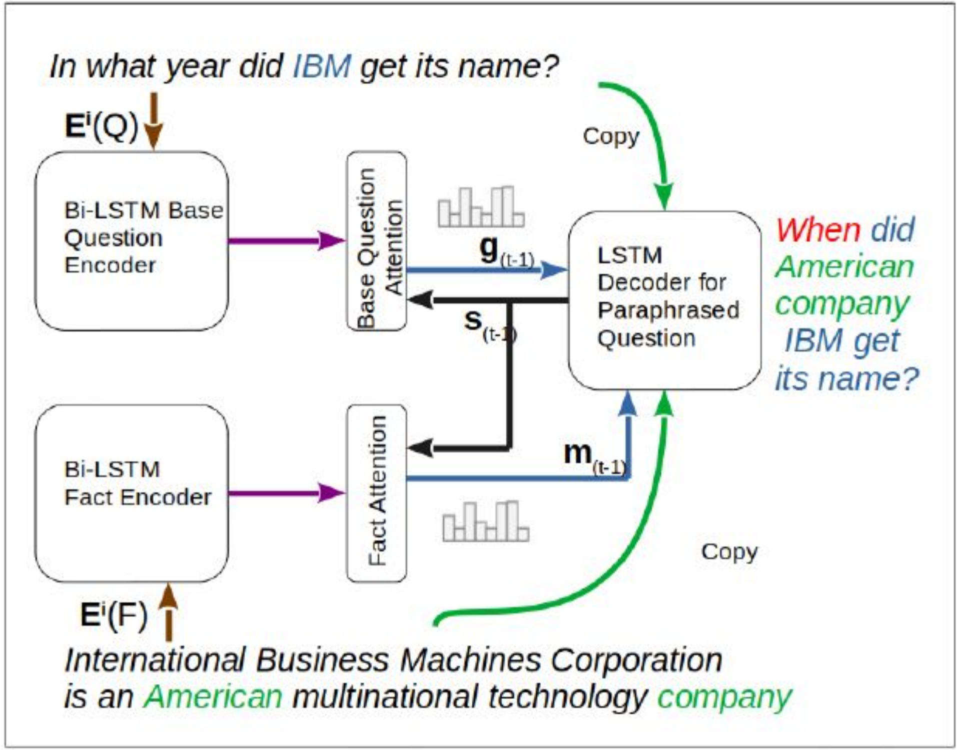
Not surprisingly, Jaccard < SBAK

Added words dominated by NOUNs or words used with NOUNs (that is, content words)



	Intra-Set	w/ Base Question
SBAK	0.6285±0.2205	0.7412±0.2061
Jaccard	0.4033±0.1448	0.5178±0.1429
POS Tag Spread of Added Words		
Nouns+Proper nouns		37.02%
Adpos+Adj+Det		32.68%
Verbs+Adverbs		9.2%
Other POS		21.1%

Fact-Infused Question Generator Network

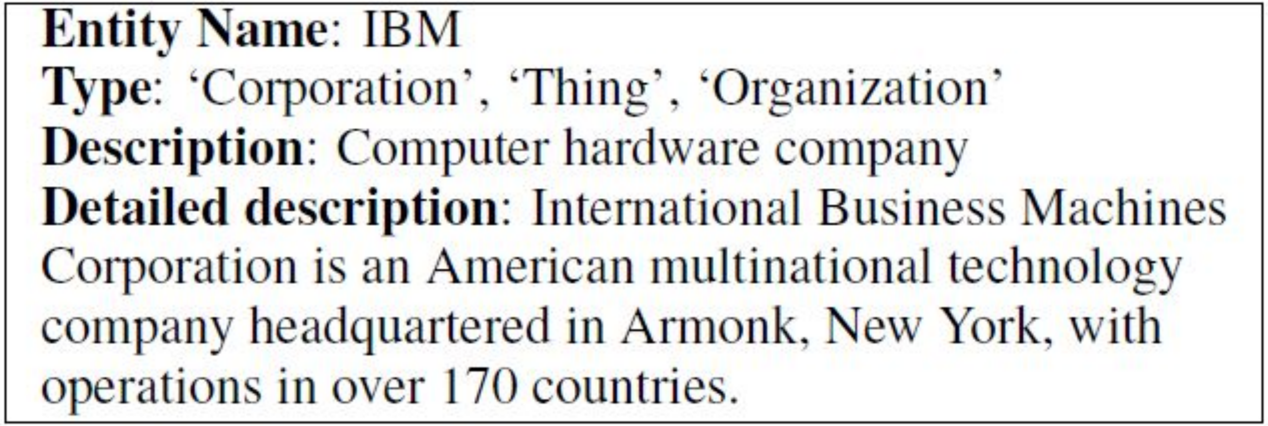

In what year did IBM get its name?
Eⁱ(Q)
Bi-LSTM Base Question Encoder
Base Question Attention
g(t-1)
s(t-1)
Copy
LSTM Decoder for Paraphrased Question
When did American company IBM get its name?
Bi-LSTM Fact Encoder
Fact Attention
m(t-1)
Copy
Eⁱ(F)
International Business Machines Corporation is an American multinational technology company

Standard encoder-decoder approach with attention mechanism

(common in QG models)

- Separate encoders for facts and for original question
- Extended copy-mechanism for incorporating words from both the original question and the fact

Caveat

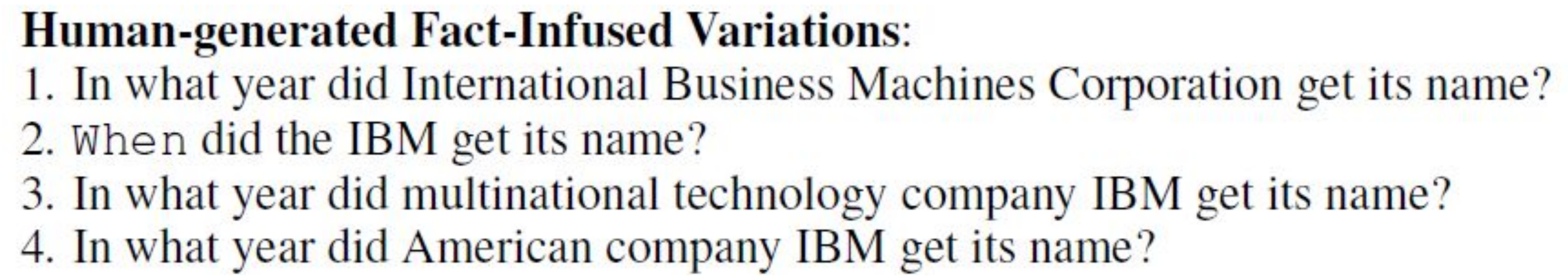


Entity Name: IBM
Type: ‘Corporation’, ‘Thing’, ‘Organization’
Description: Computer hardware company
Detailed description: International Business Machines Corporation is an American multinational technology company headquartered in Armonk, New York, with operations in over 170 countries.

Facts extracted from the entity description from IBM obtained with GES

- *IBM stands for International Business Machines Corporation*
- *IBM is an American company*
- *IBM has headquarters in Armonk, NY*
- *IBM has operations in over 170 countries*

Caveat: Extracting Facts from Entity Descriptions



Human-generated Fact-Infused Variations:

1. In what year did International Business Machines Corporation get its name?
2. When did the IBM get its name?
3. In what year did multinational technology company IBM get its name?
4. In what year did American company IBM get its name?

Originally: X=(question, entity-description), Y=(Set of Rewrites)

Training: X=(question, one-fact) y=Rewrite

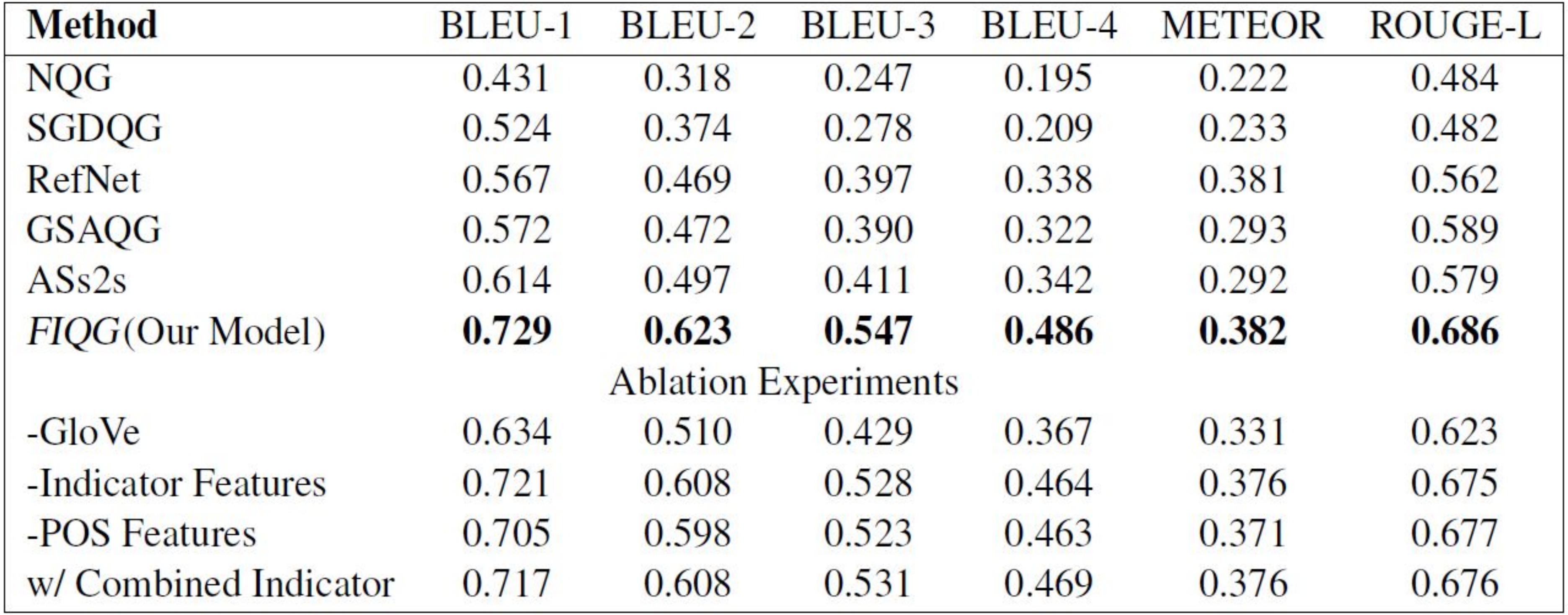
> *Facts extracted using MinIE tool (unsupervised method to extract propositions from descriptions)*

Baselines

QG models (operating on question+fact descriptions as opposed to passages)

1. **NQG** https://github.com/magic282/NQG
2. **RefNet** https://github.com/PrekshaNema25/RefNet-QG
3. **SGDQG** https://github.com/YuxiXie/SG-Deep-Question-Generation
4. **GSAQG** https://github.com/seanie12/neural-question-generation
5. **ASs2s** https://github.com/yanghoonkim/NQG ASs2s

Results



Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
NQG	0.431	0.318	0.247	0.195	0.222	0.484
SGDQG	0.524	0.374	0.278	0.209	0.233	0.482
RefNet	0.567	0.469	0.397	0.338	0.381	0.562
GSAQG	0.572	0.472	0.390	0.322	0.293	0.589
ASs2s	0.614	0.497	0.411	0.342	0.292	0.579
FIQG(Our Model)	**0.729**	**0.623**	**0.547**	**0.486**	**0.382**	**0.686**
Ablation Experiments						
-GloVe	0.634	0.510	0.429	0.367	0.331	0.623
-Indicator Features	0.721	0.608	0.528	0.464	0.376	0.675
-POS Features	0.705	0.598	0.523	0.463	0.371	0.677
w/ Combined Indicator	0.717	0.608	0.531	0.469	0.376	0.676

Observations

❑ Separately representing question and fact results in statistically significant improvements compared to the baselines that use a single encoder (combined question+fact as an input)

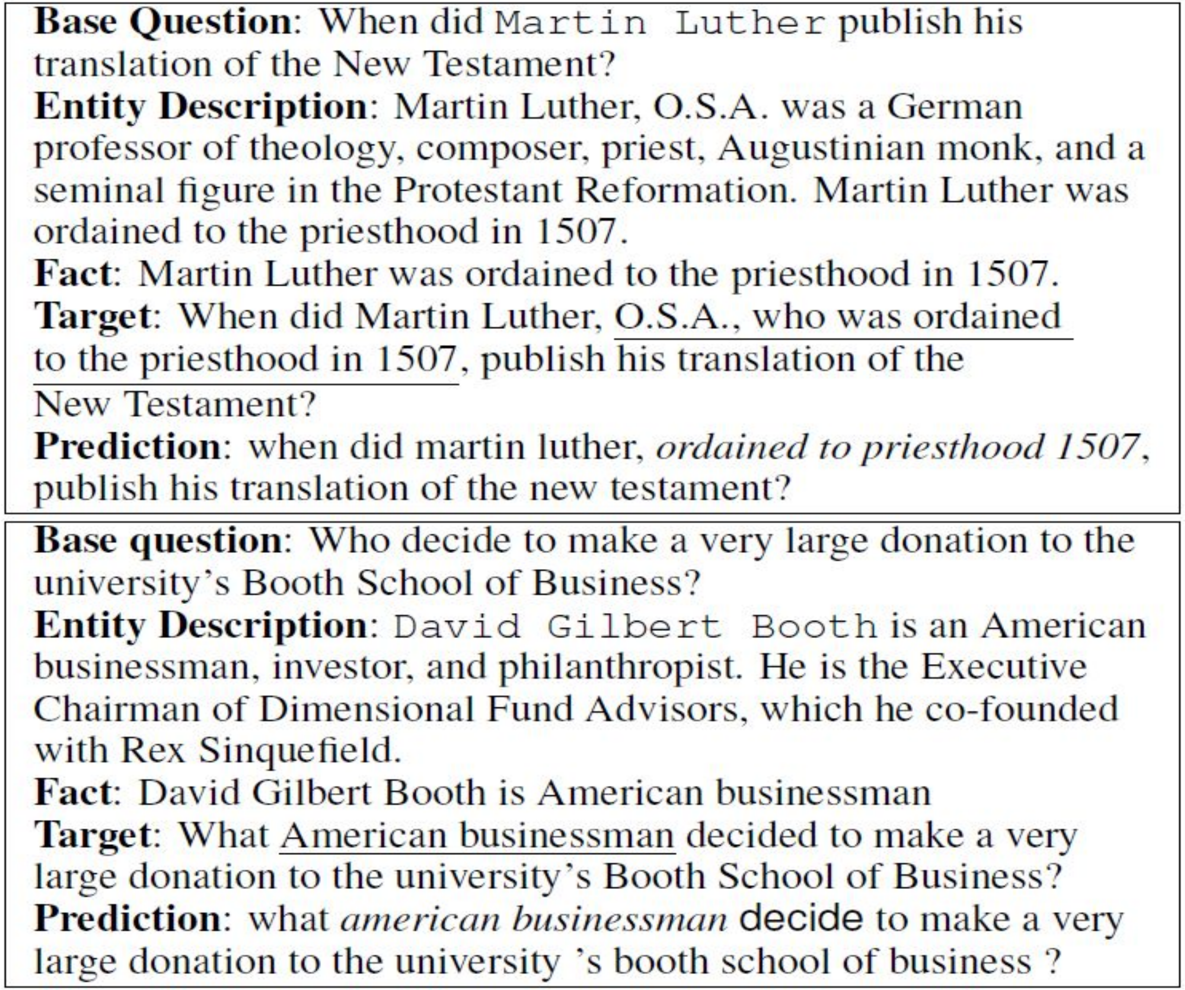
❑ POS and Indicator features: small improvements

❑ Pretrained GLoVE embeddings result in significant performance improvement

Future Possibilities for Learning Models for this dataset:

- Transformers, Reinforcement learning, Variational Autoencoders
- Incorporating the entity-knowledge as a graph and combine with VAE for using the original instances
- Study the original passage and the fact description as two passages to generate *multihop questions* a la HotpotQA

Anecdotes



Base Question: When did Martin Luther publish his translation of the New Testament?
Entity Description: Martin Luther, O.S.A. was a German professor of theology, composer, priest, Augustinian monk, and a seminal figure in the Protestant Reformation. Martin Luther was ordained to the priesthood in 1507.
Fact: Martin Luther was ordained to the priesthood in 1507.
Target: When did Martin Luther, O.S.A., who was ordained to the priesthood in 1507, publish his translation of the New Testament?
Prediction: when did martin luther, *ordained to priesthood 1507*, publish his translation of the new testament?

Base question: Who decide to make a very large donation to the university's Booth School of Business?
Entity Description: David Gilbert Booth is an American businessman, investor, and philanthropist. He is the Executive Chairman of Dimensional Fund Advisors, which he co-founded with Rex Sinquefield.
Fact: David Gilbert Booth is American businessman
Target: What American businessman decided to make a very large donation to the university's Booth School of Business?
Prediction: what *american businessman* decide to make a very large donation to the university 's booth school of business ?

Future Work

- Address the original form of the problem
  - Passage + Entity-description having multiple facts => Fact-infused Questions
- How can we use this dataset to:
  - Train adversarial QA models
  - Educational RC application

    -- don’t want to always generate the same question for testing students, grade based on difficulty-level

References

- Supplementary material, code, data at https://github.com/NUS-IDS/ranlp21-fiqv
- Google Knowledge Graph API https://developers.google.com/knowledge-graph
- The original SQuAD dataset https://rajpurkar.github.io/SQuAD-explorer/
- HotpotQA https://hotpotqa.github.io/
- MinIE: Minimizing Facts in Open Information Extraction https://github.com/uma-pi1/minie

Thank you! Any Questions?

Contact: *idssdg@nus.edu.sg*