UNDERSTANDING QUESTION TYPES FOR QUESTION ANSWERING

MORRISON ACADEMY
VINCENT CHENG
Contact: vincentcheng236@gmail.com

Introduction

NLP is a branch of AI that deals with designing systems that can understand human language. Extractive question answering is one of the most popular and difficult tasks in NLP right now. It is the task where, given a question and context, a model needs to output a coherent answer. Applications of question answering are endless

- Chatbots for customer service
- Fast document readers for students, lawyers, historians
- Enhanced search engines like Google/Yahoo

Developing strong question answering models is imperative for the development of systems with higher natural language understanding (NLU). In this project, we focus on error analysis of these models on different question types.

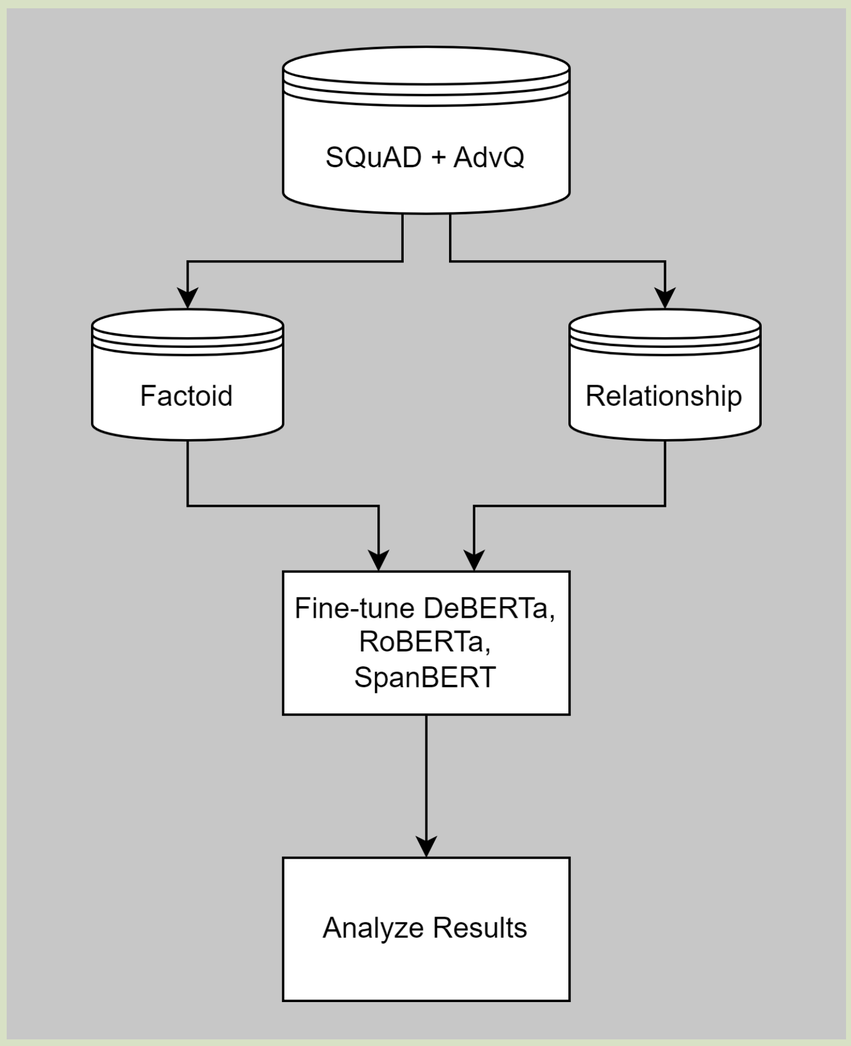
Factoid:
Information retrieval questions usually starting with one of 5 W's.
Example:
What county was Abraham Lincoln born in?

Relationship:
Questions that ask about the relation between two entities.
Example:
How did the 2001 IPCC report compare to reality for 2001-2006?

Objectives:

- **Do different question types affect the performance of QA models?**
- **Do different model architectures perform differently?**
- **What are the reasons for the possible deficiencies and how can we fix them?**

Methodology



Results

	Factoid		Relationship	
	EM	F1	EM	F1
DeBERTa	85.00	89.04	**62.63**	**72.18**
RoBERTa	84.78	88.50	61.62	69.07
SpanBERT	**88.22**	**92.00**	57.58	68.26

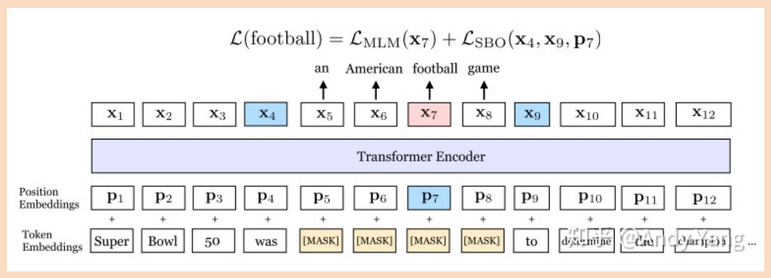
Samples Responses:

Question: Who was the main performer at this year's halftime show?
Context: ... The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars...
Correct answer: Coldplay
DeBERTa: Beyoncé and Bruno Mars
RoBERTa: Beyoncé and Bruno Mars
SpanBERT: Coldplay

Question: Which of the following words is least related to the others: hyem, home, barn or hjem?
Context: "Bairn" and "hyem", meaning "child" and "home", respectively, are examples of Geordie words with origins in Scandinavia; barn and hjem are the corresponding modern Norwegian and Danish words...
Correct answer: barn
DeBERTa: barn
RoBERTa: barn
SpanBERT: Bairn

Discussion

- From our results, **we can clearly see that different question types do indeed affect the performance of our models.** All models achieve better EM and F1 scores for factoid questions rather than relationship questions.
- A possible reason for this score discrepancy is that **relationship questions are fundamentally harder than factoids**. Relationship questions require the models to understand the relation between two entities, hence extracting two pieces of information, while factoid questions mostly require only one.
- We also provide an interesting analysis on the results of SpanBERT as it achieves the best results in factoid questions yet the worst results in relationship questions. SpanBERT is trained differently from other models as during training it masks spans of words. This may cause it to have limited ability to answer questions whose answers appear far apart, hence the low scores for relationship questions. **SpanBERT excels when answers are based on short spans but struggles when questions require connections from different parts of the context.**
- **We see that DeBERTa and RoBERTa behave similarly for both relationship questions and factoids** with DeBERTa achieving slightly higher scores for each category. SpanBERT's performance is more drastic which we provide a specific analysis for below.
- **The major deficiency we find is the underperformance for relationship questions**. We speculate that the reason for this is these models' inability to make connections between separate entities in the text. A potential solution is to add a classified in front of the model and generating further queries when the question is a relationship one. This relates to two-hop QA and can be a topic of future research.



https://arxiv.org/pdf/1907.10529.pdf

Conclusion

- In this project we develop an analysis of question answering models DeBERTa, RoBERTa, SpanBERT through question classifications. We draw potential reasons for differences in model performance with a brief case study on SpanBERT.
- In the future, we wish to explore different, more specific, types of questions by finding more diverse datasets. We also hope to explore how these results can improve new models and datasets for more durable downstream applications.