GENERATING QUESTIONS FROM TEXTUAL DATA

by

Raynard Dodzi Helegah

Outline

- Introduction
- What is question generation?
- The traditional approach
- **The multimodal approach**

Keywords

- GPT-2
- Model,
- Seq2Seq model,
- WordNet

Introduction

Why question generation

- Simulate retention of prior knowledge
- Guide for reading/learning

Introduction

Manual question setting is

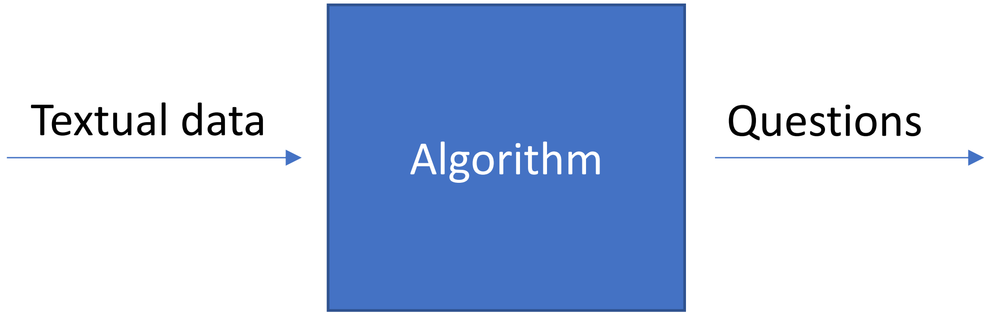
- Time consuming
- Not scalable   etc.

Introduction

How will they generate self assessment tests based on their textbooks?

What is Question Generation

- An automated process of creating questions from a textual data.



Textual data

Algorithm

Questions

Methods of Generating Questions

- Tradition method
- Using Sequence-to-sequence models

Traditional Approach

How

- Based on templates

  e.g., replace nouns with ‘what’, names with ‘who’ and rearrange

Problems

- Sometimes produces grammatically incorrect questions
- Monotonous – one style questions

Model Approach

How

- Using NLP
- Using advanced pretrained NLP models

Advantages

- Variety in question style
- On par with questions generated by humans 🤔

Types of Question

- True or False (T/F) Questions
- Multiple Choice Question (MCQ)

Generating T/F Questions

True questions

- Summarise the text using either *extractive* or *abstractive* technique.
- Each sentence in the summary is a potential question whose answer is true.

True Questions

Example

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. As the problem of information overload has grown, and as the quantity of data has increased, so has interest in automatic summarization. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. An example of the use of summarization technology is search engines such as Google. Document summarization is another.

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document.

Generating T/F Questions

False Question

- Add or remove negation
- Changing noun/verb phrases
- Changing name entities

Generating T/F Questions

False Question

- Adding or removing negation
    - E.g., Birds can fly -> Birds cannot fly.
- **Replacing the ending of the sentence**

Replacing the ending of the sentence

The idea

- Remove the ending of the sentence
- Auto complete the sentence using ***GPT-2***
- Check for similarity with the original sentence
- Pick the dissimilar one

How do you break the sentence?

Idea

  - Remove the last verb phrase or noun phrase

How

- We need a parser to break the sentence into it’s contituents
- Use **Berkley Constituency Parser**

Berkley constituency Parser

Original

The man was sitting ~~under the tree.~~

Verb phrase

Noun phrase

Goes to GPT

The man was sitting

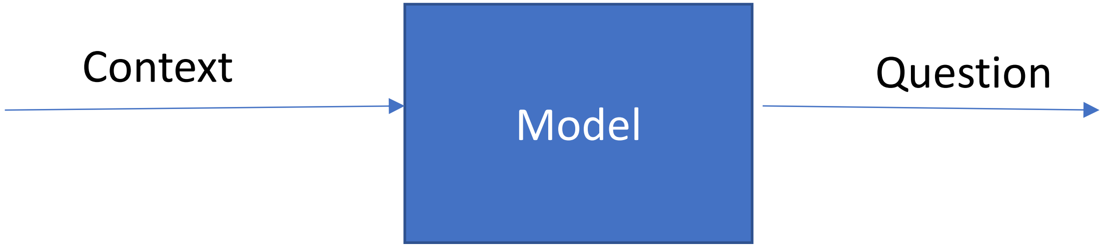
GPT gives

- in the leaving room
- beside the plant
- in the park

Generating MCQ

Idea

- Train a model that takes a statement (context) as an input
- Outputs a question



How: Models need data

Using the Standford Question Answering Dataset **(SQuaD)**

Context	Question	Answer	Answer Index
Photosynthesis is **the process by which plants use sunlight, water, and carbon dioxide to create oxygen and energy in the form of sugar**.	What is the name of processes by which plants produce their own food?	Photosynthesis	0

Model Design

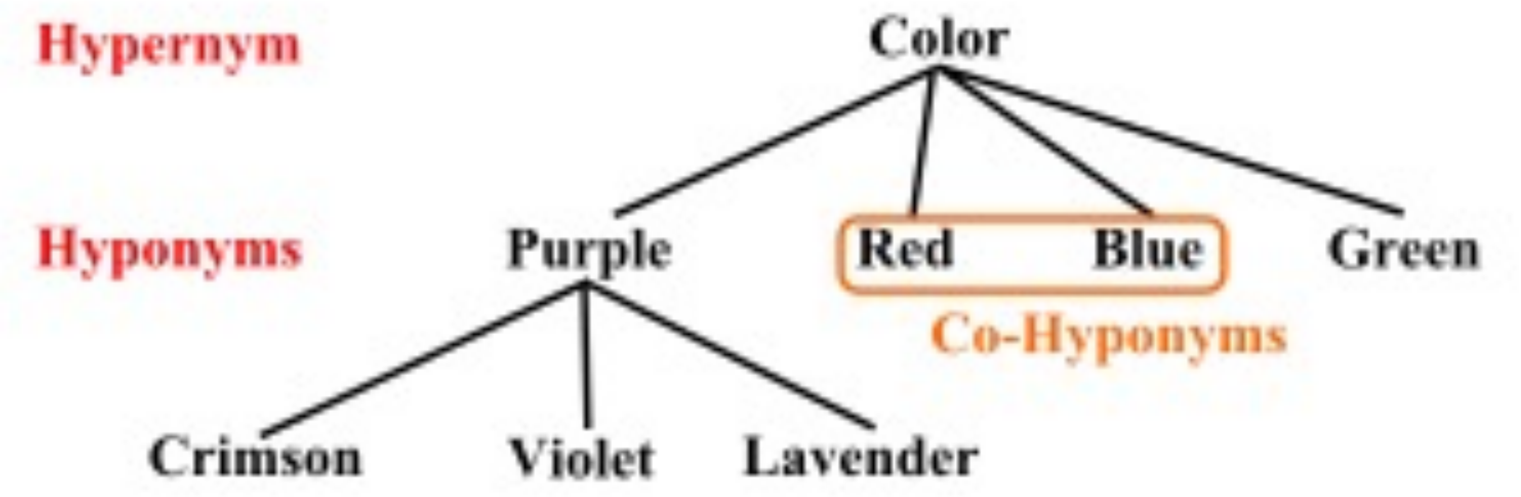
Train a *Seq2Seq* model using SQuaD with Question as our target feature

Target feature

Context	Question	Answer	Answer Index
Photosynthesis is **the process by which green plants use sunlight, water, and carbon dioxide to create oxygen and energy in the form of sugar**.	What is the name of processes by which plants produce their own food?	Photosynthesis	0

How do we get the choices?

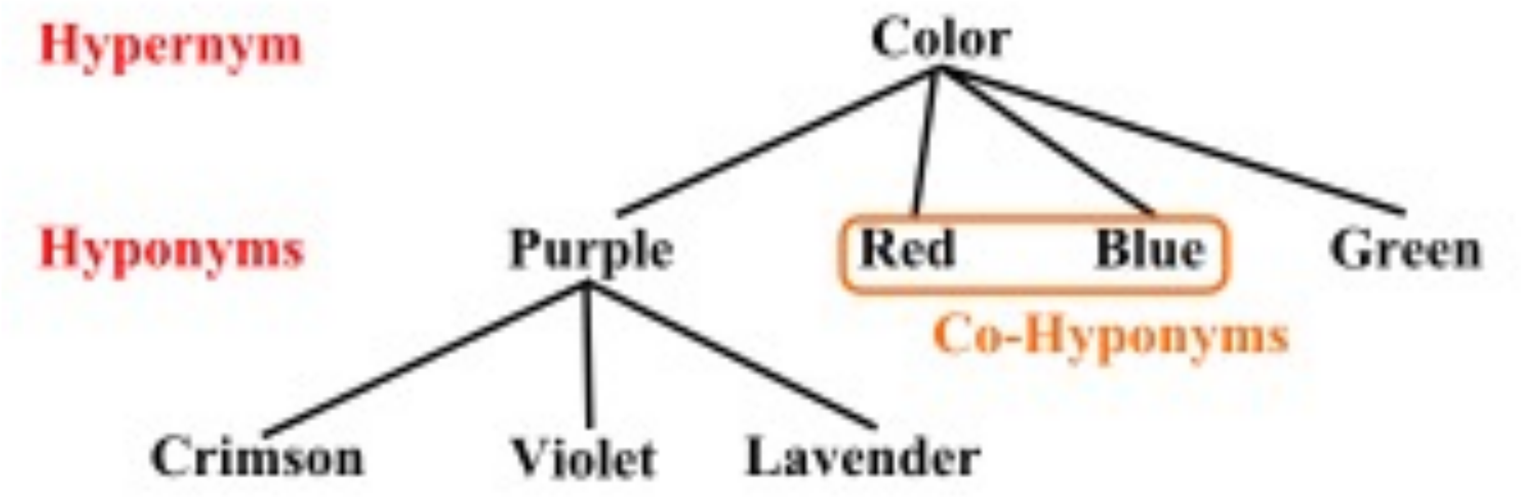
- By generating distractors (wrong answers) i.e., words that are similar to the correct answer using ***wordnet.***



Source: https://en.wikipedia.org/wiki/Hyponymy_and_hypernymy

How do we get the choices?

- By generating distractors (wrong answers) i.e., words that are similar to the correct answer using ***wordnet.***



Source: https://en.wikipedia.org/wiki/Hyponymy_and_hypernymy

Main Take-Aways

- T/F Questions are generated with the help of GPT and BERT pretrained models
- MCQs are generated by a model trained on the SQuaD dataset.
- The distractors (wrong choices) are generated using wordnet.

DEMO

Show me the code!

Contact

LinkedIn / Gmail / GitHub

dodziraynard

Issues to address

- Getting distractors for phrases and proper nouns