Team 16
Project - Question Answering
Team Members :- Rahul Padhy (2022201003)
Arun Das (2022201021)

Introduction

Question and Answering (henceforth referred to as QnA) is a computer science discipline within the fields of information retrieval and natural language processing, which focuses on building systems that automatically answer questions posed by humans in a natural language. QnA systems are now found in search engines and phone conversational interfaces, and they're fairly good at answering simple snippets of information. On harder questions, however, these normally only go as far as returning a list of snippets that the users must then browse through to find the answer to our question. QnA models are often used to automate the response to frequently asked questions by using a knowledge base (e.g. documents) as context.

QnA systems differ in the way answers are created :-

- Extractive QnA - The model extracts the answer from a context and provides it directly to the user. It is usually solved with BERT-like models.
- Generative QnA - The model generates free text directly based on the context. It leverages text generation models.

QnA systems also differ in where answers are taken from:-

- Open QnA - The answer is taken from a context.
- Closed QnA - No context is provided and the answer is completely generated by a model.

The current project aims to build a **Generative, Open QnA model**. For achieving this task, the following 3 datasets are being explored :-

- **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD)
- WikiQA Dataset
- NewsQA Dataset

SQuAD

The **SQuAD** is a popular benchmark for evaluating the performance of question answering systems. It consists of over 100,000 question-answer pairs, based on over 500 Wikipedia articles, which are used for training and evaluation of machine learning models. The answer to each question is a segment of text or span from the corresponding reading passage. Sometimes the questions might be unanswerable. SQuAD has 2 versions. SQuAD 1.1, contains 100,000+ question-answer pairs on 500+ articles. SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.

Training on SQuAD2.0 is tougher as you not only have to answer the questions, but also abstain from answering

when the question is unanswerable. So extra effort is needed to determine if the question is unanswerable.
The performance of ML models on the **SQuAD** dataset is typically evaluated using the following two metrics :- Exact Match Score and F1 Score.

The **Exact Match (EM) score** metric measures the percentage of questions for which the model produces an exact match with the correct answer. A predicted answer is considered correct if it exactly matches the gold standard answer in the dataset. The EM score ranges from 0 to 100, with higher scores indicating better performance.

The **F1 score** metric measures the average overlap between the predicted answer and the correct answer, based on the harmonic mean of precision and recall. Precision is the ratio of the number of correct answers to the total number of predicted answers, and recall is the ratio of the number of correct answers to the total number of gold standard answers. The F1 score ranges from 0 to 100, with higher scores indicating better performance.

Both **EM** and **F1** scores are commonly used to evaluate the performance of machine learning models on the SQuAD dataset. In general, higher scores on both metrics indicate better performance, and a model with high scores on both metrics is considered to be a strong performer. However, the choice of metric may depend on the specific task and application, and other metrics such as precision, recall, and accuracy may also be used to evaluate performance in different contexts.

WikiQA Dataset

The WikiQA dataset is a popular benchmark for evaluating the performance of question answering systems. It consists of a set of questions and answers, based on a subset of the English Wikipedia corpus. The dataset was released by Microsoft Research in 2015 and contains approximately 3,000 questions, each of which has 5 candidate answers. The questions cover a variety of topics and are phrased as natural language questions. The dataset is designed to evaluate systems that can retrieve accurate answers to factual questions from large collections of text. The candidate answers for each question are derived from the text in the Wikipedia articles, and the correct answer is selected from among the candidate answers by human annotators. One advantage of the WikiQA dataset is that it provides a challenging evaluation task for question answering systems, as the questions are often ambiguous and require a deep understanding of the context in order to provide accurate answers. However, the small size of the dataset is also a limitation, as it may not provide a comprehensive evaluation of the performance of question answering systems on a wide range of real-world questions.

The performance of machine learning models on the WikiQA dataset is typically evaluated using the following metrics :- **Mean Average Precision (MAP)**, **Mean Reciprocal Rank (MRR)** and **Precision at N (P@N)**.

The **Mean Average Precision (MAP)** metric measures the average precision of the system across all the questions in the dataset. For each question, the system

returns a ranked list of answers, and the average precision of the system is calculated by considering the relevance of each answer in the list. MAP ranges from 0 to 1, with higher scores indicating better performance.

The **Mean Reciprocal Rank (MRR)** metric measures the average of the reciprocal ranks of the correct answers. The reciprocal rank of an answer is the inverse of its rank in the list of answers returned by the system. MRR ranges from 0 to 1, with higher scores indicating better performance.

The **Precision at N (P@N)** metric measures the proportion of correct answers among the top N answers returned by the system. P@N is typically reported for different values of N, such as 1, 3, and 5. Higher values of P@N indicate better performance.

These metrics are commonly used to evaluate the performance of machine learning models on the WikiQA dataset. In general, higher scores on these metrics indicate better performance, and a model with high scores on these metrics is considered to be a strong performer. However, the choice of metric may depend on the specific task and application, and other metrics such as recall, F1 score, and accuracy may also be used to evaluate performance in different contexts.

NewsQA Dataset

The NewsQA dataset is a large-scale reading comprehension dataset that was created to facilitate research in natural language understanding and question-answering. It was released by the University of Washington and the Allen Institute for Artificial Intelligence in 2017. The dataset contains over 100,000 question-answer pairs based on over 14,000 news articles from CNN and the Daily Mail. The questions were written by human annotators who read the articles and created questions that required understanding of the article's content.

The questions in the NewsQA dataset are diverse and cover a wide range of topics, including politics, entertainment, and sports. The answers are often complex and require reasoning and inference skills to arrive at the correct answer.

The NewsQA dataset is useful for training and evaluating natural language understanding models, particularly those focused on reading comprehension and question-answering. It has been used in various research projects and competitions, including the Stanford Question Answering Dataset (SQuAD) and the Machine Comprehension Test (MCTest).

The most commonly used evaluation measure for the NewsQA dataset is the F1 score, which is a measure of the overlap between the predicted answer and the ground truth answer. To compute the F1 score, we first calculate the precision and recall. Precision is the fraction of the predicted answer that is correct, while

recall is the fraction of the ground truth answer that is correctly predicted. The F1 score is the harmonic mean of precision and recall, and is computed as follows :-

F1 = 2 * (precision * recall) / (precision + recall)

In the context of the NewsQA dataset, precision is the fraction of predicted answers that exactly match the ground truth answer, while recall is the fraction of ground truth answers that are correctly predicted.

The F1 score is commonly used to evaluate the performance of models on the NewsQA dataset, as it provides a single summary measure of a model's ability to accurately answer questions. However, other evaluation measures, such as accuracy and mean average precision (MAP), can also be used depending on the specific research question or application.

Estimated Workflow of Code

(1) **Question Analysis** - done through Named-Entity Recognition or Surface Text Pattern Learning or Syntactic Parsing or Semantic Role labeling.
(2) **Question Classification -** classification of the input question into a set of question types and mapping question types to the available named-entity labels.
(3) **Query Construction** - formulating a query with a high chance of retrieving relevant documents by assigning a higher weight to the question’s target

and using query expansion techniques to expand the query.

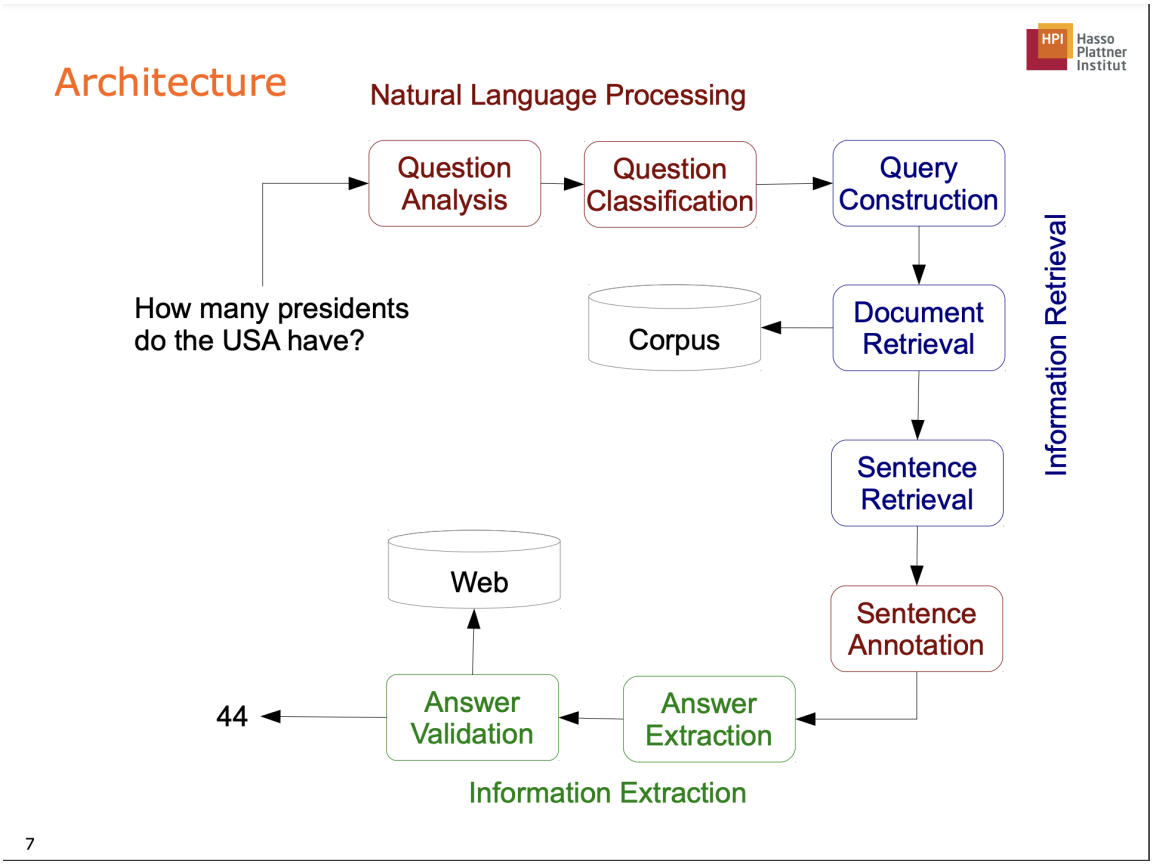
(4) **Document Retrieval** - reducing the search space for the subsequent components and retrieval relevant documents from a large corpus.

(5) **Sentence Retrieval** - finding small segments of text that contain the answer.

(6) **Sentence Annotation** - annotation of relevant sentences using linguistic analysis (such Named-entity recognition, Syntactic parsing, Semantic role labeling).

(7) **Answer Extraction** - extraction of candidate answers based on various information such as patterns, syntactic parser, semantic roles and question type.

(8) **Answer Validation** - using various measures for relevant datasets mentioned above.



[Image sourced from -> https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/NaturalLanguageProcessing/NLP2017/NLP8_QuestionAnswering.pdf]

Estimated Timelines for Completion of Various Steps

Complete exploration of all the datasets mentioned above and selection of relevant accuracy measures - by **8th March**.

Exploration of embedding models (TFIDF, word2vec, Infersent from Facebook) for better modeling of data - by **25th March**.

Exploration of various models (BERT, HuggingFace Transformer Models) - by **15th April**.

Explainability of results obtained and how it compares to the various State-of-The-Art (SOTA) models out there and also how the solution can be extended / scaled - by **20th April**.