Assignment 2:
Information retrieval and question answering system

a1909709 Shan Shan Bianca Tan

April 16, 2024

The University of Adelaide
4433_COMP_SCI_7417 Applied Natural Language Processing
Lecturer: Dr. Alfred Krzywicki

Contents

1 Introduction

Question answering (QA) dialog systems play a crucial role in enhancing the accessibility and utility of news articles for readers. With the high volume of content produced daily, it can be difficult for readers to find relevant information efficiently. QA systems allows readers to navigate lengthy articles and search for answers quickly.

Overall, question answering dialog systems play a vital role in bridging the gap between news articles and readers, offering efficient, engaging, and personalized access to information in the constantly changing digital media landscape.

In my assignment, which is the individual task of building an information retrieval system, I cover 2 different methods. One of which is using an existing pre-trained model while the other method is a hybrid model which combines existing models and functions I built to extract the infomation. I will dive deeper into both of this models below.

Some limitations of my hybrid model includes long run time and high complexity yet only being able to answer simple questions well. It produces lower accuracy for more complex questions. While my pre-trained model produced better results, it is limited by its fixed architecture and lack of domain knowledge.

2 Preprocessing

Data preprocessing is a crucial step in the machine learning pipeline as it ensures high data quality and data interpretability by the model. It lays an important foundation of building an accurate and scalable machine learning model.

2.1 Data cleaning

In my project, I carried out data cleaning such as handling incorrect punctuation, lemmatization and lower casing. This allows for my data to be more interpretable by the model.

In the imported dataset, it is noticed that there were various question marks (?) in random locations, this would lead to incorrect splitting of sentences if not correctly handled, hence, I replaced them with space.

I also did lemmatization to ease the extraction of information from the articles. Lemmatization is an important preprocessing step in information extraction models because it helps standardize words to their base form, known as a lemma. (Pykes 2023) It ensures that different inflected forms of a word are treated as the same entity, thereby improving the accuracy of information retrieval and analysis.

2.2 Tokenization

Tokenization is a fundamental preprocessing step in NLP that involves breaking down raw text into smaller units called tokens. (Awan 2023) During the lemmatization and

coreference resolution, the text was tokenized into individual words or tokens before it got processed.

Tokenization enables machines to understand and process the human language by breaking down text into individual words, phrases or punctuations. This parsing process provides a structured representation of text that can be analyzed, manipulated, and interpreted by algorithms and models. (Awan 2023)

3 System Architecture

As mentioned, there are 2 methods used, I will be referring to them as the hybrid model and pre-trained model. The same data cleaning and preprocessing is done for both the models.

3.1 Hybrid Model

For the hybrid model, Figure 1 shows the system architecture.

Figure 1: Hybrid Model System Architecture

In this model, we had functions to execute coreference resolution, text matching utility and extraction of answers from the most relevant sentence. After that, I evaluate the model using an exact match score.

A pre-trained model (from deepset/tinyroberta-squad2) is incorporated at the step of extracting entities from the sentences after finding the most relevant sentence. This is to ensure that extraction of relevant answers from the sentence is accurate.

3.2 Pre-trained Model

Figure 2 shows the system architecture of the pre-trained model.

Figure 2: Pre-trained System Architecture

I am using the model "deepset/tinyroberta-squad2" for the question and answering pipeline. I chose this model as it is computationally efficient while being able to produce accurate results. (Chan et al. 2024) The RoBERTa model shares the BERT model's architecture with slight modifications to the key hyper-parameters and tiny embedding tweaks. (Sharma 2022) I also evaluated the model using an exact match score.

4 Model Selection and Training

I have chosen to use NLP models throughout the assignment due to the nature of the task. As it is essential for the model to understand the relationships between words, the utilisation of models that are designed to understand and process natural language text was crucial. The choice of using "deepset/tinyroberta-squad2" is due to its computational efficiency and accuracy.

For the final evaluation, I chose to use MRR and MAP as the evaluation metrics.These metrics are commonly used in information retrieval tasks where the relevance of retrieved answers is crucial.

Throughout the developmental process, different models and methods were tested to measure their accuracy and efficiency. The 2 models that are presented were chosen based on its accuracy as well as ensuring that the results were not gibberish.

I chose my final model based on the computational efficiency of the models as well as the exact match scores of the top ranked answer. Exact match is a strict evaluation metric that only gives two scores, 1 if the answer provided is exactly the same as the predicted answer, else it is 0. (Shakkari 2022) In my exact match function, I allowed for lower casing of both answers during comparison as all answers produced by the models are in lowercase due to data preprocessing done.

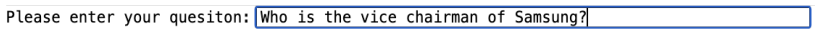
The hybrid model achieved a score of 0.17, while the pre-trained model attained a score of 0.5. The pre-trained model outperformed in predicting exact answers and demonstrated superior computational efficiency, hence it was chosen as the final model.

5 User interaction with the system

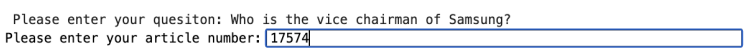
The user is able to interact with the system by entering their questions and articled ID. The program will then be executed and provide the top 5 answers as well as confidence scores.

Here is an example of the process and output:

Step 1: Input the question



Step 2: Input the article ID



Step 3: Top 5 answers are generated

```
Please enter your quesiton: Who is the vice chairman of Samsung?
Please enter your article number: 17574
Question: Who is the vice chairman of Samsung?
Expected answer:      No expected answers

Top 5 Answers and Confidence:
1: jay y. lee 0.9688
--------------------------------
Other Answers:
2: jay y. lee , 0.024
3: the de facto leader , jay y. lee 0.002
4: lee 0.0009
5: , jay y. lee 0.0007
```

6 System evaluation

As discussed, I used an exact match score to determine which model to use. After that I use Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) score to evaluate the selected model's performance taking into consideration the ranks of the top 5 answers.

Both MRR and MAP evaluates the quality of a ranked list of answers that was provided by the model. MRR measures how quickly the correct answer is found while MAP measures the considers the number of relevant answers and their position in the list. (Wang 2021)

The chosen model (pre-trained model) produced an MRR of 0.59 which indicates that correct answer is found within the top-ranked answers approximately 59% of the time. It had an MAP of 0.6 which indicates that, on average, the model retrieves relevant answers and ranks them appropriately approximately 60% of the time across all queries.(Wang 2021)

6.1 Print results of 10 questions

Here are the printed results of 10 question and answer pairs from the information extraction system:

```
Question: Who is the vice chairman of Samsung?
Expected answer:      Jay Y. Lee

Top 5 Answers and Confidence:
1: jay y. lee 0.9688
--------------------------------
Other Answers:
2: jay y. lee , 0.024
3: the de facto leader , jay y. lee 0.002
4: lee 0.0009
5: , jay y. lee 0.0007

Question: Which subway is opening in New York City on Sunday?
Expected answer:      Second Avenue subway

Top 5 Answers and Confidence:
1: second avenue subway 0.4899
--------------------------------
Other Answers:
2: second avenue 0.2715
3: the second avenue subway 0.2158
4: second avenue 0.1626
5: the second avenue 0.0716

Question: What amount did Fox News offer?
Expected answer:      20 Million

Top 5 Answers and Confidence:
1: $ 20 million 0.8697
--------------------------------
Other Answers:
2: $ 20 million 0.8697
3: 20 million 0.1175
4: 20 million 0.1149
5: $ 20 million offer 0.0071

Question: Who is Mr. Roof's lead lawyer?
Expected answer:      David I. Bruck

Top 5 Answers and Confidence:
1: david i. bruck 0.9914
--------------------------------
Other Answers:
2: david i. bruck 0.9823
3: david i. bruck , 0.0132
4: bruck 0.0125
5: standby counsel 0.0112

Question: How many students attend the Evergrande Football School?
Expected answer:      2,800 students

Top 5 Answers and Confidence:
1: 800 0.4558
--------------------------------
Other Answers:
2: 2 , 800 0.4517
3: 800 student 0.036
4: 2 , 800 student 0.0357
5: its 2 , 800 0.0046

Question: What is the main newspaper of the Communist Party in China?
Expected answer:      People's Daily

Top 5 Answers and Confidence:
1: people daily 0.97
--------------------------------
Other Answers:
2: people daily 0.9642
3: people daily , 0.0284
4: daily 0.0129
5: people daily , 0.0081

Question: Who is the spokesman?
Expected answer:      Numan Kurtulmus

Top 5 Answers and Confidence:
1: numan kurtulmus 0.9853
--------------------------------
Other Answers:
2: aiymkan kulukeyeva 0.5567
3: aiymkan kulukeyeva , 0.02
4: kulukeyeva 0.0102
5: aiymkan kulukeyeva 0.0075

Question: Where is the gunman from?
Expected answer:      Kyrgyzstan or elsewhere in Central Asia

Top 5 Answers and Confidence:
1: kyrgyzstan 0.5884
--------------------------------
Other Answers:
2: kyrgyzstan 0.5789
3: kyrgyzstan or elsewhere in central asia 0.2405
4: kyrgyzstan or elsewhere in central asia 0.2129
5: kyrgyzstan or elsewhere in central asia . 0.0114

Question: Where is Megyn Kelly moving to from Fox News?
Expected answer:      NBC

Top 5 Answers and Confidence:
1: nbc 0.9927
--------------------------------
Other Answers:
2: nbc 0.0897
3: nbc 0.0461
4: nbc      0.0024
5: nbc , 0.0017

Question: What salary was Megyn Kelly offered by the Murdoch family?
Expected answer:      More than $20 million a year

Top 5 Answers and Confidence:
1: $ 20 million a year 0.4147
--------------------------------
Other Answers:
2: more than $ 20 million a year 0.3987
3: 20 million a year 0.0998
4: 12 year 0.0371
5: $ 20 million 0.0078
```

Figure 3: Results of 10 questions

7 Conclusion

Overall, the pre-trained model produced satisfactory results for the information extraction task.

However, during the development process, several challenges were encountered. While building the hybrid model, both coreference resolution and the extraction of relevant answers from sentence codes were not accurate in producing optimal outcomes. This challenge came together with the difficulty of accessing a pre-trained model "neural-coref" for coreference resolution. Although I retained the existing coreference resolution function, I found its efficiency and accuracy to be unsatisfactory. With more time, I aim to refine this function to achieve better performance.

There are areas for potential improvement in the hybrid model, such as developing a function to accurately extract the most relevant phrase from a sentence instead of relying solely on a pre-trained model. Additionally, I aspire to build a more precise and efficient approach to coreference resolution. Regarding the pre-trained model, I intend to fine-tune it further to enhance its performance.

This assignment provided valuable insights into NLP algorithms for me. While I regret not achieving a more accurate model, I consider it a learning opportunity to improve my coding skills and deepen my understanding of NLP.

8 References

References

Awan, Abid Ali (Sept. 2023). *What is Tokenization?* URL: https://www.datacamp.com/blog/what-is-tokenization.

Chan, Branden et al. (Mar. 2024). *deepset/tinyroberta-squad2*. https://huggingface.co/deepset/tinyroberta-squad2.

Pykes, Kurtis (Feb. 2023). *Stemming and Lemmatization in Python*. URL: https://www.datacamp.com/tutorial/stemming-lemmatization-python.

Shakkari, Kaushik (July 2022). *Understanding Semantic Search — (Part 4: Answer Quality Metrics - Evaluating the Reader Models for Machine Reading Comprehension Task)*. URL: https://kaushikshakkari.medium.com/open-domain-question-answering-series-part-4-answer-quality-metrics-evaluating-the-reader-ff7fa20736bf#:~:text=Exact%20Match%20is%20a%20strict,United%20States%E2%80%9D%20is%200)..

Sharma, Drishti (Nov. 2022). *A Gentle Introduction to RoBERTa*. URL: https://www.analyticsvidhya.com/blog/2022/10/a-gentle-introduction-to-roberta/.

Wang, Benjamin (Jan. 2021). *Ranking Evaluation Metrics for Recommender Systems*. URL: https://towardsdatascience.com/ranking-evaluation-metrics-for-recommender-systems-263d0a66ef54.