Information Retrieval and Question Answering

Name and student ID:
KWONG HO LI (A1904570)

The University of Adelaide
4333_COMP_SCI_7417 Applied Natural Language Processing
Lecturer: Dr. Alfred Krzywicki

Table of Contents

1. Abstract

This report explores the development and evaluation of an Information Retrieval based Question Answering (IR-QA) system, focusing on leveraging Natural Language Processing (NLP) models to enhance answer accuracy and efficiency. Utilizing the SQuAD Dev set 2.0 for testing, the system integrated rule-based methods and the BERT model (bert-large-uncased-whole-word-masking-finetuned-squad) to process and respond to queries. Through comprehensive testing, including validation, the "en_core_web_md" model was identified as the most optimal for balancing performance and computational demands. The project revealed the superior capability of BERT in handling complex queries, highlighted the rule-based method's efficiency with straightforward questions. Key findings indicate the necessity of model selection tailored to query complexity and the importance of contextual understanding in Question Answering (QA) systems. Future work will aim to enhance the system's adaptability to diverse queries and improve its computational efficiency. This research contributes to the field by detailing the iterative process of model evaluation and integration, offering insights into the design and implementation of effective QA systems.

2. Introduction

2.1. Information Retrieval based Question Answering

IR-QA systems are pivotal in deriving accurate answers from a curated dataset, mitigating the common issue of 'hallucinations' or unreliable responses associated with Large Language Models (LLMs). By anchoring responses to existing data, these systems ensure reliability and relevance. However, rule-based QA can become intricate, as designing exhaustive question-answer mappings is challenging, potentially compromising answer accuracy. Integrating IR-QA with LLMs, utilizing external knowledge bases, enhances answer accuracy and contextuality, reducing hallucinations (Ke et al. 2024).

This system is particularly beneficial for businesses requiring a robust frequently asked question (FAQ) and simple chatbot functionality, as it adeptly retrieves and responds to user queries with the most relevant information. However, its effectiveness diminishes with complex inquiries due to its limited understanding of word relationships and inability to capture nuanced interactions within the text.

2.2. Limitations of the system

The limitations of this system are particularly evident in its rule-based questioning component, which is designed to address 'Wh' questions like Who, Where, When, What, and How many. This design specificity means it falls short when questions deviate from these starters or when the underlying text lacks identifiable named entities aligned with the question's focus, often resulting in

a 'NO ANSWER' outcome. This system struggles significantly with questions seeking explanations (such as 'Why'), as it lacks the mechanism to deduce underlying reasons or motivations from the text.

Furthermore, questions probing relational dynamics or abstract concepts pose a challenge due to the system's rudimentary interpretive capabilities. On the BERT side, while offering a more nuanced understanding of context, the model faces its bottleneck with the 512-token limit, which restricts its operational scope, especially for lengthy texts.

Additionally, the information retrieval process, heavily reliant on TF-IDF and cosine similarity. It demands exact lexical matches, which can overlook the nuances of synonym usage, thereby affecting the system's ability to capture and respond to the semantic essence of queries.

2.3. Overview of the system workflow

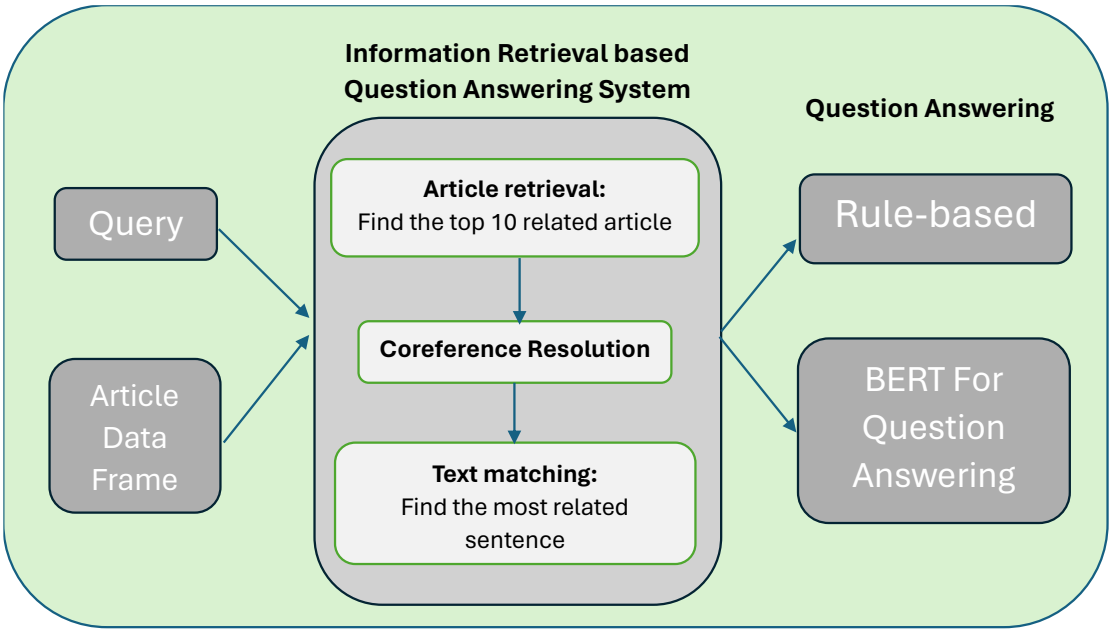

Figure 1. The graph of IR-QA workflow

According to Figure 1, task execution involves preparing the query and the article database, with the system designed to identify the top 10 related articles for any given query than proceed with Coreference Resolution. The most pertinent sentence is then extracted from these articles to answer the query.

The answer extraction method varies: the rule-based approach uses predefined criteria to generate responses, while the BERT-based approach employs deep learning to deduce answers from the context and query, demonstrating a flexible and user-adaptive answering strategy.

2.4. Testing dataset

Self-defined questions:
20 self-defined questions were created for model selection, formulated from randomly selected articles, to evaluate the system's response accuracy, especially for 'Wh' questions excluding 'Why' and contain 4 questions has “NO ANSWER” to testing the confident level of the similarity.

SQuAD dataset:
Additionally, the Stanford Question Answering Dataset (SQuAD) Dev v2.0, a benchmark in reading comprehension, was employed to test the system's proficiency in handling real-world data. It required converting the dataset from JSON to a more manageable data frame format, enabling a structured testing environment where each article's first question and answer were extracted to assess the system's effectiveness in understanding and responding to varied queries (The Stanford NLP Group 2024).

3. Preprocessing

In IR-QA systems utilizing pretrained models, preprocessing is minimal to maintain text integrity, as these models are trained on raw text and expect similar input to ensure accurate output. Traditional steps like tokenization are avoided to preserve natural language context. The main preprocessing task is to enhance data quality and relevance, notably by removing articles without author information, thus ensuring the dataset's credibility. This results in a refined dataset aligned with the model's training, optimizing the IR-QA system's accuracy and effectiveness.

4. System Architecture

4.1. The overall architecture of the system

The IR-QA system initiates its process once the article dataset and user query are inputted. It commences with the article retrieval phase, identifying the top 10 articles most pertinent to the query. These articles are then subjected to coreference resolution to substitute pronouns like "he", "she", and "it" with their actual references, thereby clarifying the text. Subsequent text matching identifies the most relevant sentence across the articles. The final answer extraction can operate via two distinct pathways: Type-based Answer Extraction or BERT for Question Answering, depending on the method chosen, guided by the relevance and context provided by the preceding stages.

Article retrieval

Utilizes TF-IDF for text vectorization and Nearest Neighbors algorithm, configured with cosine similarity, to select the top 10 related articles. This phase ensures that both articles and the query are appropriately vectorized and compared to retrieve the most pertinent content.

Coreference Resolution

To refine the accuracy of subsequent text matching, the coreference resolution employs the "en_core_web_md" model and the "fastcoref" package (Otmazgin, Cattan & Goldberg, 2022). This combination enhances text clarity by resolving referential expressions, ensuring a coherent and unambiguous article context (spaCy 2024).

Text matching

Employing cosine similarity, TF-IDF vectorization, and the nltk package, this step synthesizes the top 10 articles into a cohesive text, facilitating the extraction of the most relevant sentence through nuanced similarity analysis with the query.

Type-based Answer Extraction

Utilizes the “en_core_web_md” spaCy model, which provides robust Named Entity Recognition (NER) capabilities. These entities are crucial for determining the answer to the user's query based on the question's type (who, where, when, what, how many).

The system first converts the question to lowercase to standardize the format for processing. It then analyzes the question's starting phrase (e.g., who, where, when, what, how many) to determine the type of entity to look for in the most relevant sentence. This approach is essential for aligning the question type with the corresponding entity type. Based on the question's leading word, the system identifies the answer by searching for the first relevant entity in the text: PERSON for "who," LOC or GPE for "where," DATE for "when," ORG for "what," and MONEY or DATE for "how many," corresponding to the query's focus on people, places, times, organizations, or quantities.

The threshold of low similarity sentence is calculated by the “NO ANSWER” self-defined question. The average similarity of these questions is 0.31. Therefore, the threshold of trustworthy sentence was set as 0.35.

BERT for Question Answering

Leverages the 'bert-large-uncased-whole-word-masking-finetuned-squad' model for a deep learning-based approach to QA. This component adapts to the nuanced context and content interplay between the query and the selected text, ensuring accurate and contextually appropriate responses.

4.2. The machine learning models in the system

Three principal machine learning models underpin the IR-QA system, each pivotal at different stages:

"en_core_web_md"

Integral to both Coreference Resolution and Type-based Answer Extraction, this model excels in unified entity recognition and text clarification, forming the backbone of the system's understanding and contextual grounding.

"fastcoref"

Augments the "en_core_web_md" model within the Coreference Resolution phase, enhancing entity recognition consistency and referential clarity, crucial for accurate text matching and context interpretation (Otmazgin**,** Cattan & Goldberg**,** 2022).

'bert-large-uncased-whole-word-masking-finetuned-squad'

Serves as the core of the BERT Question Answering module, employing advanced text processing and machine learning to deduce contextually relevant answers, effectively handling the complexities of natural language query responses.

5. Model Selection

5.1. The choice of the NLP models

When selecting the appropriate NLP model for a question-answering system, accuracy, model size and usage of the model should be considered to ensure that the system is effective, efficient, and scalable. These considerations are vital for the overall performance and usability of the QA system:

Accuracy

Importance of Precision and Recall: Accuracy is paramount in a QA system as it directly affects the trustworthiness and reliability of the answers provided. Precision is the rate of relevant answer retrieval and recall is the ability to retrieve all relevant answers are key metrics in evaluating the accuracy of an NLP model.

Model Size

Larger NLP models, while potentially more accurate due to their complexity and depth, require significant computational power and memory. This can lead to higher operational costs and slower response times, especially in resource-constrained environments.

Usage of the Model

Depending on the application domain of the QA system, certain NLP models might be preferable. For instance, models pre-trained on legal or medical corpora may perform better in specialized QA systems for these fields.

5.2. The evaluation metrics

F1 score, Exact Match and output timing was used to evaluate the model. Since the usual F1 from sklearn is designed for calculating numeric data, therefore it is not applicable in the NLP task. Referring to the Evaluation Scripts v2.0 from SQuAD GitHub, it provides the calculation of F1 score and Exact Match. In this assignment will use it as an evaluation metrics (The Stanford NLP Group 2024).

F1 score

The F1 score assesses the balance between precision and recall, two key performance indicators in IR-QA systems. Precision measures the accuracy of the returned results, indicating the percentage of relevant answers among all provided answers. Recall, on the other hand, assesses the system's ability to retrieve all relevant answers from the dataset. The F1 score, being the harmonic mean of these two metrics, ensures that both are considered equally. A high F1 score indicates that the system effectively balances finding as many relevant answers as possible (recall) while minimizing the number of irrelevant answers returned (precision). This balanced view is essential in QA systems to ensure that they provide accurate and comprehensive answers without overwhelming the user with incorrect or irrelevant information.

Exact Match

Exact Match is a stringent metric that evaluates the accuracy of the QA system by comparing the exactness of the system's answers to the true answers. It counts the instances where the predicted answer perfectly matches the ground truth, considering the entire answer string. This metric is important because it directly reflects the system's ability to provide precise answers as expected by the users. In the context of user satisfaction and system utility, having a high Exact Match score signifies that users can rely on the system to provide exact and complete answers, enhancing trust and efficiency in the system's usage.

Output timing

Output timing measures the speed at which the QA system processes and responds to queries. In practical applications, this metric is crucial because it affects the user experience directly. Users typically expect quick responses from an automated system, and delays can lead to frustration and disengagement. Furthermore, in real-time applications or scenarios where timely information retrieval is essential, such as in customer support or live interactive sessions, having a system that delivers prompt responses is vital. Efficient processing and quick output generation ensure that the system can be effectively integrated into operational workflows, thereby enhancing its usability and user satisfaction.

5.3. The evaluation results

Model	Method	F1	Exact Match	Time (sec)
en_core_web_sm	Rule-based	0.68	0.65	0.31
	BERT	0.64	0.5	1.79
en_core_web_trf	Rule-based	0.78	0.75	1.49
	BERT	0.64	0.5	1.76
en_core_web_md	Rule-based	0.78	0.75	0.74
	BERT	0.64	0.5	1.82
en_core_web_lg	Rule-based	0.73	0.7	0.93
	BERT	0.64	0.5	1.88

Table 1. The results of validation

To identify the most effective model for coreference resolution and type-based answer extraction, a validation method was employed. This approach allowed for a comprehensive evaluation of each model's performance across different subsets of the data, ensuring a robust and reliable selection process. Testing with 20 self-defined question and models included “en_core_web_sm”, “en_core_web_trf”, “en_core_web_md”, and “en_core_web_lg”, which are all part of the spaCy NLP library (spaCy 2024).

The evaluation focused on key metrics, particularly the F1 score and Exact Match, to assess both the accuracy and precision of the models in replicating human-like understanding in coreference resolution and extracting accurate answers. As per the results detailed in Table 1, “en_core_web_md” emerged as the standout performer, demonstrating the highest efficiency in both coreference resolution and answer extraction processes.

The decision to select “en_core_web_md” over other models was influenced by its balanced performance in terms of speed and accuracy. While both “en_core_web_trf” and “en_core_web_md” achieved comparable results in F1 score and Exact Match, “en_core_web_md” was notably faster, offering twice the speed in rule-based prediction tasks and only marginally slower by 0.1 in BERT prediction scenarios compared to “en_core_web_trf”. This speed advantage is significant in a real-world application where response time is crucial for user satisfaction.

6. User Interaction with the System

6.1. The steps of using the IR-QA system

The system is already integrated the Article retrieval, Coreference Resolution, Text matching when user initialize the system. The user provides the data frame which has article content and the

question. The system will run through the process when user initializes the system. The user needs to select between Rule-base or BERT to answer the question. The system will return the answer.

6.2. Example of user session

```
df_clean = df.copy()
question = 'Who is the vice chairman of Samsung?'
ir = Information_Retrieval_System(df_clean,question)
print(ir.Rule_base())
print(ir.Bert())
```

Output

```
Jay Y. Lee
jay y . lee
```

7. System evaluation:

7.1. The testing results of using SQuAD dataset

Method	F1	Exact Match	Avg Time (s)
Rule-based	0.15	0.15	1.52
BERT	0.30	0.25	1.98

Table 2. The testing results of using SQuAD dataset

For evaluating the performance of the IR-QA system, 20 questions from the SQuAD Dev set 2.0 were utilized. These questions were extracted from the original JSON format and converted into a data frame, with the first 500 observations selected to streamline the computation during testing. Specifically, questions numbered 234th to 253rd were used for system evaluation, where the system attempted to find answers within a corpus of 500 articles.

As per the Table 2, the testing revealed that BERT outperformed the rule-based method in QA, demonstrated twice the accuracy in terms of F1 score and a 10% higher rate in Exact Match. This indicates a more nuanced understanding and retrieval capacity by BERT, especially in complex query contexts.

7.2. Discussion

During the model evaluation stage, both question answering methods have over 0.6 accuracy in F1 and rule-based has a faster and more accurate result than the BERT. However, during the SQuAD data set testing, Both method has low accuracy, with BERT slightly outperforms the rule-base.

The Performance in Model Evaluation
Initially, the rule-based method exhibited better performance than BERT during the model evaluation, primarily due to its straightforward mechanism of responding to "Wh" questions (except why and which). This approach, however, lacks contextual relationship analysis, making it effective for simple, directly stated questions but less so for complex queries.

BERT, particularly the “bert-large-uncased-whole-word-masking-finetuned-squad” model, is specifically trained for QA tasks and inherently considers the relational context within the text, contributing to its superior performance in the SQuAD dataset testing.

The Performance in Testing with SQuAD
The discrepancy in performance between self-defined and SQuAD questions highlights the limitation of the rule-based method and TF-IDF in handling rephrased or contextually rich queries. Self-defined questions, often directly extracted from articles, are easier for TF-IDF and cosine similarity to resolve, unlike the more diverse and complex SQuAD questions.

BERT's effectiveness stems from its training on the SQuAD dataset, enabling it to better grasp the essence of the questions and the logical structure of potential answers, thus ensuring higher performance in QA tasks.

Overall Accuracy
The observed decrease in BERT's accuracy from approximately 80% in original configurations to 30% in this report can largely be attributed to differences in the testing environment. Unlike the original tests conducted by the Stanford NLP team with the SQuAD dataset, where articles were specifically selected for the model, this report involves a more complex selection process where articles are chosen from multiple sources and each paragraph is treated as a separate article. This approach likely led to high similarity among paragraphs, complicating the information retrieval process and affecting the system's ability to discern and retrieve the most contextually relevant sentences. Additionally, the system struggled with sophisticated question structures and altered phrasing, further impacting its accuracy in identifying precise content.

8. Conclusion

8.1. The key findings and contributions of the project

Key Findings

The project underscored the optimal model selection with "en_core_web_md" outperforming in coreference resolution and type-based answer extraction for its balance of computational efficiency and accuracy. BERT, especially the "bert-large-uncased-whole-word-masking-finetuned-squad" model, showcased superior accuracy in complex question answering, highlighting the limitations of rule-based methods in understanding context and relationships, and differential performance testing revealed that rule-based methods are effective for straightforward queries, whereas BERT excels in navigating the complexities and nuances of the SQuAD dataset.

Contributions

The project advanced QA system design by illustrating the impact of different NLP models on system efficiency and accuracy, offering an in-depth analysis of model appropriateness relative to question complexity. Practical insights were derived from the system's implementation, highlighting the crucial balance between speed and accuracy in enhancing user experience. A rigorous evaluation framework was established through the use of self-defined and SQuAD questions, creating a comprehensive method for assessing QA system performance. Finally, the project provided valuable guidance for future research, emphasizing the necessity for QA systems to dynamically adapt to query complexity and the significance of contextual understanding in automated answering.

8.2. The challenges faced during the development process

During the development of the IR-QA system, several challenges were encountered. Selecting the optimal NLP model posed a significant challenge due to the need to balance accuracy with computational efficiency. Integrating BERT for complex query understanding required fine-tuning to avoid performance bottlenecks, especially with tokenization limits. The rule-based method, while efficient for direct queries, struggled with complex and nuanced questions, revealing limitations in handling varied linguistic contexts. Additionally, processing and converting large datasets, like SQuAD, into a usable format demanded substantial preprocessing efforts to ensure data quality and relevance.

8.3. Potential areas for future improvements.

Future improvements for the IR-QA Answering system could focus on enhancing its understanding of complex queries and contextual nuances. Incorporating advanced NLP techniques such as deep learning and context-aware algorithms could improve the system's ability to interpret and answer questions more accurately. Expanding the data set and including more diverse question types would also help in training the system to handle a broader range of inquiries. Optimizing computational efficiency, perhaps through more efficient algorithms or hardware acceleration, would improves response times. Additionally, integrating continuous learning mechanisms would allow the system to evolve and adapt to new information and user feedback, enhancing its overall performance and reliability.

9. References

Ke, Y.H., Jin, L., Elangovan, K., Abdullah, H.R., Liu, N., Sia, A.T.H., Soh, C.R., Tung, J.Y.M., Ong, J.C.L. & Ting, D.S.W. 2024, 'Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report', ArXiv, pp. 1-22

Otmazgin, S., Cattan, A. & Goldberg, Y. 2022, 'F-coref: Fast, Accurate and Easy to Use Coreference Resolution', Proceedings of the AACL

spaCy. 2024, Trained Pipelines English, viewed 29 March 2024, <https://spacy.io/models/en>

The Stanford Natural Language Processing Group. 2024, Dev Set v2.0, SQuAD2.0 The Stanford Question Answering Dataset, viewed 29 March 2024, <https://rajpurkar.github.io/SQuAD-explorer>.

The Stanford Natural Language Processing Group. 2024, Evaluation Script 2.0, SQuAD2.0 The Stanford Question Answering Dataset, viewed 29 March 2024, <https://rajpurkar.github.io/SQuAD-explorer>.

10. Appendix

Self-defined questions

id	question	answer
17574	Who is the vice chairman of Samsung?	Jay Y. Lee
17344	What was Chen Zhongshu the head of?	Panzhihua Land and Resources Bureau
17579	What unexpected product contains added sugar?	NO ANSWER
17598	What does the Trump administration frequently raise?	NO ANSWER
17620	What date did Steve Harvey meet Donald Trump?	NO ANSWER
17619	What was the specific law discussed in the call?	NO ANSWER
17311	Where is the senior Republican from?	Oklahoma
17339	When is Donald Trump inaugurated?'	Jan. 20
17620	What issues did Steve Harvey discuss with Donald Trump?	Housing issues
17618	Who vetoed a similar package to the one passed by Republicans in 2015?	President Obama
17575	Where were the light weapons seized that appeared to have been manufactured in Iran?	Near Yemen's coast
17300	When is the Turkish newspaper Hurriyet reported about the gunman?	Monday
17300	Who said at a news conference that investigators believed they found the assailant?	Numan Kurtulmus
17570	Who is the spokesman for the Taiwan Affairs Office in Beijing?	Ma Xiaoguang
17570	Who is naval affairs researcher at the Shanghai University of Political Science and Law?	Ni Lexiong
17568	Where is the first began to investigate Volkswagen early in 2014?	the United States
17623	Who is the Los Angeles bureau chief for California Today?	Adam Nagourney
17619	Who is the chief executive of Lockheed Martin?	Marillyn Hewson

17619	When was top Bush administration officials walked Mr. Obama?	2009