Web Data Processing Systems

Ikram Ait Taleb Naser

1 Introduction

This report presents a method for extracting and validating structured knowledge from the output of a Large Language Model. Specifically, the system utilizes Llama2 7B to generate raw text output and the BERT-based model for Named Entity Recognition (NER) to extract relevant entities from the input and generated text. The program's functionality is structured into four main tasks:

1. **Task 1:** Retrieve the raw output text from the Llama model.
2. **Task 2:** Extract and list entities from both the input text and the raw text output.
3. **Task 3:** Extract an answer (either "yes/no" or a Wikipedia entity link).
4. **Task 4:** Evaluate the correctness of the extracted answer.

2 Methodology

2.1 Entity Extraction

In this step, a pre-trained BERT model fine-tuned to extracts named entities like people, locations, and organizations from text. The NER model is implemented as follows:

```
def extract_entities(text):
    ner_results = ner_pipeline(text)
    entities = list(set(result["word"] for result in ner_results))
    return entities
```

The `extract_entities` function takes the input text and applies the NER pipeline. The result is a list of entities found in the text. Each entity is returned as a unique word extracted from the text.

2.2 Querying Llama Model

Once entities are extracted from the input text, we proceed to query the Llama2 model, which generates raw output text. The Llama model is queried as follows:

```
def query_llama(question):
    output = llm(question, max_tokens=128, echo=True)
    return output["choices"][0]["text"]
```

The `query_llama` function sends the user's question to the Llama model and retrieves the raw text response. A parameter of 128 is set to limit the length of the response.

2.3 Entity Linking and Disambiguation

The next step is to link the extracted entities to their corresponding Wikipedia pages using the Wikipedia API. The main goal then is to resolve ambiguities in the entity name based on the question's context:

```
def disambiguate_entity(entity_name, context):
    url = "https://www.wikidata.org/w/api.php"
    params = {
        "action": "wbsearchentities",
        "format": "json",
        "language": "en",
        "search": entity_name
    }
    response = requests.get(url, params=params)
    if response.status_code == 200:
        results = response.json().get("search", [])
        for result in results:
            description = result.get("description", "")
            if any(word.lower() in description.lower() for word in context.split()):
                return result["label"], result["description"]
    return entity_name, "No disambiguation found"
```

2.4 Answer Extraction

Once output and linked entities are obtained, I extract the answer to the question. This process differs depending on whether the question is asking for a "yes/no" answer (for questions 1,2 and 5) or for a specific entity (such as a person or location). The answer extraction function is as follows:

```
def extract_answer(question, raw_text, linked_entities):
    def classify_yes_no(text):
        yes_patterns = [r"\b(yes|yeah|yep|correct|true|indeed|absolutely|definitely|of course)\b", r"it
            is", r"that's right", r"without a doubt"]
        no_patterns = [r"\b(no|nope|false|not at all|incorrect|never|absolutely not)\b", r"it is not", r
            "that's wrong", r"under no circumstances"]

        if any(re.search(pattern, text, re.IGNORECASE) for pattern in yes_patterns):
            return "yes"
        if any(re.search(pattern, text, re.IGNORECASE) for pattern in no_patterns):
            return "no"
        return "Answer not found"

    if question.lower().startswith(("is", "does", "are", "was", "were", "can", "should")):
        return classify_yes_no(raw_text)

    for entity in linked_entities:
        if entity[0] in raw_text:
            return entity[0]

    return "Answer not found"
```

The `classify_yes_no` function identifies patterns in the raw text that indicate a "yes" or "no" response. The `extract_answer` function handles the classification and returns either a "yes/no" answer or the name of the entity found in the raw text.

2.5 Step 5: Fact-Checking

Finally, the extracted answer is fact-checked by comparing it to the information found on the corresponding Wikipedia page. The fact-checking function is as follows:

The `fact_check_answer` function checks if the extracted answer matches the expected entity and verifies it by retrieving the summary of the entity's Wikipedia page.

3 Results

The system will be evaluated on 5 questions:

1. "Is Managua the capital of Nicaragua?",
2. "Is it true that China is the country with most people in the world?",
3. "The largest company in the world by revenue is Apple.",
4. "Who is the director of Pulp Fiction?",
5. "Is it true that the monarch of England is also the monarch of Canada?"

The evaluation showed that the system performed well in extracting factual information when the context was clear and straightforward, as demonstrated in Questions 2 and 5. However, it faced challenges in more complex situations where the context wasn't directly confirmed in the raw text, as seen in Questions 1, 3 and 4. This indicates that the Yes/No classification mechanism is functioning properly when the LLM provides direct, affirmable answers. However, for other questions requiring entity-based extraction, such as "Who is the director of Pulp Fiction?" and "Is Managua the capital of Nicaragua?", the system failed to extract an answer.

A Appendix: Results

Here are the raw results for the example questions processed by the system.

```
Processing: question-001 Is Managua the capital of Nicaragua?
Result:
Input (A): Is Managua the capital of Nicaragua?
Entities in Input (A): ['Managua', 'Nicaragua']
Raw Text (B): Is Managua the capital of Nicaragua?
 kwiet 24, 2022 at 11:46 pm
What is the capital of Nicaragua?
```

```
Is Panama the capital of Nicaragua?
What is the biggest city in Nicaragua?
Is Nicaragua richer than Costa Rica?
What is the capital of Nicaragua?
Is there a capital of Central America?
What is the largest city in Central America?
What is Nicaragua known for?
What are 5 things Nicaragua is known for?
What is Central America known for?
What is the capital of Gu
Entities in Raw Text (B): ['Panama', 'Managua', 'Nicaragua', 'Central America', 'Gu', 'Costa
    Rica']
Linked Entities: [('Panama City', 'https://en.wikipedia.org/wiki/Panama_City', 'capital of
    Panama'), ('Managua', 'https://en.wikipedia.org/wiki/Managua', 'capital of Nicaragua'),
    ('MTV Hits (Latin America)', 'https://en.wikipedia.org/wiki/MTV_Hits_(Latin_America)',
    'Latin American pay-television music channel.'), ('Central America',
    'https://en.wikipedia.org/wiki/Central_America', 'subregion of the Americas'),
    ('Gujarati', 'https://en.wikipedia.org/wiki/Gujarati', 'Indo-Aryan language that is
    spoken on the state of Gujarat'), ('Costa Rica',
    'https://en.wikipedia.org/wiki/Costa_Rica', 'episode of Wildboyz (S2 E3)')]
Extracted Answer: Answer not found
Correctness: incorrect

Processing: question-002 Is it true that China is the country with most people in the world?
Result:
Input (A): Is it true that China is the country with most people in the world?
Entities in Input (A): ['China']
Raw Text (B): Is it true that China is the country with most people in the world?
 gepr ft.de
There are actually many sites offering the same service. They are very similar to each other
    and the quality is the same. The prices can differ a little but they are not very
    different. I've used all of them and I don't have any recommendations as to one in
    particular. I've tried many sites and the quality is the same.
Chinese are the most populous nationality in the world. I've used many sites and I think
    there are many sites that can be used to write a dissertation. I've tried a lot of sites
    and I think I can
Entities in Raw Text (B): ['China', 'Chinese']
Linked Entities: [("People's Republic of China",
    'https://en.wikipedia.org/wiki/People%27s_Republic_of_China', 'country in East Asia'),
    ('Chinese', 'https://en.wikipedia.org/wiki/Chinese', 'language group of the Sinitic
    languages')]
Extracted Answer: yes
Correctness: correct

Processing: question-003 The largest company in the world by revenue is Apple.
Result:
Input (A): The largest company in the world by revenue is Apple.
Entities in Input (A): ['Apple']
Raw Text (B): The largest company in the world by revenue is Apple. sierp. 15, 2020.
Is Apple the largest company in the world?
What is the biggest company in the world 2020?
What company is the biggest?
What is the biggest company in the world?
What is the biggest company ever?
What is the biggest company 2020?
What is the biggest company?
What is the biggest company in the USA?
Is Apple the richest company in the world?
What company has the most employees?
How many employees does Apple have 2021?
What company has the
Entities in Raw Text (B): ['USA', 'Apple']
Linked Entities: [('United States', 'https://en.wikipedia.org/wiki/United_States', 'country
    primarily located in North America'), ('Apple', 'https://en.wikipedia.org/wiki/Apple',
    'American multinational technology company based in Cupertino, California')]
Extracted Answer: Apple
Correctness: incorrect

Processing: question-004 Who is the director of Pulp Fiction?
Result:
Input (A): Who is the director of Pulp Fiction?
Entities in Input (A): ['Pulp Fiction']
Raw Text (B): Who is the director of Pulp Fiction?
 gepr ft on 12/7/2013
What is the best movie of the 2000s?
Ratatouille on 9/19/2009
```

```
What are the top 100 movies ever made?
Who directed the movie Pulp Fiction?
What is the movie Pulp Fiction about?
What does the movie Pulp Fiction mean?
What was Quentin Tarantino's first movie?
What was the first Quentin Tarantino movie?
When was Tarantino's movie Pulp F
Entities in Raw Text (B): ['Tarantino', 'Ratatouille', 'Quentin Tarantino', 'Pulp F', 'Pulp
    Fiction']
Linked Entities: [('Santa Fe-class submarine',
    'https://en.wikipedia.org/wiki/Santa_Fe-class_submarine', '1931 class of Argentine Navy
    submarines'), ('ratatouille', 'https://en.wikipedia.org/wiki/Ratatouille', 'French
    dish'), ('pulp fiction', 'https://en.wikipedia.org/wiki/Pulp_fiction', 'popular fiction
    dealing with lurid or sensational topics and originally printed on inexpensive paper made
    from wood pulp'), ('pulp fiction', 'https://en.wikipedia.org/wiki/Pulp_fiction', 'popular
    fiction dealing with lurid or sensational topics and originally printed on inexpensive
    paper made from wood pulp')]
Extracted Answer: Answer not found
Correctness: incorrect

Processing: question-005 Is it true that the monarch of England is also the monarch of Canada?
Result:
Input (A): Is it true that the monarch of England is also the monarch of Canada?
Entities in Input (A): ['Canada', 'England']
Raw Text (B): Is it true that the monarch of England is also the monarch of Canada?
 Hinweis: Der Beitrag ist nur auf Englisch verf gbar.
Queen Elizabeth II is the monarch of Canada.
Queen Elizabeth II is the Head of State of Canada. Canada is a constitutional monarchy in
    which the role of the monarch is strictly defined by law, and the monarch acts on the
    advice of her ministers. The Queen does not directly control the government or have veto
    power over legislation, though some of her prerogative powers (such as the power to
    dissolve Parliament) have become largely decorative due to constitutional conventions.
    The position of the monarch is to be
Entities in Raw Text (B): ['Canada', 'England', 'Englisch', 'Elizabeth II', 'Der Be']
Linked Entities: [('Newfoundland and Labrador',
    'https://en.wikipedia.org/wiki/Newfoundland_and_Labrador', 'province of Canada'),
    ('England', 'https://en.wikipedia.org/wiki/England', 'country in north-west Europe, part
    of the United Kingdom'), ('Englischer Garten',
    'https://en.wikipedia.org/wiki/Englischer_Garten', 'a large public park in the centre of
    Munich'), ('Elizabeth II', 'https://en.wikipedia.org/wiki/Elizabeth_II', 'Queen of the
    United Kingdom from 1952 to 2022'), ('Der Bergdoktor',
    'https://en.wikipedia.org/wiki/Der_Bergdoktor', 'German medical drama television series')]
Extracted Answer: yes
Correctness: correct
```