Practical Assignment

Web Data Processing Systems

Jacopo Urbani

Department Computer Science
Vrije Universiteit Amsterdam, The Netherlands

2023/2024

The assignment

Goal: Put into practice what you have learned during the lectures

In a nutshell

You are asked to develop a method for translating the output of a large language model (LLM) into clean formal statements to check if the LLM's output is correct

The assignment

Goal: Put into practice what you have learned during the lectures

In a nutshell

You are asked to develop a method for translating the output of a large language model (LLM) into clean formal statements to check if the LLM's output is correct

Why?

LLMs are becoming very popular and people start to use them extensively. Unfortunately, the output of the LLMs is often **noisy** and/or **incorrect**! Knowledge bases (e.g., Wikidata) contain high-quality knowledge, hence they can help to check if the LLM's output is correct.

The assignment

Goal: Put into practice what you have learned during the lectures

In a nutshell

You are asked to develop a method for translating the output of a large language model (LLM) into clean formal statements to check if the LLM's output is correct

Why?

LLMs are becoming very popular and people start to use them extensively. Unfortunately, the output of the LLMs is often **noisy** and/or **incorrect**! Knowledge bases (e.g., Wikidata) contain high-quality knowledge, hence they can help to check if the LLM's output is correct.

Practical info (check also the Canvas page!)

- Deadline: **22/12/2023** at 23:59 (Amsterdam time)
- You must work in groups of four students
- Submission done via Canvas
  - Source code
  - Report (max 2 pages), whatever format
  - (Link to a) video presentation (max 10 minutes)

Example

Let's assume we want to query a LLM and ask whether the capital of Italy is Rome.

Input to a LLM "Is Rome the capital of Italy?"

Output of a LLM surely it is but many don't know this fact that Italy was not always called as Italy. Before Italy came into being in 1861, it had several names including Italian Kingdom, Roman Empire and the Republic of Italy among others. If we start the chronicle back in time, then Rome was the first name to which Romans were giving credit. Later this city became known as "Caput Mundi" or the capital of the world...

The goal is to answer the following questions

Task 1: Which are the entities mentioned in the text?	Rome, Italy,...
Task 2: What's the answer?	Yes/No
Task 3: Is the returned answer correct?	Yes/No

Example

Let's assume we want to query a LLM and ask whether the capital of Italy is Rome.

Input to a LLM	"Is **Rome** the capital of **Italy**?"
Output of a LLM	surely it is but many don't know this fact that **Italy** was not always called as **Italy**. Before **Italy** came into being in 1861, it had several names including **Italian Kingdom**, **Roman Empire** and the **Republic of Italy** among others. If we start the chronicle back in time, then **Rome** was the first name to which **Romans** were giving credit. Later this city became known as **"Caput Mundi"** or the capital of the world...

Task 1

Entities are important to identify the meaning of the text. You must recognize all entities mentioned in the text (input **and** output) and reporting the corresponding Wikipedia page

Desired output

Rome	⇒	`https://en.wikipedia.org/wiki/Rome`
Italy	⇒	`https://en.wikipedia.org/wiki/Italy`
Roman Empire	⇒	`https://en.wikipedia.org/wiki/Roman_Empire`
Italian Kingdom	⇒	`https://en.wikipedia.org/wiki/Kingdom_of_Italy`
...		

Example

Let's assume we want to query a LLM and ask whether the capital of Italy is Rome.

Input to a LLM	"Is Rome the capital of Italy?"
Output of a LLM	surely it is but many don't know this fact that Italy was not always called as Italy. Before Italy came into being in 1861, it had several names including Italian Kingdom, Roman Empire and the Republic of Italy among others. If we start the chronicle back in time, then Rome was the first name to which Romans were giving credit. Later this city became known as "Caput Mundi" or the capital of the world...

Task 2

The LLM returns: "surely it is but many don't know this ..." ⇒ Is it a yes or a no?

First, a big **challenge** is to first determine if the answer to a given input is either "yes/no" or an entity. For example, a similar request could have been "The capital of Italy is ...".

Once you determined the type of question, you must extract the answer from the returned text

Example

Let’s assume we want to query a LLM and ask whether the capital of Italy is Rome.

Input to a LLM “Is Rome the capital of Italy?”

Output of a LLM surely it is but many don’t know this fact that Italy was not always called as Italy. Before Italy came into being in 1861, it had several names including Italian Kingdom, Roman Empire and the Republic of Italy among others. If we start the chronicle back in time, then Rome was the first name to which Romans were giving credit. Later this city became known as “Caput Mundi” or the capital of the world...

Task 3

Once you determined the answer and the involved entities, your goal is to **validate** the answer, for instance by consulting an external knowledge base. In this example, the answer turned out to be correct.

Important

This task hinges on the quality of the previous ones. If you wronged extracting the answer from the text, the validation process may return a different output. This dependency is very common in real extraction pipelines

The role of the assignment within this course

The assignment fulfills **two** main objectives:

1. It allows you to develop a method to extract symbolic knowledge from unstructured sources (text produced by language models),
2. ...and to use the knowledge to validate factual statements (made by the language model)

The role of the assignment within this course

The assignment fulfills **two** main objectives:

1. It allows you to develop a method to extract symbolic knowledge from unstructured sources (text produced by language models),
2. ...and to use the knowledge to validate factual statements (made by the language model)

To this end, you can use techniques for:

- clean noisy text so that it can be used for further processing (NLP pre-processing)
- recognize and disambiguate entities (topic 4)
- interact with a state-of-the-art LLM (topic 5)
- validate knowledge (topic 9)

Moreover, you may also want to use techniques like WSD or statistical inference if the KB does not contain relevant knowledge, etc.

Working in a group

In the real world, you often have to work together. Group assignments is a way to learn to do so

Some advices

- Frequent communication is fundamental
- Let your team know what you expect from them
- Divide the work optimally
- If things don't go as you hoped, address the issue(s) immediately

I expect that every group member can explain the content of the submission. So let the others know what you have done!

Development constraints

Which language should I use?

Whatever language you want. I strongly advice you to use Python. If you use another language, make sure that it works in the provided Docker image so that I can test your solution

Development constraints

Which language should I use?

Whatever language you want. I strongly advice you to use Python. If you use another language, make sure that it works in the provided Docker image so that I can test your solution

Which techniques should I use?

Feel free to use any technique that we studied during the course. You can also use other libraries (e.g., for entity recognition) provided they don't implement key tasks like entity linking

Important: Better a simple solution than a complex one that does not!

Knowledge Bases and LLMs

KBs

You must return links to **Wikipedia** entities

WIKIPEDIA
The Free Encyclopedia

You can use whatever KB you like. The ones built from Wikipedia are:

- Wikidata
- YAGO
- DBPedia

You can implement a solution that uses an online service, but this will impact scalability

LLMs

You can use whatever LLM you like (provided it's a state-of-the-art one)

If you decide to use *(Chat)GPT* or any other non-free service, you must also provide a free alternative that I can use to test your solution

Be aware that using such services will impact your grade since they do some of the cleaning that you are supposed to do

My advice: Use the LLaMA language model, which is free and reasonably good. A small version of this model is in the Docker image.

Docker

Docker (docker.com) is a popular framework to run code in virtual machines. It works on all major operating systems and it is a way to deploy a single environment on different machines

Why?

Two reasons

- **It's good for you.** The image that I provide contains a small LLM with a minimal example that you can use to start the development
- **It's good for me.** With so many groups, I need a single platform for testing all solutions

Important!

During the development, you don't have to use Docker all the time. You can use the programs you are already familiar with. In that case, make sure that your program will work in the Docker environment that I provide (e.g., by giving me instructions on how install external libraries)

Grading

I will grade your submission considering the following criteria

- Originality (5%)
- Performance on entity disambiguation (10%)
- Performance on answer extraction (10%)
- Performance on the fact checking (10%)
- Code quality (15%)
- Documentation (10%)
- Compliance (15%)
- Scalability (15%)
- Presentation (10%)

If I discover that some group members did not contribute, I'll grade them differently

What if a group member quits the course?

Unfortunately this happens all the time. Go on and finish the implementation without them. I'll take it into account while grading

Conclusion

More advices

- 1:1 meetings with them are useful to discuss issues or to get some advice on how to proceed
- don't wait until the last minute!
- Don't go immediately for very complicated solutions. First develop an easy solution and then build from it

Good luck!