Natural Language Processing

HW3 - Decoding, Prompting and Instruction Tuning

Computer Software Engineering(컴퓨터소프트웨어학부)

2020081421

Kim Sung Hyuk(김성혁)

1. [Model Explanation]

Model : Llama 2 7B-hf(Lagre Language Model Meta AI)

- **Base model :** "meta-llama/Llama-2-7b-hf"
- **Instruction Fine-tuned Model :** "meta-llama/Llama-2-7b-chat-hf"

LLaMA 2 (Large Language Model Meta AI) is a family of open-source large language models developed by Meta. It offers various parameter options, allowing the selection of a suitable model(7B parameters, most smaller one in LLaMA2) based on the given hardware(e.g. GPU, RAM)

As the model is based on a Transformer Decoder, it supports auto-regressive generation. Although LLaMA 3 supports multiple languages, only English will be used for this homework3 , that making LLaMA 2 sufficient.

LLaMA 2 stands out for its variety of options, ease of use, and strong performance, which were key factors in selecting this model for the task.

2. [Decoding algorithms]

- **3 unfinished test sentences :**

Test Sentences

```
[ ]  # Define 3 candidate sentences
     candidate_sentences = [
         'Hanyang University is',
         'Difficulty of learning Korean compare to English is',
         'Alpine ski is'
     ]
```

- 'Hanyang University is~'
- 'Difficulty of learning Korean compare to English is~'
- 'Alpine ski is~'

2-1. Greedy Search

- It is simplest decoding method.
- Selects the highest probability as its next word **at each timestep** t.

[Output Result]

```
Output for Greedy Search:
--------------------------------------------------------------------------------------------------------------
Input Sentence: Hanyang University is
Generated Output: Hanyang University is a private research university in Seoul, South Korea. It was established in 1939 as the first technological university in Korea.

--------------------------------------------------------------------------------------------------------------
Input Sentence: Difficulty of learning Korean comapre to English is
Generated Output: Difficulty of learning Korean comapre to English is 2.6 out of 10.
Which is easier to learn, Korean or English?
Is Korean or English easier to learn
--------------------------------------------------------------------------------------------------------------
Input Sentence: Alpine ski is
Generated Output: Alpine ski is a form of skiing where skiers descend a mountain using alpine skis, fixed-heel bindings, and poles. Alpine ski
--------------------------------------------------------------------------------------------------------------
```

[Pro]

- Simple and fast

[Con]

- Repetition Happen
- Miss high probability words hidden begind a low probability word -> We can alleviate by using **Beam Search(2-2)**

2-2. Beam Search

- On each step of decoder, keep track of the k most probable partial translations. (k : beam size)
- All hypothesis has score(log probability)
- Decode untill **reach timestep T** or **at least n completed hypothesis**

[Output Result]

```
Output for Beam Search with Early Stopping:
--------------------------------------------------------------------------------------------------------------
Input Sentence: Hanyang University is
Generated Output (Beam Search): Hanyang University is located in Seoul, South Korea. It was founded in 1939. It is accredited by Ministry of Education, Science and
--------------------------------------------------------------------------------------------------------------
Input Sentence: Difficulty of learning Korean comapre to English is
Generated Output (Beam Search): Difficulty of learning Korean comapre to English is about the same.
Difficulty of learning Korean comapre to Japanese is about the same.
Difficulty of learning Korean comap
--------------------------------------------------------------------------------------------------------------
Input Sentence: Alpine ski is
Generated Output (Beam Search): Alpine ski is a winter sport in which competitors descend a snow-covered slope on skis with fixed-heel bindings, using poles for balance and
--------------------------------------------------------------------------------------------------------------
```

[Pro]

- Reduces the risk of missing hidden high probability word sequences
- Work well in task such as Machine Translation or Text Summarization

[Con]

- Does not guarantee to find optimal solution(or most likely output)
- Repetition Happen -> we can solve it by **no repeat n-gram(2-3)**
- Not good at open-ended generation such as dialog and story generation

2-3. No Repeat N-Gram

- Simple option for not to generate(or reduce) repetition on Beam Search Decoding

[Output Result]

```
Output for Beam Search with Multiple Sequences:
--------------------------------------------------------------------------------------------------------------
Input Sentence: Hanyang University is
Sequence 1: Hanyang University is a private research university located in Seoul, South Korea. It was founded in 1939 as the first technological university in the country
--------------------------------------------------------------------------------------------------------------
Input Sentence: Difficulty of learning Korean comapre to English is
Sequence 1: Difficulty of learning Korean comapre to English is 1.5 out of 5 (hard).
Which is easier to learn, Korean or English?
Is Korean easier than English to
--------------------------------------------------------------------------------------------------------------
Input Sentence: Alpine ski is
Sequence 1: Alpine ski is a winter sport in which competitors descend a snow-covered slope on skis with fixed-heel bindings, using poles for balance and
--------------------------------------------------------------------------------------------------------------
```

[Pro]

- Help avoid generating repetitive phrases or words by restricting the model from repeating n-grams (a sequence of n consecutive words)
- By limiting repeated structures, the generated text often sounds more fluent and human-like

[Con]

- Preventing repetition too aggressively might result in a loss of context or coherence
- If the restriction is too stringent, it might force the model to generate less optimal or less accurate outputs

2-4. Vanilla Sampling

- Randomly picking the next word wt according to its conditional probability distribution

[Output Result]

```
Output for Sampling with Top_k=0:
--------------------------------------------------------------------------------------------------------------
Input Sentence: Hanyang University is
Generated Output (Sampling): Hanyang University is located in Seoul, South Korea. It is locally known as 한양대학교. The university was established in 19
--------------------------------------------------------------------------------------------------------------
Input Sentence: Difficulty of learning Korean comapre to English is
Generated Output (Sampling): Difficulty of learning Korean comapre to English is the same
A lot of people have a misconception that learning Korean is difficult. This is because Korean has a lot of characters and some of
--------------------------------------------------------------------------------------------------------------
Input Sentence: Alpine ski is
Generated Output (Sampling): Alpine ski is a winter sport in which participants descend a snow-covered slope on skis with fixed-heel bindings, using poles for propulsion and
--------------------------------------------------------------------------------------------------------------
```

[Pro]

- Repetitiion is not occuring as much as in Beam Search or Greedy Decoding Algorithms

[Con]

- Model often generate incoherent gibberish beacuse Vanilla sampling makes every token in the vocabulary an option.

2-5. Top - K Sampling

- Sample from the top k tokens in the probability distribution.
- Incresing k -> Diverse, risky output
- Decresing k -> Safe, generic output

[Output Result]

```
Output for Sampling with Top_k=50:
--------------------------------------------------------------------------------------------------------------
Input Sentence: Hanyang University is
Generated Output (Sampling, Top_k=50): Hanyang University is a private research university in Seoul, South Korea. It was established in 1939 as a technical school, and was formall
--------------------------------------------------------------------------------------------------------------
Input Sentence: Difficulty of learning Korean comapre to English is
Generated Output (Sampling, Top_k=50): Difficulty of learning Korean comapre to English is same as learning English from a Korean?
I have a question that I have been wondering for a long time. I am learning Korean and I am
--------------------------------------------------------------------------------------------------------------
Input Sentence: Alpine ski is
Generated Output (Sampling, Top_k=50): Alpine ski is one of the most popular sports in the world. Alpine skiing is a very exciting sport that requires a lot of training. Alpine sk
--------------------------------------------------------------------------------------------------------------
```

[Pro]

- Much more human like text compare to vanila Sampling

[Con]

- Cannot **dynamically adapt** the k value

2-6. Top - p (Nucleus) Sampling

- Sample from all tokens in the top p cumulative probability mass

[Output Result]

```
Output for Sampling with Top_k=0 and Top_p=0.92:
--------------------------------------------------------------------------------------------------------------
Input Sentence: Hanyang University is
Generated Output (Sampling, Top_p=0.92, Top_k=0): Hanyang University is one of the top private universities in Korea. It is also one of the top 5 universities in Korea. It was ran
--------------------------------------------------------------------------------------------------------------
Input Sentence: Difficulty of learning Korean comapre to English is
Generated Output (Sampling, Top_p=0.92, Top_k=0): Difficulty of learning Korean comapre to English is hard.
There is a lot of homophones and grammar rules.
So, I recommend to learn basic Korean before learning advanced Korean.

--------------------------------------------------------------------------------------------------------------
Input Sentence: Alpine ski is
Generated Output (Sampling, Top_p=0.92, Top_k=0): Alpine ski is a popular winter sport, especially in the Alps. It is a sport in which skiers descend a snow-covered mountain using
--------------------------------------------------------------------------------------------------------------
```

[Pro]

- Number of words in the set can dynamically increase and decrese
- Diversity increses

[Con]

- Sensitive to p value.

3. [Prompting Base Model vs Instruction fine-tuned Model]

- Let's look how our base model and instruction fine tuned model react to the complex question

- **Using Top-k decoding algorithm** to generate answer

- Basic question, Few-Shot prompt, CoT prompt

Complex step by step question :

"Kim has 3 boxes of apples, with each box containing 10 apples. Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes."

How many apples does Kim have now?"

3-1. [Base Model]

-model : "meta-llama/Llama-2-7b-hf"

3-1-1. Base Model

[Input : Complex question]

```
# Test sentence
complex_questions ="""
Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?
"""
```

[Output Result]

```
Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

## Solution

Kim has 6 apples.

## Solution

Kim has 6 apples.

## Solution

Kim has 6 apples.

## Solution

Kim has 6 apples.

## Solution

Kim has 6 apples.

## Solution

Kim has 6 apples.

## Solution

Kim has 6 apples.

##
```

- **Resulted in the model generating repetitive sequences**
- **Do not give a proper answer to the given complex question**

3-1-2. Few-Shot Prompting

- Using Few-Shot Prompting to our Base Model
- Few-shot prompting involves providing the model with a small number of examples (in this case, 3) within the prompt to help it understand the task. By showing the model how a problem is solved with a few specific examples, the model can generalize from these examples and apply the learned patterns to new inputs

[Input : Few-Shot Prompt]

```
# Few-Shot Prompt
few_shot_prompt = """
Question1: What is 6 - 4 ?
Answer1: Answer is 2.

Question2: Alex has 5 pens, and he buys 3 more pens. How many pens does Alex have now?
Answer2: Alex now has 8 pens.

Question3: Kim has 10 apples and gives 4 apples to Jung. How many apples does Kim have now?
Answer3: Kim has 6 apples left.

Quserion4: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question4?
Answer4:
"""
```

[Output]

```
Question1: What is 6 - 4 ?
Answer1: Answer is 2.

Question2: Alex has 5 pens, and he buys 3 more pens. How many pens does Alex have now?
Answer2: Alex now has 8 pens.

Question3: Kim has 10 apples and gives 4 apples to Jung. How many apples does Kim have now?
Answer3: Kim has 6 apples left.

Quserion4: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question4?
Answer4:
Kim has 10 apples left.

Answer4:
Kim has 10 apples left.

Answer4:
Kim has 10 apples left.

Answer4:
Kim has 10 apples left.

Answer4:
Kim has 10 apples left.

Answer4:
Kim has 10 apples left.

Answer4:
Kim has
```

- **Same as the above case, repetition occurred.**
- **Do not give a proper answer to the given complex question**

3-1-3. Chain-of-Thought(COT) Prompt

- Using **Chain-of-Thought(COT)** Prompting to our Base Model
- Chain-of-Thought (CoT) prompting encourages the model to explicitly break down the reasoning process step-by-step, rather than jumping straight to an answer. This is particularly useful for complex tasks that involve multi-step reasoning. By **prompting the model to reason** through the problem, CoT helps the model provide more accurate and logically coherent answers, especially for tasks that require intermediate steps, like math problems(like this case) or logical deductions.

[Input : Chain-of-Thought(COT) Prompt]

```
# COT Prompt
cot_prompt = """
Question1:Question1: Sarah has 100 dollars. She buys a book for 40 dollars and a pencil for 5 dollars.
How much money does Sarah have left?
Answer1: Let's go through this step-by-step.
Sarah starts with 100 dollars. After buying the book for 40 dollars, she has 100 - 40 = 60 dollars.
Then, after buying the pencil for 5 dollars, she has 60 - 5 = 55 dollars. So, Sarah has 55 dollars left.

Question2: John's flight departs at 3:00 PM and arrives at 6:30 PM. How long is John's flight?
Answer2: Let's go through this step-by-step.
John's flight departs at 3:00 PM and arrives at 6:30 PM. From 3:00 PM to 6:00 PM is 3 hours.
From 6:00 PM to 6:30 PM is 30 minutes. So, the total flight duration is 3 hours and 30 minutes.

Question3: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question3?
Answer3:

"""
```

[Output]

```
Question1:Question1: Sarah has 100 dollars. She buys a book for 40 dollars and a pencil for 5 dollars.
How much money does Sarah have left?
Answer1: Let's go through this step-by-step.
Sarah starts with 100 dollars. After buying the book for 40 dollars, she has 100 - 40 = 60 dollars.
Then, after buying the pencil for 5 dollars, she has 60 - 5 = 55 dollars. So, Sarah has 55 dollars left.

Question2: John's flight departs at 3:00 PM and arrives at 6:30 PM. How long is John's flight?
Answer2: Let's go through this step-by-step.
John's flight departs at 3:00 PM and arrives at 6:30 PM. From 3:00 PM to 6:00 PM is 3 hours.
From 6:00 PM to 6:30 PM is 30 minutes. So, the total flight duration is 3 hours and 30 minutes.

Question3: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question3?
Answer3:

Let's go through this step-by-step.
Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.

After giving the box to Jung, Kim has 2 boxes of apples, with each box containing 10 apples.
After eating 3 apples from one of the remaining boxes, Kim has 1 box of apples, with each box containing 10 apples.

So, Kim has 10 apples in the box that he gave to Jung and 10 apples in the box that he ate from.
So, Kim has 20 apples.
```

- **Model attempts to solve the problem step by step in response to the question**
- **Do not give a proper answer to the given complex question.**
- **Closer to the correct answer compared to the few-shot prompt approach**

3-1. [Instruction fine-tuned model]

- “meta-llama/Llama-2-7b-chat-hf”

3-2-1. Instruction fine-tuned model

[Input : Complex Quetion]

```
# Test sentence
complex_questions ="""
Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?
"""
```

[Output]

```
Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

Answer:
Kim has 2 boxes of apples left, with each box containing 10 apples. So, Kim has 20 apples left in total.

Explanation:
Kim gives 1 box of apples to Jung, so Kim has 3 boxes of apples left.
Kim eats 3 apples from one of the remaining boxes, so Kim has 2 boxes of apples left.
Each box contains 10 apples, so Kim has 20 apples left in total.
```

- **Model not only attempts to generate an answer but also tries to provide an explanation**
- **However, it failed to deliver the desired exact answer**
- **Repetition does not occur**

3-2-2. Few-Shot Prompting

[Input : Chain-of-Thought(COT) Prompt]

```
# Few-Shot Prompt
few_shot_prompt = """
Question1: What is 6 - 4 ?
Answer1: Answer is 2.

Question2: Alex has 5 pens, and he buys 3 more pens. How many pens does Alex have now?
Answer2: Alex now has 8 pens.

Question3: Kim has 10 apples and gives 4 apples to Jung. How many apples does Kim have now?
Answer3: Kim has 6 apples left.

Quserion4: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question4?
Answer4:
"""
```

[Output]

```
Question1: What is 6 - 4 ?
Answer1: Answer is 2.

Question2: Alex has 5 pens, and he buys 3 more pens. How many pens does Alex have now?
Answer2: Alex now has 8 pens.

Question3: Kim has 10 apples and gives 4 apples to Jung. How many apples does Kim have now?
Answer3: Kim has 6 apples left.

Quserion4: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question4?
Answer4:
Kim has 2 boxes of apples left, and each box contains 7 apples. So, Kim has 14 apples left.
```

- **Model finally provided the correct answer!!!**
- **Additionally, in the Few Shot 3 example provided as input to the model, a detailed explanation was omitted in the answer**
- **Similarly, the model was able to generate the answer while providing only the necessary explanation**

3-2-3. Chain-of-Thought(COT) Prompt

[Input : Chain-of-Thought(COT) Prompt]

```
# COT Prompt
cot_prompt = """
Question1:Question1: Sarah has 100 dollars. She buys a book for 40 dollars and a pencil for 5 dollars.
How much money does Sarah have left?
Answer1: Let's go through this step-by-step.
Sarah starts with 100 dollars. After buying the book for 40 dollars, she has 100 - 40 = 60 dollars.
Then, after buying the pencil for 5 dollars, she has 60 - 5 = 55 dollars. So, Sarah has 55 dollars left.

Question2: John’s flight departs at 3:00 PM and arrives at 6:30 PM. How long is John’s flight?
Answer2: Let's go through this step-by-step.
John’s flight departs at 3:00 PM and arrives at 6:30 PM. From 3:00 PM to 6:00 PM is 3 hours.
From 6:00 PM to 6:30 PM is 30 minutes. So, the total flight duration is 3 hours and 30 minutes.

Question3: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question3?
Answer3:

"""
```

[Output]

```
Question1: Sarah has 100 dollars. She buys a book for 40 dollars and a pencil for 5 dollars.
How much money does Sarah have left?
Answer1: Let's go through this step-by-step.
Sarah starts with 100 dollars. After buying the book for 40 dollars, she has 100 - 40 = 60 dollars.
Then, after buying the pencil for 5 dollars, she has 60 - 5 = 55 dollars. So, Sarah has 55 dollars left.

Question2: John’s flight departs at 3:00 PM and arrives at 6:30 PM. How long is John’s flight?
Answer2: Let's go through this step-by-step.
John’s flight departs at 3:00 PM and arrives at 6:30 PM. From 3:00 PM to 6:00 PM is 3 hours.
From 6:00 PM to 6:30 PM is 30 minutes. So, the total flight duration is 3 hours and 30 minutes.

Question3: Kim has 3 boxes of apples, with each box containing 10 apples.
Kim gives 1 box of apples to Jung and eats 3 apples from one of the remaining boxes.
How many apples does Kim have now?

What is the answer of Question3?
Answer3:

Let's go through this step-by-step.
Kim has 3 boxes of apples, with each box containing 10 apples. So, Kim has a total of 30 apples (3 x 10 = 30).
Kim gives 1 box of apples to Jung, so Kim has 3 - 1 = 2 boxes of apples left.
Each box of apples contains 10 apples, so Kim has a total of 2 x 10 = 20 apples left.
Kim eats 3 apples from one of the remaining boxes, so Kim has 20 - 3 = 17 apples
```

- **Also, Model finally provided the correct answer!!!**
- **Compared to other prompts, it shows significantly better results**
- **Model accurately understood the intent of the question and generated the most appropriate answer, step by step, along with explanations as intended.**

4. [Compare result with Base model and Instruction fine-tuned model]

Base Model

- Base models are pre-trained on large datasets that cover a broad range of topics and tasks but are not specifically fine-tuned for particular instructions or objectives

- Output might not always align with the user's specific needs or instructions

- Since they are not instruction-tuned, their responses may lack clarity, precision, or relevance to a particular task

Instruction Fine-Tuned Model

- Models are better at interpreting complex instructions and providing more relevant and focused outputs

- With instruction tuning, the model can provide more accurate, contextually appropriate, and consistent responses compared to a base model

For commercial applications, instruction fine-tuned models are generally more suitable

- Instruction-tuned models offer more user-friendly and context-aware outputs, enhancing the overall user experience

- In tasks where efficiency is essential, instruction fine-tuning enables the model to follow directions more effectively, reducing the need for extensive post-processing or human intervention.

5. [Truble Shooting]

- **Import Llama2 Model :**
  We cannot use Llama2 the same way as GPT-2 (an open model). To use Llama2, we need to log in to Hugging Face, obtain a token, and use that token to import the Llama2 model.
- **Lack of Memory :**
  When running the Llama2 7B model, our GPU and RAM may not have enough memory to handle it. By setting torch_dtype=torch.float16, we can utilize the model efficiently on the given hardware, as it reduces the memory requirements.