IN4080 exercises (autumn 2023)

Week 10: Chatbots (continued)

Retrieval-based chatbots

If you have not finished some of the exercises from last week, you can go through them now.

Generative models

We will test the use of generative models using the transformers library from HuggingFace. If you have not downloaded the library, do it first (using *pip install* or *conda*). To test the library without requiring access to GPUs, we will use the slightly older GPT-2 model.

Tasks:

1. We first start with tokenization. Find how to load the GPT-2 tokenizer and run it to tokenize the following text: "I love the IN4080 course because".

2. Load the GPT-2 model using the *AutoModelForCausalLM* class.

3. Generate a continuation of that sentence with a maximum of 30 new tokens, using the generate function.

4. The default implementation for the *generate* function is greedy decoding. Try running beam search and showing the 5 best results instead of just one.
   NB: you need to set both *num_beams* and *num_return_sequences*.

5. Change the temperature and investigate what happens when you do so.

6. You might notice some repetition in the generated output. Look at the documentation of the *generate* function to find out how to prevent it.

Chatbot models (previous exam question)

Assume you wish to develop a chatbot to answer questions about the current time in various locations around the world. The chatbot should for instance be able to respond to queries such as "What is the current time in Buenos Aires?", "What is the

time difference between Boston and Oslo?" or "If it is 6:00 AM in Oslo, what time is it in Tokyo?".

Your first task is to decide what kind of chatbot development strategy you wish to follow. We have covered four alternative approaches during the course: handcrafted chatbots, IR-based chatbots, generative chatbots and NLU-based chatbots.

Answer the questions below:

1) Which type of approach (among the four approaches above) do you think is most suitable for this task? Motivate your choice in one or two paragraphs.

2) Which data will you need to collect to train/develop your chatbot, based on the approach chosen above? What type of annotations would you need to add to this data, if any? Explain in a few sentences.

3) Which system modules (machine learning models or rule-based components) would you need to integrate in your chatbot? Describe in one or two paragraphs the general processing pipeline of your chatbot.

Obligatory assignment 3

If you need help with the obligatory assignment, don’t hesitate to ask us!