FIT5217 (S1 2024) - Assignment 1

Marks	Worth 50 marks, and 25% of all marks for the unit
Due Date	Week 7 - Lecture Date, 11:55 PM
Extension	An extension could be granted under some circumstances. A special consideration application form must be submitted. Please refer to the university webpage on special consideration.
Lateness	For all assessment items handed in after the official due date, and without an agreed extension, a 10% penalty applies to the student's mark for each day after the due date (including weekends) for up to 7 days. Assessment items handed in after 7 days without special consideration will not be considered.
Authorship	This is an individual assessment. All work must be your own. All submissions will be placed through Turnitin. This makes plagiarism remarkably easy to identify for us.
Submission	All answers should be typed. For part 1 and 2, a single pdf report needs to be submitted (report page should be A4 size with standard margins and 11 point font). For part 3, you need to submit a CSV file, a power point file, and a 5 minute video presentation (see last part for details). The name of the files must be `Assignment_1_FIT5217_012345678.pdf` and `Assignment_1_FIT5217_012345678.csv` and `Assignment_1_FIT5217_012345678.pptx` and `Assignment_1_FIT5217_012345678.mp4` where "`012345678`" is replaced by your own student ID.

Table 1: Instructions for Assignment 1

Part 1 - POS Tagging (Total 10 Marks)

Consider the following HMM with three possible observations "snow", "fell", "storm" and three possible Part-of-Speech (POS) tags ("N", V", "J"):

State transition probabilities (A) – e.g., $a_{N,J} = P(s_{i+1} = J|s_i = N) = 0.1$

A	N (noun)	V (verb)	J (adj)
N	0.4	0.5	0.1
V	0.5	0.1	0.4
J	0.5	0.1	0.4

Emission probabilities (B) – e.g., $b_N(snow) = P(o_i = snow|s_i = N) = 0.4$

B	snow	fell	storm
N	0.4	0.2	0.4
V	0.3	0.5	0.2
J	0.2	0.4	0.4

Initial state distributions (π) – e.g., $\pi[J] = P(s_1 = J|s_0 =< S >) = 0.3$

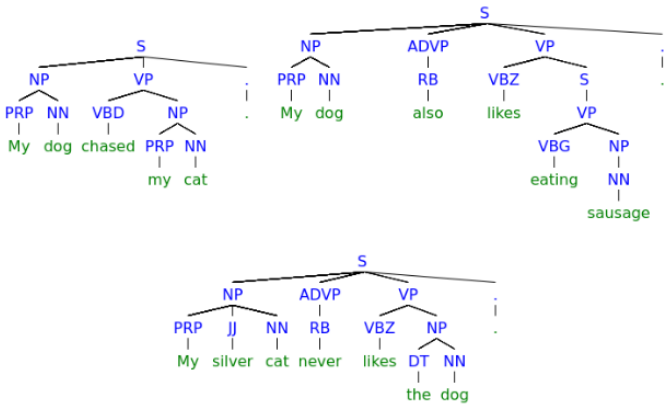
	N	V	J
π	0.4	0.3	0.3

Question 1.1. Draw the forward trellis diagram of the sentence "snow fell" using the given HMM. Clearly show the forward arrows to illustrate the computation of the forward algorithm. (5 marks)

Question 1.2. Find the most likely state sequence for the sentence "snow fell" using the Viterbi algorithm (show your steps clearly). What is the joint probability of the sentence "snow fell" and its most likely state sequence? (5 marks)

Part 2 - PCFG (Total 10 Marks)

Consider the following parse trees:



Question 2.1. Convert the grammar rules into the CNF format and list them. Then estimate their corresponding probabilities (based on all trees) using Maximum Likelihood estimation. (5 Marks)

Note. Structure you response as one rule per line:
[0.2] A → B C
[0.1] D → E
...
where A → B C is the grammar rule, and [0.2] is its probability.

Question 2.2. Using the estimated probabilities and the CKY parsing algorithm, calculate the probability of the following sentence (and report the final CKY chart): (5 Marks)

`The dog chased my silver cat .`

Part 3 - Large Language Models (Total 30 Marks)

For this assignment you need to create accounts to use OpenAI's ChatGP (`https://chat.openai.com/chat`), Cohere's Chat Only LLM (`https://cohere.com/`), and Anthropic's Claude3-opus (`https://claude.ai/chats`). These are 3 well-established commercial LLMs on the market.

In this part of the assignment you need to work on the following 2 categories of reasoning, and come up with a total of 5 reasoning questions (could be: short answer/essay, reading comprehension, or multiple choice questions) for which ChatGPT fails. All 5 questions could belong to 1 category. The 2 categories and examples per each category:

- Category 1: Numerical/Logical Reasoning - Examples of questions:
  - *If the zookeeper had 100 pairs of animals in her zoo and if two pairs of babies are born for each and every one of the original animals, and then sadly 23 animal don't survive, how many animals do you have left in total?*
- Category 2: Commonsense/Physical/Temporal Reasoning - Examples of questions:
  - *I landed on the planet gooblygoob9m2, my flootenwooten slipped off my hand, and hit the ground. Does gooblygoob9m2 have gravity?*
  - *There is an apple inside a blue box. There is also a red box inside the blue box. The red box has a lid. How can I get the apple?*

Once you found these 5 examples which make ChatGPT fail, you need to try them again with Cohere and Anthropic LLMs and assess whether these two other LLMs fail or succeed at your reasoning questions.

Important Note. When interacting with LLMs, each question should be typed into a New Chat (i.e., do NOT type more that 1 question into a single chat session). If you do not know why this is necessary, ask the CE.

Submission format for part 3 : For this part of the assignment, there is no designated report section. You are required to submit three specific files. Two template files provided with the assignment - a CSV template and a PPTX template - must be completed according to their structure and specified content in below. Additionally, you are required to create and submit a 5-minute video presentation that comprehensively covers the content outlined in the PPTX template. This video, along with the completed CSV and PPTX files are the 3 files that should be uploaded to Moodle for part 3.

Mark break down for part 3 :

- For each question you will get 3 marks: 1 mark for the question, 1 mark for the analysis of ChatGPT failure on the question, and 0.5 mark for providing the results from Anthropic's Claude3-Opus LLM, and 0.5 mark for providing the results from Cohere LLMs. This adds up the mark to a total of 15 marks. The results of LLMs from Cohere and Anthropic do not require any analysis. These will provide the content of the PPTX and CSV.
- The questions should NOT come from the internet or published articles, but if you were inspired by a resource cite[1] it. Variations of the same question, even with minor modifications, are NOT acceptable. Additionally, questions should have objective responses, subjective answers are NOT acceptable.
- The analysis of ChatGPT failure must be grounded on the published works in the literature and have proper citation of relevant papers.
- The presentation and the 5 minutes recording is worth 15 Marks. The presentation needs to be clear and easy to understand, and address the points requested in the pptx template. Recordings above 5 minutes or below 4.5 minutes are NOT acceptable.

Useful Readings To better be prepared for the analysis part of the failed cases, you may want to scan the following works and cite them or any other work that supports your speculation for the ChatGPT's behaviour.

- Benchmarks for Automated Commonsense Reasoning: A survey
- LogiQA: A Challenge Dataset for Machine Reading Comprehension with Logical Reasoning
- HellaSwag: Can a Machine Really Finish Your Sentence?
- PIQA: Reasoning about Physical Commonsense in Natural Language
- TruthfulQA: Measuring How Models Mimic Human Falsehoods
- WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale
- Online Resource: Gary Markus and Ernest Davis Experiments testing GPT-3's ability at commonsense reasoning: results.
- Training Verifiers to Solve Math Word Problems
- Online Resource: Google Big Bench
- Language Models are Few-Shot Learners, sec 3.5
- PaLM: Scaling Language Modeling with Pathways, section 6
- Sparks of Artificial General Intelligence: Early experiments with GPT-4
- Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4

[1] Use Chicago Citation Style: https://guides.lib.monash.edu/citing-referencing