FIT5217 (T3, 2022) - Assignment 1

Marks	Worth 80 marks, and 20% of all marks for the unit
Due Date	Friday, 15 July 2022, 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. A single pdf file needs to be submitted. The name of the file must be `Assignment_1_FIT5217_012345678.pdf` where "`012345678`" is replaced by your own student ID. The pages should be A4 size with standard margins and 11 point font or similar.

Part 1 - Language Model (Total 20 Marks)

Take the corpus which contains the following three sentences - pad the sentences with start and end tags (only once) then answer the following questions:

```
YI READ BALL LIGHTNING
JINMING READ A DIFFERENT BOOK
SHE READ A BOOK BY LIU
```

Question 1.1. List all the unigrams, bigrams and trigrams. (4 Marks)

Question 1.2. Using a bigram language model and maximum likelihood estimation (from the above corpus), calculate the probability of each of the following 2 sentences. List all the bigram probabilities as well as the final probabilities of the sentences. (8 Marks)

```
YI READ A BOOK
LIU READ A BOOK
```

Question 1.3. Similar to the above question, using a bigram language model and add-1 Smoothing, calculate the probability of each of the following 2 sentences. List all the bigram probabilities as well as the final probabilities of the sentences. (8 Marks)

```
YI READ A BOOK
LIU READ A BOOK
```

Part 2 - POS Tagging (Total 15 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 2.1. Explain how to obtain the values in the matrices A, B and π given above in a supervised manner. (3 marks)

Question 2.2. 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. (4 marks)

Question 2.3. 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? (4 marks)

Question 2.4. What are the two main assumptions that are built into an HMM? What is the major downside if we construct an HMM without (or relaxing) these assumptions? You may need to read the text book for this. (4 marks)

Part 3 - Syntactic Parsing (Total 10 Marks)

Question 3.1. Answer with True or False (you need to do a bit of research for some of these questions):

a. A lexicalized PCFG is the PCFG which does not have any non-terminals on the right-hand-side of the grammar rules.

b. The Inside-Outside algorithm is a version of Expectation-Maximisation.

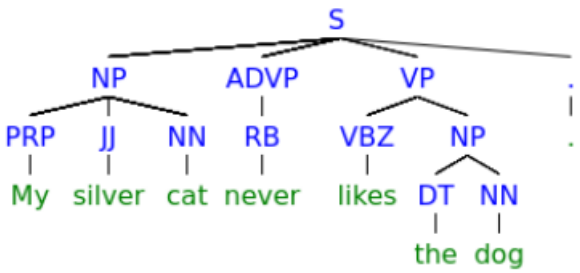
c. The Inside-Outside algorithm is used for unsupervised learning of a PCFG.

d. CKY algorithm has a complexity of $O(n^3)$ where n is the length of the input sentence.

e. Tree Adjoining Grammars (TAGs) are more expensive to parse compared to Context Free Grammars (CFGs).

Part 4 - Chomsky Normal Form (Total 5 Marks)

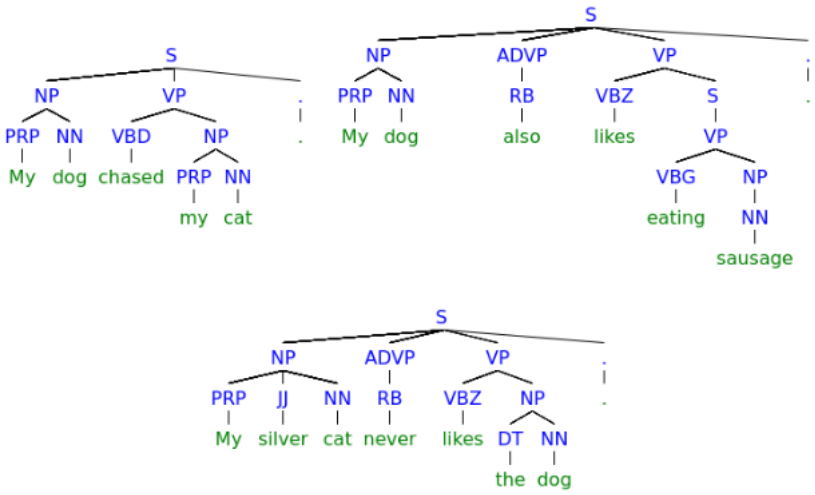
Consider the following parse tree:



Question 4.1. Write down all the grammar rules extracted from the parse tree. Then convert the rules into Chomsky Normal Form (CNF).

Part 5 - PCFG (Total 15 Marks)

Consider the following parse trees:



Question 5.1. List all the grammar rules and estimate their corresponding probabilities (based on all trees) using Maximum Likelihood estimation. (8 Marks)

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

Question 5.2. Using the estimated probabilities, calculate the probability of the following sentence: (8 Marks)

```
The dog chased my silver cat .
```

Hint. Need to do CNF in 5.1 before estimating the rules probabilities, then apply CKY in 5.2 to get all parse trees for the sentence (no need to include this in your response), then use the grammar rules and their probabilities to calculate the probability of the sentence (sum of the probabilities of its trees).

Part 6 - Short Essay (Total 15 Marks)

Read the following paper, then write a short essay, no more than 2 paragraphs, summarising:

- Problem Statement
- Motivation and Novelty
- Method
- Key Results

Paper: Pauls, Adam, and Dan Klein. Large-scale syntactic language modeling with treelets. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.