UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11125 ACCELERATED NATURAL LANGUAGE PROCESSING

Friday 18 th December 2015

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer any TWO questions.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: F. Keller
External Examiners: A. Burns, S. Denham, P. Healey, T. Norman

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **Words and tags**

(a) Consider the sentence $S =$ `<s> the cat chased the dog </s>`.

i. Suppose we use a *unigram* language model to compute sentence probabilities. Give a different sentence with the same unigram probability as S. [*2 marks*]

ii. Now suppose we use a *bigram* language model. The following table gives smoothed probability estimates for $P(w_i|w_{i-1})$ under this model, with w_{i-1} labelling the rows and w_i the columns.

	the	a	cat	dog	girl	chased	</s>
<s>	0.0067	0.0067	0.00056	0.0002	0.00007	0.0045	0
the	0.00012	0.00009	0.03	0.02	0.02	0.005	0.00008
a	0.00012	0.00009	0.05	0.01	0.03	0.00041	0.00009
cat	0.006	0.0081	0.00008	0.00009	0.0004	0.003	0.003
dog	0.0081	0.006	0.00008	0.00009	0.0004	0.004	0.003
girl	0.0016	0.007	0.00005	0.00013	0.00022	0.0061	0.003
chased	0.024	0.006	0.00041	0.0001	0.00037	0.00002	0.0003

A. Is this table complete, or does the model also include probabilities for bigrams that aren't shown in this table? Justify your answer. [*2 marks*]

B. Give a different sentence that has the same probability as S under this bigram model. (*Hint:* you should not need to perform arithmetic to answer this question.) [*4 marks*]

(b) Suppose we want to estimate the parameters of an HMM using add-α smoothing.

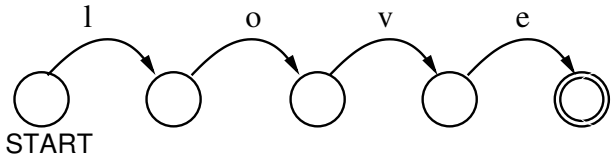
i. Give the formulas we would use to estimate the transition probabilities and emission probabilities, and say what each term in the formulas refers to. [*4 marks*]

ii. In class we discussed some problems with add-α smoothing. For an HMM, where would these problems be worse: for estimating the transition probabilities, or for estimating the emission probabilities? Justify your answer. [*4 marks*]

(c) In social media such as Twitter, users sometimes emphasize words by repeating letters. For example, a user might write `I llooooovveee it!` to emphasize the word `love`. However, this non-standard spelling can present problems for NLP systems. This question considers some ways to deal with the problems.

QUESTION CONTINUES ON NEXT PAGE

i. The diagram below shows a finite-state automaton that recognizes the word `love`. Convert this to a finite-state *transducer* that corrects non-standard to standard spelling for the word `love`. Assume that the non-standard spelling contains the characters `l`, `o`, `v`, `e` in that order, but might have more than one of each character. Your FST need not be deterministic. [*3 marks*]



ii. Suppose you designed similar transducers for *all* English words in order to convert emphasized (repeated letter) versions of words into standard spelling. What problem(s) might you have with the results? Outline how you might try to solve these problem(s) using other methods from the course, assuming you have access to the full text where the words are used. Include specific examples in your answer where possible, and discuss whether your solution requires other resources beyond those already mentioned, and any other pros/cons of your solution. [*6 marks*]

2. **Grammar and parsing**

(a) Context-free Phrase Structure grammars (CFGs)
What is **attachment ambiguity**?

Explain the term, identifying and briefly describing two aspects of the syntax of English where attachment ambiguity arises.

Include at least one pair of examples to illustrate each of your points, giving rules and trees.

Note that this is *not* a question about statistical grammars, or about processing or ranking alternative analyses. Nor should you make any attempt at attaching explicit semantics to your trees. [*5 marks*]

(b) CFG parsing

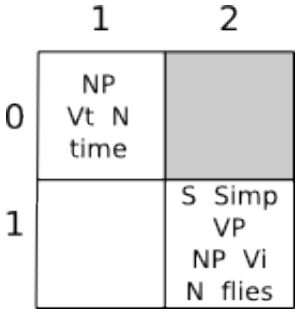
i. Very briefly define **dynamic programming**. How can it be used to address a major source of inefficiency in **recursive descent parsing** of context-free grammars? Explain what is meant by a (passive) **well-formed substring table** and how it addresses that inefficiency. [*3 marks*]

ii. What does an **active** chart parser add to the well-formed substring table? [*2 marks*]

iii. By what means can an active chart parser be configured to search for parses **depth-first**? To search **breadth-first**? [*2 marks*]

QUESTION CONTINUES ON NEXT PAGE

(c) Cocke-Kasami-Younger (CKY) parsing
The figure below shows an intermediate state of a CKY parse of the ambiguous string "time flies".



i. What are the rules that must be in the grammar used in this parse so far, and to get it to a successful conclusion with two analyses of the whole string as **S**, one with a structure which would also be appropriate for "open the door!" and the other appropriate for "water evaporates"? [*3 marks*]

ii. What happens next? That is, what gets inserted into the upper-right-hand corner? Why? That is, explain in each case how the CKY algorithm determines what gets added. [*3 marks*]

Note that you are *not* required to stick to Chomsky Normal Form for your rules. That is, you may use any symbols on the right-hand side of rules, as long as there are only one or two of them.

(d) Probabilistic context-free grammar (PCFG)

i. What is the inescapable weakness of a PCFG constructed from a treebank such as the Penn treebank, which uses simple atomic non-terminals, with respect to attachment ambiguity? Illustrate your answer with parse trees for two simple noun-phrases with the the pre-terminal tags `Adj Nom Conj Nom`. [*4 marks*]

ii. What is **pointwise mutual information**? How might it be used to address the above-mentioned weakness? Refer to your examples in your answer. You do *not* need to provide the precise equation for pointwise mutual information, but you should explain what that equation captures in terms of probabilities. [*3 marks*]

3. **Semantics and discourse**

(a) Suppose we have the following grammar fragment with semantic attachments. Grammar rules are numbered so you can refer to them easily.

	Rule	**Sem. attachment**		**Rule**	**Sem. attachment**
1.	Det →`a`	$\lambda P.\lambda Q.\exists x.P(x) \wedge Q(x)$	6.	Nom →N	N.sem
2.	N →`green`	$\lambda x.green(x)$	7.	Nom →Adj Nom	Adj.sem(Nom.sem)
3.	Adj →`green`	$\lambda P.\lambda x.green(x) \wedge P(x)$	8.	NP →Det Nom	Det.sem(Nom.sem)
4.	N →`flower`	$\lambda x.flower(x)$	9.	VP →V	V.sem
5.	V →`flower`	$\lambda x.\lambda y.flower(x, y)$	10.	VP →V NP	V.sem(NP.sem)

i. Is the meaning representation language in this grammar *compositional*? Explain why or why not. [*2 marks*]

ii. Which rules are needed to parse the phrase `a green flower`? [*1 mark*]

iii. Show how the meaning representation for `a green flower` is obtained. You can explain each step, show a parse tree, or use some combination of those to illustrate the method. [*4 marks*]

(b) Suppose you want to build a sentiment analysis system to classify reviews as either positive or negative.

i. Explain how you could use a Naive Bayes classifier to solve the task. Your answer should include what features you would use, how you would train the model, and how you would classify a new document d at test time. [*5 marks*]

ii. Do you think it would be better to evaluate your system using *accuracy* or *precision and recall*? Why? [*2 marks*]

QUESTION CONTINUES ON NEXT PAGE

(c) Discourse structure

The brief passage below is the beginning of a lightly-edited transcript of a verbal description of a short (silent) film.

1 It opens with I guess a farm worker, picking pears, in a tree.
2 And you see him picking the pears off the leaves, and putting them in a white apron, and he walks down the ladder, and dumps the pears into a basket.
3 Then you see him going back up in the tree.
4 And you see a guy leading a goat, past the tree where he's picking the pears.
5 Then a little boy on a bicycle, comes riding past the tree, and sort of goes past the pears in the baskets and then stops and looks up at the guy in the tree, he's still on the ladder, and hes not watching him, so he puts his bike down, he walks over, and he picks up the whole basket of pears, and puts it on the front fender of his bike, and holds on to it and he takes off with them.

i. Define **cohesion** and **coherence** and **coreference chain**, making their relation to one another clear. List all the coreference chains which *end* in sentences 4 and 5, including the ones which begin and end in a single sentence. [*5 marks*]

ii. Propose a grouping of the sentences in this passage into sub-discourses. Describe the reasons for your choice. Relate those reasons to the terms defined in the previous question. (There is no single 'correct' grouping—you will be assessed on your discussion of your choice, not on the choice itself) [*3 marks*]

iii. Pick one of the discourse structure/coreference approaches discussed in class, and discuss how well you think it would succeed with one or two specific aspects of this text. [*3 marks*]

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11125 ACCELERATED NATURAL LANGUAGE PROCESSING

Wednesday 14th December 2016

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

Answer any TWO questions.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: F. Keller
External Examiners: A. Burns, P. Healey, M. Niranjan

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **Words and tags**

(a) How many **types** and how many **tokens** are in the following sentence?

`the cat looked at the other cats` [*2 marks*]

(b) What **affixes** are in the word "reactors"? Identify whether each affix is *derivational* or *inflectional*. [*2 marks*]

(c) Explain what is meant by the **bag-of-words assumption**. Give an example of a probabilistic model that makes this assumption and a task for which that model might be appropriate. [*3 marks*]

(d) Suppose I have a large corpus, and I make a list of all the words that occur **exactly once in the first half** of the corpus. I then count how many times each of those words occurs in the **second half** of the corpus. Which of the following should I expect to find? (Write the correct letter in your exam book, an explanation is not needed.)

A. On average, these words occur *less than once* in the second half.

B. On average, these words occur *more than once* in the second half.

C. On average, these words occur *exactly once* in the second half.

D. There is no way to predict; any of the above are equally likely. [*2 marks*]

(e) Suppose I'm building **5-gram language models over characters**. I build two models using the same corpus: one uses *add-α smoothing*, and the other uses *backoff smoothing*. If _ represents the space character, and the sequence `xing_` never occurs in the corpus, which model is likely to assign a higher value to $P(\_|\texttt{xing})$, and why? [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(f) Suppose we are **tagging named entities** using an **HMM**. We want to tag each word as either `PER` for persons, `ORG` for organizations, or `OTH` for all other tokens. We also have special start- and end-of-sentence markers `<s>` and `</s>` in both the tag sequences and word sequences.

Below are the transition matrix and part of the output matrix for the HMM:

$t_{i-1} \backslash t_i$	PER	ORG	OTH	</s>
<s>	0.6	0	0.4	0
PER	0.5	0.05	0.45	0
ORG	0	0.7	0.3	0
OTH	0.02	0.03	0.9	0.05

$t \backslash w$	Mr	Craft	David	winked	...
PER	1×10^{-2}	3×10^{-3}	8×10^{-3}	3×10^{-8}	
ORG	2×10^{-3}	7×10^{-4}	6×10^{-4}	0	
OTH	0	3×10^{-7}	0	5×10^{-7}	

i. If $\vec{t}$ = `<s> PER ORG OTH </s>` and $\vec{w}$ = `<s> Mr Craft winked </s>`, what is $P(\vec{t}, \vec{w})$ for this HMM?
You should give the general expression for computing this probability, then fill in the correct values. You do not need to reduce your answer to a single number. So, answers of the form $(.5)(.8)/(.3)$ or $(.3)(.6) + (.7)(.1)$ are fine. [*3 marks*]

ii. Below is a partly completed chart computed using the **Viterbi algorithm**:

	<s>	Mr	Craft	winked	</s>
<s>	1	0			
PER	0	6×10^{-3}			
ORG	0	0			
OTH	0	2×10^{-9}			
</s>	0	0			

Write down the computation that needs to be done in order to fill in the single cell [`PER`, `Craft`] using the Viterbi algorithm. What does the numerical value in this cell represent, and where will the backpointer point? [*4 marks*]

QUESTION CONTINUES ON NEXT PAGE

(g) This question is about handling **unseen words** in sequence models. As in question 1(f),we want to tag named entities using the `PER`, `ORG`, and `OTH` tags.

The following sentences occur in our test set, and the words `Xfinity`, `GBX`, `Altoona`, `Slovensky`, and `semiconscious` have not been seen in training:

```
<s> Xfinity is going further . </s>
<s> GBX acquired Altoona Corporation for $100 million . </s>
<s> Mr Slovensky looked at the semiconscious attendees . </s>
```

i. Consider the following method for handling unseen words using an HMM. In the training corpus, use the new token `UNK` to replace all words that occur exactly once, and then estimate probabilities as usual. To tag the test sentences, temporarily replace any unseen word with `UNK`, and then run the Viterbi algorithm. In the output, change each `UNK` back to its original word. Using this method, which of the unseen words in the example sentences seem most likely to be tagged correctly, and why? (That is, what information might lead the HMM to predict the correct tag?) [*3 marks*]

ii. Now consider using a discriminative sequence model instead, so we can include arbitrary features. Describe two features you think could improve the model's performance, and explain why you think so, with reference to the examples above and your knowledge of English. [*3 marks*]

2. **Grammar and parsing**

(a) Context-free Phrase Structure grammars (CFGs)

Consider the following joke by Groucho Marx:

One morning I shot an elephant in my pajamas.
How he got into my pajamas I don't know.

i. Sketch two alternative trees and accompanying grammar rules for the sentence "I shot an elephant in my pajamas" which might be used to explain the joke quoted above. [*3 marks*]

You can assume the following rules to start with, but will need to add more of your own:

N $\longrightarrow$ elephant | pajamas
D $\longrightarrow$ an | my
P $\longrightarrow$ in
NP $\longrightarrow$ D N
PP $\longrightarrow$ P NP
S $\longrightarrow$ NP VP

You do *not* need to use Chomsky Normal Form for your rules. Also, this is *not* a question about statistical grammars, or about processing or ranking alternative analyses. Nor should you make any attempt at attaching explicit semantics to your trees.

ii. Briefly explain the aspects of grammar discussed in the course that are illustrated by the first sentence in the above joke and your trees. [*2 marks*]

iii. Explain the term **compositional semantics**, using the quote above and your analysis to illustrate. Include paraphrases of the meanings associated with your two trees and explain them by reference to your grammar rules and your explanation of compositionality. [*5 marks*]

(b) Active Chart Parsing

i. Describe in detail bottom-up depth-first (i.e., LIFO) active chart parsing using context-free phrase-structure grammars. Include descriptions of:

- The chart and the agenda, and what is stored in each
- The way the grammar and input are used
- The processing rules which create edges: that is, the bottom-up rule and the fundamental rule
- The overall process of parsing: how it starts, how the agenda and chart interact, how it finishes

[*5 marks*]

ii. Contrast the top-down and bottom-up versions of active chart parsing. How does an active chart parser (even when running top-down) avoid the left-recursion trap that a simple recursive-descent parser is vulnerable to? [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(c) Probabilistic context-free grammar (PCFG)

i. The Penn treebank uses simple atomic symbols as non-terminal node labels. What is the inescapable weakness of a PCFG constructed from this kind of treebank with respect to attachment ambiguity? Illustrate your answer with parse trees for two simple noun-phrases with the pre-terminal tags `Adj Nom Conj Nom`. [*4 marks*]

ii. What is **pointwise mutual information**? How might it be used to address the above-mentioned weakness? Refer to your examples in your answer. You should provide the equation for pointwise mutual information, and explain what that equation captures in terms of probabilities. [*3 marks*]

3. **Semantics and discourse**

(a) What is meant by the **distributional hypothesis** in lexical semantics? [*2 marks*]

(b) Fill in the blanks: (Write the answers in your script book, not on this paper!)

i. In terms of **lexical semantic** relationships, "fruit" is a _____ of "apple". [*1 mark*]

ii. In **WordNet**, "buy" and "purchase" belong to the same _____ because in many contexts, they are _____. [*1 mark*]

(c) Suppose you want to build a **sentiment analysis** system for **movie reviews**. That is, given the text of a movie review, your system should classify the review as either positive or negative.

Describe how you would build such a system. Your description should include the following aspects:

- What data or other resources would you need, how you might get them, and how you would use them.
- What model or algorithm you would use, and how you would apply it in this situation. Your description should be high-level and does **not** need to include equations.
- How you would evaluate the results of the system.
- What, if any, ethical issues might arise in developing the system.

In some cases there might be multiple options for each choice. If so, mention briefly the **strengths and weaknesses** of each approach, or say what additional information you might need in order to decide which approach is best. [*8 marks*]

QUESTION CONTINUES ON NEXT PAGE

(d) **Discourse structure**

The brief passage below is the part of a lightly-edited transcript of a verbal description of a short (silent) film. In the preceding scene a boy sees a basket of pears at the edge of a road running past an orchard.

1 The boy picks up the whole basket of pears, and puts it on the front fender of his bike and takes off with them.
2 Then a girl on a bicycle comes riding towards him, in the opposite direction.
3 And as they pass, he sort of turns, and looks at her, and his hat flies off, and his bicycle hits a rock...
4 Because he's looking at the girl.
5 He falls over, and then these three other little kids about his same age come walking by.
6 One guy has a paddle with a ball attached with a string.
7 And they see that he's fallen off his bike, and his pears have scattered.
8 And they walk over to help him, they gather all the pears and put them in the basket, and one of the guys helps him brush off the dust, and another guy picks up the rock, and he throws it out of the road.
9 And he gets all situated again, and he takes off, and the boys keep walking in their direction.

i. Define **cohesion** and **coherence** and **coreference chain**, making their relation to one another clear. List all the coreference chains which *end* in sentences 8 and 9, including the ones which begin and end in a single sentence. [*6 marks*]

Show sentence boundaries with vertical bars in the chains. For example, there's a chain ending in sentence 4 that would be reported like this:

girl | they - her | girl

ii. Propose a grouping of the sentences in this passage into sub-discourses. Describe the reasons for your choice. Relate those reasons to the terms defined in the previous question. (There is no single 'correct' grouping—you will be assessed on your discussion of your choice, not on the choice itself) [*4 marks*]

iii. Pick one of the discourse structure/coreference approaches discussed in class, and discuss how well you think it would succeed with one or two specific aspects of this text. [*3 marks*]

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11125 ACCELERATED NATURAL LANGUAGE PROCESSING

Friday 8 th December 2017

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY. If both QUESTION 2 and QUESTION 3 are answered, only QUESTION 2 will be marked.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: G. Sanguinetti
External Examiners: W. Knottenbelt, M. Dunlop, M. Niranjan, E. Vasilaki

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. THIS QUESTION IS COMPULSORY

(a) What is the key feature of a **dynamic programming** algorithm? (Write the letter corresponding to the best answer in your exam book) [*2 marks*]

A. It uses a chart to store partial results so that for inputs of size n, it runs in $O(n^2)$ time.

B. It stores information in a chart so that the computation can be dynamically updated over time.

C. It avoids redundant computation by storing partial results in a chart and combining them to form the full solution.

D. It improves accuracy by using a chart to filter out incorrect results.

E. It fills a chart bottom-up to avoid the left recursion problem and produce a solution efficiently.

(b) **HMMs** are sometimes used for *chunking*: identifying short sequences of words (chunks) within a text that are relevant for a particular task. For example, if we want to identify all the person names in a text, we could train an HMM using annotated data similar to the following:

> On/**O** Tuesday/**O** ,/**O** Mr/**B** Cameron/**I** met/**O** with/**O** Chancellor/**B** Angela/**I** Merkel/**I** ./**O**

There are three possible tags for each word: **B** marks the beginning (first word) of a chunk, **I** marks words inside a chunk, and **O** marks words outside a chunk. We also use **SS** and **ES** tags to mark the start and end of each sentence, respectively.

Crucially, the **O** and **SS** tags may not be followed by **I** because we need to have a **B** first indicating the beginning of a chunk.

Answer the following questions about this HMM. [*4 marks*]

i. Write down an expression for the probability of generating the sentence `Henry saw Barack Obama` tagged with the sequence **B O B I**.

ii. What, if any, changes would you need to make to the Viterbi algorithm in order to use it for tagging sentences with this **BIO** scheme? How can you incorporate the constraint on which tags can follow which others?

(c) Consider the trigrams below. (*Note:* a *private eye* is slang for a private detective.)

A. `private eye maneuvered`
B. `private car maneuvered`

Suppose that neither of these trigrams has been observed in a particular corpus, and we are using **backoff** to estimate their probabilities. What are the bigram probabilities that we will back off to in each case? In which case is the backoff model likely to provide a more accurate estimate of the trigram probability? Why? [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(d) Answer the following questions about this ambiguous sentence:

`I ate the fish in the freezer` [*7 marks*]

i. Give **unambiguous paraphrases** of each of the two meanings and say which of them is the more likely meaning.

ii. Look at the three **syntax trees** shown in Figure 1 on the following page. Which of them (A, B, or C) best corresponds to the likely meaning of the sentence? (Write the correct letter in your exam book.)

iii. Is the ambiguity in this sentence also reflected in its dependency structure? If so, give the two alternative **labelled dependency parses** and say which corresponds to the likely meaning. If not, explain why not.

(e) Suppose I want to classify restaurant reviews into those that complain about the restaurant's service and those that do not. I evaluate two different **text classification** systems on the same dataset of reviews. Compared to the gold standard, system A gets a precision of 92% and system B gets a precision of 94%. I then claim that system B is better than system A for my task.

Give two reasons why you shouldn't believe my claim (unless I provide further evidence or explanation). [*3 marks*]

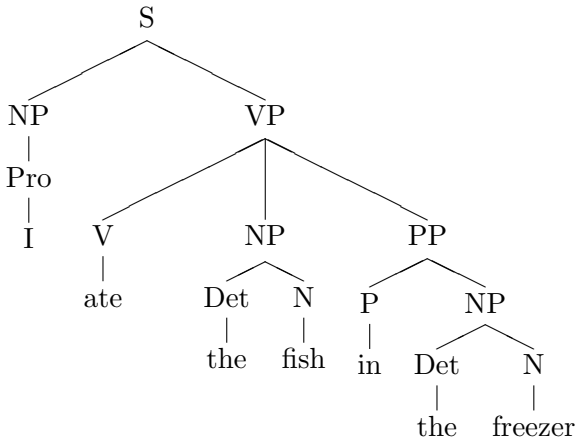
(f) The following table shows which of five different 50-word documents contain each of two words, x and y. For example, the ✓ under document 1 for word x means that x occurred (at least once) in document 1.

doc:	1	2	3	4	5
x:	✓	✓	✓	✓	
y:	✓		✓	✓	

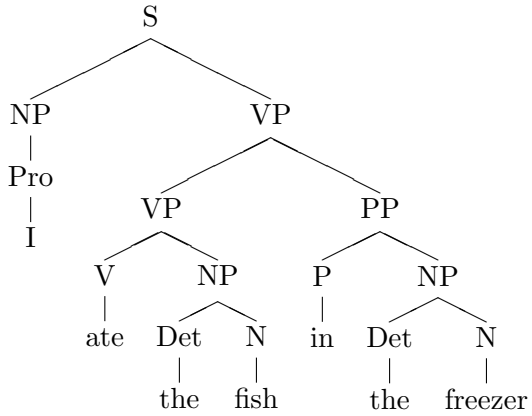
Based on this data, what is the **pointwise mutual information** (PMI) between the events "x occurs in a document" and "y occurs in a document", assuming maximum-likelihood estimation of probabilities? You should write down the general expression for PMI and plug in the appropriate numbers, but you do not need to reduce your answer to a single numerical value. [*3 marks*]

(g) Give a **semantic representation** for the sentence `Alex watched Hamlet yesterday` using first-order logic with event variables. Using this example, explain the advantage of using event variables with respect to identifying entailment relationships. [*3 marks*]

A.



B.



C.

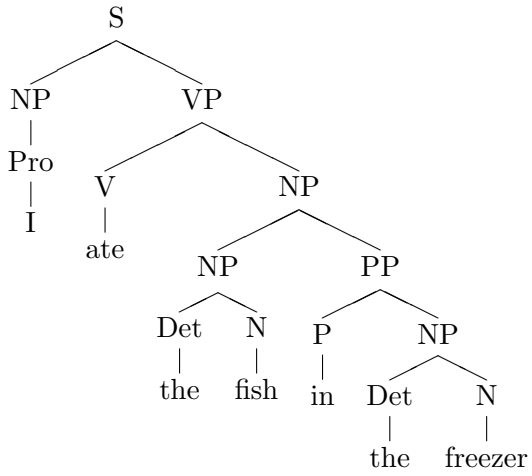


Figure 1: The trees used for question 1(d)ii and question 3a.

2. ANSWER EITHER THIS QUESTION OR QUESTION 3

This question addresses **text normalization, logistic regression, annotation**, and **evaluation**.

Parts (a)–(c) ask about how you could develop NLP tools to help historians and social scientists study historical documents. NLP tools developed on modern text have difficulty with these documents because historical words may have inconsistent spellings, or spellings that differ from their modern forms. One way to address this problem is to **normalize** the text before doing further processing.

Here is an excerpt of a historical text (1a) and a normalized version of it (1b):

(1a) `Myn adversarie is become bysshop of Cork in Irland, and ther arn ii other persones provided to the same bysshopriche yet lyvyng, beforn my seyd adversarie;` (William Paston, 1426)

(1b) `My adversary has become bishop of Cork in Ireland, and there are two other persons provided to the same bishopric yet living, before my said adversary;`

And here is a simple method to normalize historical text:

Dictionary method:

1	Look up each historical word token h in a modern dictionary.
2	If h appears in the dictionary, leave it as is.
3a	Else find the dictionary word d that has the most similar spelling to h, and replace h with d.

(a) If you implement the dictionary method above, what well-known **algorithm** could you use to determine how similar the spellings of each word pair are? (You do not need to explain how the algorithm works, just give its name.) [*2 marks*]

(b) Sometimes, d will not be the right normalization for h, so the dictionary method will produce an error. A possible improvement might be:

N-best method: Follow steps 1 and 2 as above, but replace 3a with:

3b	Else generate a list of the N dictionary words that are spelled most similarly to h. For each of the N words, use a **logistic regression** model to compute the probability that this word is the correct normalization, and pick the most probable one.

i. Give two different types of **features** that would be important to include in the model, and say why. [*4 marks*]

QUESTION CONTINUES ON NEXT PAGE

ii. What sort of **training data** would you need to train your model? [*2 marks*]

iii. Would you also need a **development set**? If so, explain what you would use it for in this particular context. If not, explain why not. [*2 marks*]

(c) Describe *two other* types of **errors** that can occur with the dictionary method, which can *not* be solved by the N-best method. Use examples from the provided text to illustrate each type. Pick *one* of these error types and propose a way you could try to solve it. Discuss any new difficulties or weaknesses your method might introduce. [*5 marks*]

Like historical text, Twitter data often contains non-standard spellings, and some researchers have suggested normalization as a way to improve performance of downstream NLP systems.

(d) Consider the following two tweets, and imagine you are trying to get annotators to produce "normalized" versions of them.

(1) `@SwizzOnaRampage lol no comment bro... can't say if I disagree or agree. lol`

(2) `Really wish 1 of the #sixxtards won. Haha mayb a book? Wld b my 3rd copy. My 1st copy got worn out had to buy a new1`

Give two examples of difficult **annotation decisions** from this data: that is, tokens where annotators might disagree about the correct normalized form unless very specific guidelines are given. For each example explain briefly why it is difficult (i.e. provide two or more alternative normalizations and say why each might be reasonable). [*4 marks*]

Training an NLP system often involves some **random choices**, for example random initialization of the parameters. The rest of this question addresses this randomness and its implications.

(e) Give two examples of models or algorithms used in NLP where different **initialization** during training may lead to differences in the trained systems, because the training objective has multiple local optima. [*2 marks*]

(f) Suppose you implement a change to a system where different initializations matter. You want to compare the original system to the system with your change. You train each version with 10 different random initializations, and evaluate the trained systems on a development set. Answer the following questions: [*4 marks*]

i. A friend suggests you use a **t-test** to assess whether there is a statistically significant difference between the performance of the two systems on the development set. What assumptions does this test make, and are these assumptions justified for your data?

ii. Name a method or tool that would be appropriate for assessing the likely size of any performance difference between the two systems.

3. ANSWER EITHER THIS QUESTION OR QUESTION 2

This questions addresses **Syntax, semantics**, and **parsing**.

(a) Look again at the parse trees in Figure 1. Assume we have a grammar with all the rules needed to generate these trees. Which rule(s) used in these trees could cause non-termination if we tried to parse with a **recursive descent parser**? [*2 marks*]

(b) Answer the following questions about the same sentence used in Question 1d:

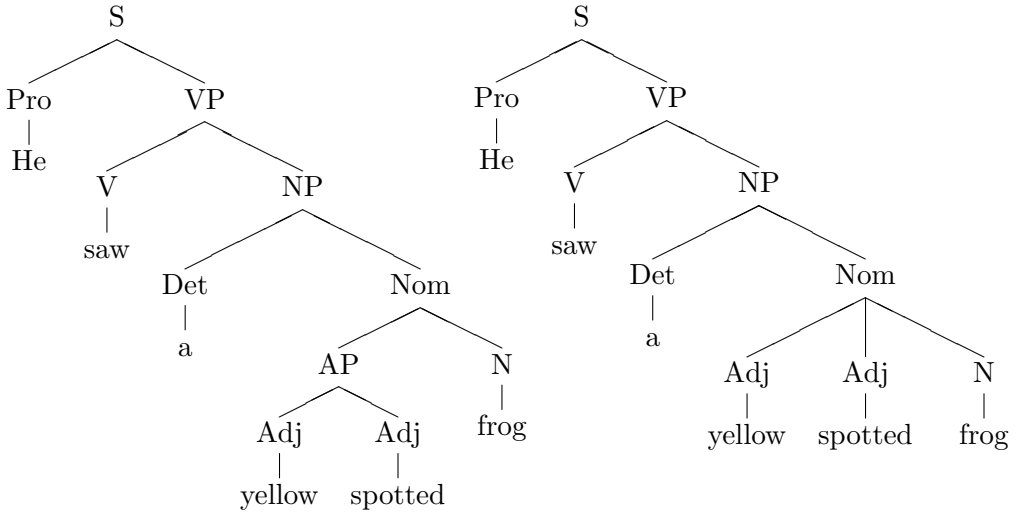
`I ate the fish in the freezer` [*7 marks*]

i. Give an example of a contrasting sentence that illustrates the fundamental weakness of vanilla PCFGs for **attachment disambiguation**, and explain what that weakness is, with reference to your contrast pair. You can refer to the tree structures in Figure 1 by letter (A, B, C) if you wish.

ii. What **extension** to the vanilla PCFG model could potentially disambiguate both sentences correctly? Justify your answer using syntactic trees and/or rules as appropriate.

(c) You are developing a question answering system that users will run on their mobile devices. The system requires parsing each query in real time. Would you choose a **dependency parser** or a **constituency parser** in this situation, and why? Also give one potential *disadvantage* of the system you chose. [*3 marks*]

(d) Compare the following two parses. On the left is the gold standard parse, and on the right is the output of a parser.



What are the parser's **labelled precision and recall** scores on this example? [*2 marks*]

QUESTION CONTINUES ON NEXT PAGE

(e) Answer the following questions about lexical semantics. [*3 marks*]

i. What is the name of the **lexical semantic relationship** between the word *chair* and the word *furniture*?

ii. Suppose you are building a **question-answering system**. Using an example, explain why it is important for your system to be able to identify this kind of relationship (between *chair* and *furniture*).

(f) Name one lexical semantic relationship that is *easy* to identify using **distributional word representations**, and one that is *hard* to identify this way. [*2 marks*]

(g) Suppose we have the following grammar fragment with semantic attachments. Grammar rules are numbered so you can refer to them easily.

	Rule	**Sem. attachment**		**Rule**	**Sem. attachment**
1.	Det →`a`	$\lambda P.\lambda Q.\exists x.P(x) \wedge Q(x)$	6.	Nom →N	N.sem
2.	N →`country`	$\lambda x.country(x)$	7.	Nom →Adj Nom	Adj.sem(Nom.sem)
3.	Adj →`country`	$\lambda P.\lambda x.country(x) \wedge P(x)$	8.	NP →Det Nom	Det.sem(Nom.sem)
4.	V →`walk`	$\lambda x.\lambda y.walk(x, y)$	9.	VP →V	V.sem
5.	N →`walk`	$\lambda x.walk(x)$	10.	VP →V NP	V.sem(NP.sem)

Answer the following questions about this grammar. [*6 marks*]

i. Is the meaning representation language in this grammar **compositional**? Explain why or why not.

ii. Which **rules** are needed to parse the phrase `country walk`?

iii. Show how the **meaning representation** for `country walk` is obtained. You can explain each step, show a parse tree, or use some combination of those to illustrate the method.

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11125 ACCELERATED NATURAL LANGUAGE PROCESSING

Tuesday 11 th December 2018

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **CALCULATORS MAY NOT BE USED IN THIS EXAMINATION.**

MSc Courses

Convener: M.Mistry
External Examiners: W. Knottenbelt, M. Dunlop, M. Niranjan, E. Vasilaki

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. (a) Answer the following questions about this sentence: [*3 marks*]

`<s> the chair chaired the meeting </s>`

i. Not counting the begin and end markers, how many tokens and how many types are in this sentence?

ii. Identify which of the words (if any) have derivational morphology, and which (if any) have inflectional morphology.

(b) The tables below show some of the transition probabilities (top) and output probabilities (bottom) from an HMM. (As usual, row labels indicate the conditioning information.) Use them to answer the following questions. [*4 marks*]

	DT	NN	VBD	VBG	...	</s>
<s>	0.6	0.024	0.004	0.001		0
DT	0.005	0.4	0.0016	0.0002		0.01
NN	0.12	0.1	0.16	0.21		0.2
VBD	0.25	0.031	0.009	0.005		0.07
VBG	0.33	0.03	0.00014	0.00015		0.05
...						

	<s>	</s>	the	chair	chaired	meeting	...
<s>	1	0	0	0	0	0	
DT	0	0	0.3	0	0	0	
NN	0	0	2×10^{-5}	3×10^{-6}	0	5×10^{-7}	
VBG	0	0	0	0	0	6×10^{-5}	
VBD	0	0	0	0	4×10^{-7}	0	
</s>	0	1	0	0	0	0	

i. Assuming the values in the output probability table were estimated from an annotated corpus, can you say whether the word `meeting` is tagged more often as `VBG` or `NN` in the corpus? Justify your answer.

ii. Using the model defined by the tables above, compute $P(\vec{w}, \vec{t})$, where $\vec{w}$ is the sentence from part (a) and $\vec{t}$ is `<s> DT NN VBD DT VBG </s>`. In your answer, you should give the general expression for computing this probability, and expand the expression using the values given above, but you do not need to reduce your answer to a single number. That is, answers of the form (.5)(.8)/(.3) or (.3)(.6) + (.7)(.1) are fine.

(c) The transition and output tables of an HMM allow us to compute $P(\vec{w}, \vec{t})$. Mathematically, how do we use this information to compute $P(\vec{w})$? (That is, write down an equation.) What is the name of the algorithm that allows us to actually do this computation in practice? [*2 marks*]

(d) Give the definition of a "forward probability" in an HMM. During each iteration of the forward-backward algorithm, we combine forward and backward probabilities in order to compute what? [*2 marks*]

2. Statement S1 below entails statement S2:

   (S1) `Alec scratched his chin thoughtfully.`
   (S2) `Alec scratched his chin.`

   Answer the following questions about these statements. [*6 marks*]

   (a) Give logical meaning representations for these two sentences using reified events.

   (b) Explain why using reified events makes it easier to model the entailment relationship, compared to a meaning representation without reified events.

   (c) Give another statement that is entailed by S1, but where the entailment does *not* follow naturally from the logical meaning representations. How does your example illustrate a weakness of logical meaning representations?

3. Consider the sentence `The boy in the shop likes reading`, and answer the following questions about it. [*7 marks*]

   (a) This sentence contains an agreement relationship that a trigram model cannot capture. What is it?

   (b) Using examples that build on the given sentence, explain why *no* N-gram model (for any fixed N) is sufficient to model this type of relationship in English.

   (c) Name two different models of syntax discussed in class that *can* capture this type of relationship. Then pick *one* of these, give a parse of the sentence using this type of model, and use it to explain how the agreement relationship is captured.

4. Below is a grammar that covers a tiny fragment of English. It generates sentences such as "the kids take the toys and books" and "the big kids take the toys".

Freq	Rule	Freq	Rule	Freq	Rule	Freq	Rule
8	Conj→ and	2	V→ books	16	S→ NP VP	31	NP→ Det Nom
12	N→ kids	6	V→ take	3	S→ VP	4	Nom→ Adj Nom
5	N→ toys	11	V→ takes	7	VP→ V	8	Nom→ Nom Conj Nom
10	N→ books	31	Det→ the	9	VP→ V NP	27	Nom→ N
		4	Adj→ big	3	VP→ V NP NP		

(a) Assume that the frequencies (Freq) shown next to the rules are counts from a treebank, and answer the following questions. [*3 marks*]

i. Suppose the rules/counts provided are the complete set of rules/counts collected from the corpus. What are the maximum-likelihood estimated probabilities of the two rules VP→ V NP and N→ books, and what conditional probabilities do these correspond to? That is, for each rule give an expression of the form $P_{MLE}(\cdot|\cdot) = \cdot$ where you replace the dots with the correct answers.

ii. Now suppose the grammar also contains other rules (not shown above) that may not have been observed in the treebank, and you want to apply add-alpha smoothing. Give the formula you would need to use to estimate the smoothed probability of the VP→ V NP rule and say what the terms mean in this context.

(b) The provided grammar overgenerates. Give an example demonstrating this, and explain using your example how subcategorization could help to address the problem. What new problem is introduced by subcategorization? [*3 marks*]

(c) Consider the following sentence, which is ambiguous under this grammar:

```
take the big toys and books
```

Suppose we are parsing this sentence using an algorithm like CKY (you can assume that the algorithm extends CKY to handle ternary rules, but is otherwise the same as CKY). Answer the following questions. [*6 marks*]

i. Draw the alternative parse trees for the sentence and put a box around the node in each parse that is at the root of the ambiguous subtree. Which location in the chart will store this node?

ii. How does the algorithm represent ambiguity in the chart while also being efficient? You should describe the information that is stored in the chart location you identified in part (i), and explain how this information is used when building higher nodes in the tree. Full details of the computation are not needed but your answer must indicate how the efficiency is achieved.

5. Google's Gmail service recently added a new feature. In some cases, when a user receives an email, three possible short responses are suggested. If the user clicks on one of the responses, it is copied into the reply window. (The user can then choose to add more text before sending.) This feature allows the user to avoid typing common responses.

   Some example emails and suggested responses produced by the real system are shown in Figure 1 (next page). Note that suggested responses are not always provided, but if they are, there are always three options.

   Imagine that it is 2016 and you are one of the employees helping to develop this system. Your boss suggests that you build a supervised text classifier to produce the desired results.

   (a) Describe the following aspects of building the system. [*6 marks*]

   (*Note:* this question is not asking for mathematical definitions or technical details of the classifier, nor is it asking you to choose a particular classifier. But you may need to say something about what kind of information the classifier must provide.)

   i. How would you define the classes (labels) that this classifier needs to predict?

   ii. What data is needed to train the supervised classifier and how would you obtain it?

   iii. Once you have built and trained the classifier, how can you use it to produce the desired behaviour?

   (b) You will need to evaluate different versions of your system during development, and eventually convince your boss that your system is worth deploying. What evaluation method(s) and approach(es) would you use to accomplish these goals, and why? Briefly discuss any weaknesses of your approach. [*5 marks*]

   (c) Identify two ethical issues that should be considered when developing this system. That is, identify two aspects of developing or deploying the system that could have unintended negative consequences for individuals, groups, or society as a whole. What are the possible negative consequences and for whom? [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(a) Hi ***,
I just realised that I won't be here on the 19th of Nov. So please make it the 12th of November

Ok, no problem.	Ok, thanks for letting me know.	Will do.

(b) Argh sorry, forgot the attachment - here it is!

Got it, thanks!	Thank you!	Looks great!

(c) Hi ***,
You can hand it into the ITO if there's not a servitor around.
Kind regards,
***

Yes, I have one.	Sorry, I don't have it.	I don't, sorry.

(d) Happy New Year to you too! It was great to see you over the summer and [... long text with personal updates ...]
Let us know if you will be in *** at all--it was great to re-connect in May.
Love to everyone,
***

Figure 1: Several example emails, for Question 5. In examples (a)-(c), the system's suggested responses are shown in boxes and the user can click on one (if desired). In example (d), the system does not provide any suggested responses. Identifying details in all examples are replaced here by ***.

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11125 ACCELERATED NATURAL LANGUAGE PROCESSING

Friday 13 th December 2019

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **CALCULATORS MAY NOT BE USED IN THIS EXAMINATION.**

MSc Courses

Convener: V.Nagarajan
External Examiners: W.Knottenbelt, M.Dunlop, M.Niranjan, E.Vasilaki

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. **Short answer questions (20 marks total).** Many of these questions can be answered with a word or two, and none should require more than a few sentences.

   (a) **N-gram language models.** [*4 marks*]

   For each scenario below, I train a bigram and a trigram model, and compute perplexities. In each case, say which model you would normally expect to have lower perplexity: the **bigram model**, the **trigram model**, or **either model could have lower perplexity**. Justification is *not* needed.

   i. Both models are trained using maximum-likelihood estimation (MLE) on the same corpus. I compute perplexity on the training corpus.

   ii. Each model is trained using MLE on a different corpus. I compute the perplexity of each model on the corpus I used to train it.

   iii. I collect a very large corpus of English Wikipedia data and set aside some random sentences for testing. I use Kneser-Ney smoothing and build models using standard methods with the non-test data, then compute perplexity on the test data.

   (b) Use the following sentence to answer this question. [*5 marks*]

   `The president arrived in Hawaii after a long trip`

   i. Give the **lemma** for any word(s) in the sentence where the lemma is different from the word itself.

   ii. Draw the **dependency parse** above the sentence and label each arc using Universal Dependency style labels.
   Then, underneath each word in the sentence, write its **part of speech**, using Penn Treebank style tags. The relevant tags are all included in the table below as a reminder.

CC	coordinating conjunction	PRP	personal pronoun
CD	cardinal number	RB	adverb
DT	determiner	RBR	comparative adverb
IN	preposition	TO	"to"
JJ	adjective	VB	verb base form
JJR	comparative adjective	VBD	verb past tense
MD	modal	VBG	verb gerund
NN	singular or mass noun	VBN	verb past participle
NNS	noun, plural	VBP	verb non-3g present
NNP	proper noun	VBZ	verb 3sg present
NNPS	proper noun, plural		

QUESTION CONTINUES ON NEXT PAGE

(c) **Evaluating text classification.** [*2 marks*]

Suppose I have a spam classifier that returns the predicted probability that each email is spam. I run it on a small set of emails where I know the gold standard label for each email (1 = spam; 0 = not spam). I then rank the emails according to the system's predicted probability (1 = highest probability; 7 = lowest probability), as shown below.

System ranking	Gold label
1	1
2	1
3	0
4	1
5	0
6	0
7	0

The classifier will label an email as spam if its predicted probability is above some threshold t. Suppose I want the classifier's **precision** on these emails to be at least 80%, and I set t accordingly. What will the classifier's highest possible **recall** be?

(d) What does the **Viterbi algorithm** require as input, and what does it produce as output? If you include equations in your answer, you must explain what they mean. [*2 marks*]

(e) **Edit distance.** [*2 marks*]

I use dynamic programming to compute the weighted edit distance between the strings `BALL` and `BILE`. The weights (costs) for each edit operation are between 0 and 1, and yield the chart shown below.

	B	A	L	L
B	0	←0.3	←0.7	←1.1
I	↑0.2	**↖**0.3	←0.7	←1.1
L	↑0.6	↑0.7	**↖**0.3	↖↑0.7
E	↑0.8	↖↑0.9	**↑**0.5	**←**0.9

i. What is the alignment corresponding to the backtraces shown in bold?

ii. What is the cost of inserting the letter E?

QUESTION CONTINUES ON NEXT PAGE

(f) Use the following three very short documents to answer this question. [*5 marks*]

Doc1: `The president arrived in Hawaii after a long trip. He rested in a local hotel.`

Doc2: `Hawaii is beautiful, and you should take a trip there. I can tell you what my favorite hotel in Hawaii is.`

Doc3: `The president announced another trip yesterday.`

i. Find two (distinct) examples in these documents that illustrate different kinds of linguistic **agreement**. What are the two kinds of agreement your examples illustrate and which words agree in each case?

ii. Consider the two events:
X = "`Hawaii` occurs in a document"
Y = "`president` occurs in a document"

Using MLE on the documents above, find the estimated probabilities $P(X)$ and $P(Y)$ and determine whether X and Y are independent in this data. What does the answer tell us about the value of PMI(X,Y)? (*Hint:* you don't need to compute PMI(X, Y) to answer this question.)

2. **Text classification. (10 marks total)**

You're working for a startup that has partnered with the National Health Service (NHS). NHS want to explore whether they can partially automate diagnosis of depression by asking patients to write a short essay, and applying NLP to classify the text. If the classifier flags the patient as being likely to have depression, then they would be referred to a doctor for further diagnosis and treatment.

(a) Your boss asks you to build a simple baseline system first using Naive Bayes. She thinks the style and sentiment of the document are more discriminative for this task than the topic, so she suggests you use the following features: [*4 marks*]

- All stopwords from a standard stopword list.
- The 75 most frequent positive words and 75 most frequent negative words from a sentiment lexicon.

i. In some of the practical work for the course, we used a sparse vector representation to store word co-occurence counts for a large data set. Would there be a benefit to using a sparse vector representation to store the word-document co-occurence counts in this situation? Explain why or why not.

ii. To train a Naive Bayes classifier, you need to estimate $P(f_i|c_j)$ for each feature f_i and class c_j. Say what the classes are in this scenario, and give an expression for the add-alpha smoothed estimate of $P(\texttt{happy}|c_1)$. Make sure you define all of the terms in your equation and, where possible, say what actual values they have in this case.

(b) Together with the NHS, how would you obtain data to train the classifier and what would you need to ensure before using it? [*3 marks*]

(c) Your boss thinks that if the system is successful, perhaps your company could also sell it to Facebook so they can apply it to users' posts and suggest mental health counselling if a user appears to be depressed. [*3 marks*]

Briefly describe *three* potential problems with this plan. At least one must be technical, and at least one must be ethical.

3. **Co-reference and gender bias. (10 marks total)**

Consider the following two sentences for this question.

S1: `The nurse called the plumber to ask her about the blocked drain.`

S2: `The nurse called the plumber to advise about her back pain.`

(a) Is S1 a pro-stereotypical or anti-stereotypical sentence? By changing a single noun phrase, turn it into the opposite kind of sentence. [*2 marks*]

(b) Explain how sentences like these have been used to measure gender bias in co-reference systems. Which sentence would you expect to be more affected by gender bias in a co-reference system, and why? [*4 marks*]

(c) In class, we discussed a way to reduce gender bias in English co-reference systems by adding artificial gender-swapped examples to the training data. Using concrete examples, explain how this is done in English, and what difficulties you might run into if you tried to use the same method on a morphologically complex language. [*4 marks*]

4. **Parsing and logistic regression. (10 marks total)**

It is 2021 and aliens have landed on earth. The linguist Panini is trying to understand their language and discovers it has some similarities to human languages. For example, there are four parts of speech (verb: `V`, noun: `N`, adjective: `A`, and determiner: `D`) and these combine to form syntactic constituents.

Panini hypothesizes that there are three types of constituents (verb phrases: `VP`, noun phrases: `NP` and adjective phrases: `AP`), and that the type of phrase is determined by the following rules:

R1 If there are more than two words in the phrase labelled as `V`, it is a `VP`.

R2 Otherwise, if the leftmost word in the phrase is labelled as `N`, it is an `NP`.

R3 Otherwise, if the rightmost word in the phrase is labelled as `A`, it is an `AP`.

R4 Otherwise, the sequence of parts of speech (or words) is not a valid constituent (`None`).

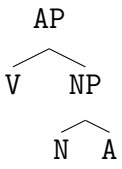
(a) Panini decides to formulate the above rules as a logistic regression model $P(y \mid x)$ where $y \in \{\texttt{VP}, \texttt{NP}, \texttt{AP}, \texttt{None}\}$ and x is a sequence of POS tags over $\{\texttt{V}, \texttt{N}, \texttt{A}, \texttt{D}\}$ (for example $x =$ `D N V P` or $x =$ `V V`). [*3 marks*]

i. Define four features $f_1(x, y)$, $f_2(x, y)$ and $f_3(x, y)$ and $f_4(x, y)$ such that the label y will always be predicted correctly given the weights $w_1 = 8$, $w_2 = 4$, $w_3 = 2$ and $w_4 = 1$.

ii. Write down the model's final decision rule: that is, the formula for choosing y. (*Hint*: your answer should have the form $y = \arg\max_y[\ldots]$)

(b) Panini wants to build a parser using this logistic regression model. The parser takes a sequence of POS tags as input, predicts a phrase label (or `None`) for each span in the sequence, and then tries to join up all phrases labelled as `VP`, `NP`, or `AP` into a larger parse structure. [*2 marks*]

For example, for the string `V N A`, Panini's rules lead to phrases consisting of the three POS spans (0, 1, `V`), (1, 2, `N`), (2, 3, `A`) and the two additional phrases (0, 3, `AP`) and (1, 3, `NP`). The tree built from all these phrases would be:



If this approach is applied to the sequence `V V V D N A`, will it be possible to create a valid hierarchical phrase structure tree using all the phrases generated by Panini's rules? If so, give the tree. If not, explain why not.

QUESTION CONTINUES ON NEXT PAGE

(c) Eventually Panini realizes that his original idea of the grammar was wrong. In fact, the grammar consists of two parts: [*5 marks*]

- a phrase structure grammar with a wide variety of phrase types, which can be expressed in Chomsky Normal Form (CNF), and
- rules R1, R2, and R3 from above.

This new grammar can be parsed by just modifying the CKY algorithm.

i. Write down pseudocode (or a formula) for what happens in the standard CKY algorithm when chart cell (i, j) is processed.

ii. Now show how you would modify the algorithm to constrain the parse so that a phrase structure is only included if it matches *both* the CNF grammar *and* the three additional rules (R1-R3). Will your modification affect the efficiency of the parser? Explain why or why not.