UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Friday 18$^{\underline{\text{th}}}$ May 2018

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer all of Part A and TWO questions from Part B.

Part A is COMPULSORY.

The short answer questions in Part A are each worth 3 marks, 24 marks in total. Each of the three questions in part B is worth 13 marks; answer any TWO of these.

Use one script book for part A and one book for each question in Part B; that is, three books in all.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 3 Courses

Convener: C. Stirling
External Examiners: S.Rogers, A. Donaldson, S. Kalvala

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

Part A

Answer ALL questions in Part A.

Your answers should be thorough—as opposed to being as brief as possible—but they need not be lengthy: from one or two sentences up to a paragraph. When two terms are contrasted, make sure your short definitions of each make clear where the contrast lies. Each question is worth three marks, 24 marks in total for this section.

1. What is a **frequency distribution**? Draw a graph showing the most common frequency distribution of natural language items. What is its name? Give two examples of items that have this distribution.

2. In implementing probablistic models, we often compute with **negative log probabilities** (costs) instead of actual probabilities. What are the benefits to doing so? Why *can't* we do this when implementing the forward algorithm for HMMs?

3. Name and define two types of **syntactic ambiguity**, give a linguistic example of each, and briefly explain why they present a major challenge for **parsing**.

4. Name two types of linguistic ambiguity where knowing the POS tags of the sequence of words can *not* completely eliminate the ambiguity, even in principle. Explain why not.

5. Assuming a **trigram language model** with $\langle s \rangle$ and $\langle /s \rangle$ respectively marking the beginning and end of a sentence, give an expression for the joint probability of the sentence "Kim gave Sandy a book" made up of simpler probabilities.

6. Assuming a **POS tagger** as a Hidden Markov Model (HMM), give the probability that the phrase "Time flies" is tagged NNP VBD in terms of probabilities from the transition model and output model of the HMM.
   Note: assume that the start and end of a sentence is marked $\langle s \rangle$ *and* $\langle /s \rangle$ *respectively.*

7. What problem are **backoff** and **smoothing** trying to solve? Briefly describe how each of them works.

8. Assume that you are building a **word sense disambiguation** model that represents the context of each occurrence of the target word as a vector, consisting of the lemmas of the previous two open class words and the lemmas of the next two open class words, in that order. Then if the target word is *market*, what would the vector be for representing its context in the sentence "The stocks went up as much as the market expected them in the current climate"?

$t_{i-1} \backslash t_i$	NN	VB	JJ	DT	$\langle /s \rangle$
NN	20.6	3.1	45.0	10.0	3.0
VB	15.0	20.0	20.0	2.0	2.0
JJ	9.0	35.0	8.0	45.0	25.0
DT	6.0	45.0	3.0	50.0	55.0
$\langle s \rangle$	8.0	30.0	20.0	2.0	∞

Table 1: Transition Model

$t \backslash w$	*the*	*boy*	*likes*	*boring*	*relatives*
NN	70.0	2.0	4.0	15.0	2.0
VB	∞	15.0	2.0	2.0	25.0
JJ	∞	∞	45.0	8.0	30.0
DT	2.0	∞	∞	∞	∞

Table 2: Output Model

Part B

ANSWER TWO QUESTIONS IN PART B.

1. **Dynamic Programming and Hidden Markov Models**

   (a) What independence assumptions are made by a Hidden Markov Model POS tagger? *[1 mark]*

   (b) Assuming a set T of POS tags, a tag $\langle s \rangle$ for the start of a sentence and a tag $\langle /s \rangle$ for the end of a sentence, derive the Hidden Markov Model (HMM) formula for identifying the most likely tag sequence t_1^n of a word sequence w_1^n. Derive the formula using probabilities, and justify each step in the derivation. Then state the derived formula using *costs*—that is, negative log probabilities. *[4 marks]*

   (c) Assuming that the POS tagger is defined by the transition and output models in Tables 1 and 2, give the negative log probability C that the sentence "the boy likes boring relatives" receives the POS tag sequence DT, NN, VB, JJ, NN.
   Note: All numbers are costs—that is, negative log probabilities. *[5 marks]*

   (d) What happens when tagging "teh boy likes boring relatives" with the transition and output models in Tables 1 and 2? Describe briefly how one might meet the challenge of POS tagging unrestricted and unedited text. *[3 marks]*

2. **Spell Checking**

Suppose that you are developing a spell checker. Your aim is two-fold. First, you need to build a (binary) classifier, which maps each character string in the text that occurs between two spaces to either *correct* (i.e., the character string is spelled correctly) or *incorrect*. Secondly, you need to build a probabilistic model which maps each character string that is classified *incorrect* by the binary classifier to a ranked set of alternative correct spellings.

(a) Minimum edit distance will clearly be an important source of information for estimating the ranked set of alternative correct spellings for an incorrectly spelled word. Assume that it costs 1 to delete or insert a character in a string, 2 to substitute one character for another, and 0 to leave a character unchanged.

Draw a dynamic programming lattice to find the lowest-cost transformation from `form` to `from`. Show your work. Include the backtraces. [*4 marks*]

(b) Describe the kinds of contextual features and language resources that would potentially be informative for building the binary classifier for identifying incorrectly spelled words, and also those that would be informative for building a probabilistic model for ranking candidate corrections. Use illustrative examples, including examples of the character string and examples of the textual context in which it occurs, to justify your answer. [*9 marks*]

3. **Parsing**

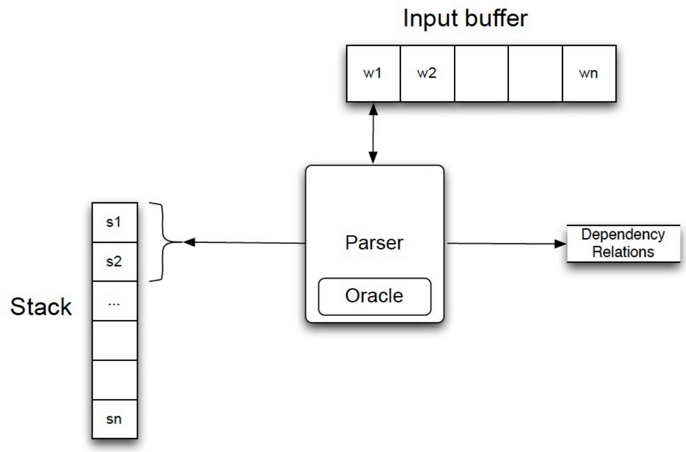
(a) Describe why parsing the following sentences with a probabilistic context free grammar (PCFG) is problematic:

John saw birds with a fish
John saw birds with a telescope

[2 marks]

(b) Draw the correct unlabelled dependency parse for each of the above examples. *[4 marks]*

(c) A shift reduce parser has the following components:



The parser examines the top two elements of the stack and uses the oracle to decide which action to take, based on the current configuration. Define the three actions. *[4 marks]*

(d) Draw the correct dependency parse for *John saw birds*. Use the sequence of actions that a shift-reduce parser would assign to this analysis. *[3 marks]*