UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR09028 FOUNDATIONS OF NATURAL LANGUAGE PROCESSING

Friday 16 th May 2014

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 each question, that is, three books in all.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 3 Courses

Convener: S. Viglas
External Examiners: A. Cohn, T. Field

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. Name and define two types of **syntactic ambiguity**. Which one can a **POS tagger** *not* completely eliminate, even in principle, and why not?

2. Assuming a **bigram language model** with s and e respectively marking the beginning and end of a sentence, give an expression for the joint probability of the sentence "keep calm don't panic" made up of simpler (conditional) probabilities.

3. What is a **frequency distribution**? Draw simple sketches of a **normal** distribution and a **Zipf's law** distribution. For which one are **parametric** statistics appropriate for testing significance?

4. What company built the first credible statistically-based Machine Translation system, based on what corpus? Why did that corpus make the breakthrough possible?

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

6. We looked at two key dynamic programming computations with respect to **Hidden Markov Models**: finding the most probable state sequence for a sequence of observations (**Viterbi decoding**) and finding the total probability of a given observation sequence (the **forward algorithm**). What is the key difference between the two in the computation of each cell in the dynamic programming matrix? For which one is using **costs** instead of **probabilities** to be preferred, and why?

7. What is a **well-formed substring table**? Identify two problems with **recursive descent** parsing that it solves?

8. Consider the following log_2 (maximum likelihood) probabilities of some words and bigrams based on their frequencies in the Brown corpus:

the	−3.8
court	−12.1
superior	−14.4
the court	−13.8
superior court	−16.1

What's interesting about the values for the bigrams, given the values for the words? What do we call the simple measure which captures this?

Part B

Answer two questions in Part B.

1. **HMM POS Tagging**

An HMM POS tagger consists of a *language model* (the *priors*) and a *channel model* (the *likelihoods*) shown in the tables below. These show the *cost*—that is, negative log probabilities—so that (a) you can sum them rather than multiplying, and (b) you are looking to minimise the total cost.

Language model: Gives the conditional cost distribution $-log(P(t_i|t_{i-1}))$, with t_{i-1} down the left and t_i across the top. 's' and 'e' are used respectively for sentence start and sentence end.

$t_{i-1}\backslash t_i$	D	N	V	P	A	e
s	0.7	3.3	4.3	5	2.3	∞
D	∞	1.3	∞	5	1.7	∞
N	4	2.3	0.7	1.3	7	0.3
V	0.3	1.3	∞	2.3	3.3	0.7
P	0.3	3.3	4	5	3.3	5
A	∞	1.3	5	∞	2.3	4

Channel model: Gives the conditional cost distribution $-log(P(w_i|t_i))$.

	D	N	V	P	A
the	0.0	∞	∞	∞	∞
judge	∞	0.2	4	∞	∞
spoke	∞	∞	0.2	∞	5
to	∞	∞	∞	0.0	∞
primary	∞	0.2	∞	∞	1.2
jury	∞	0.0	∞	∞	∞

Given a **Hidden Markov Model**, Viterbi search can be used to determine the most probable state sequence s_1^n for any given sequence of observations o_1^n, that is

$$\underset{s_1^n}{\text{argmax}}\ P(s_1^n|o_1^n)$$

(a) Give the formula for Bayes' Rule, and use it and other simplifying assumptions, step by step, to transform the $P(...)$ expression above into its more useful form consisting of the product of two probabilities. Briefly justify each simplification as you make it. Identify which of the two parts of final formula is the **prior** and which the **likelihood**. [*5 marks*]

(b) Consider the following sentence:

The judge spoke to the primary jury.

i. Give the likelihood, in terms of cost, that this sentence receives the POS tag sequence DNVPDAN, using the costs in tables above. Show your working. [*5 marks*]

ii. Is tagging *primary* as an A or an N more likely, assuming that the tags for the other words in the sentence remain DNVPD_N? Justify your answer. [*3 marks*]

2. **Determining text authorship**

The *Federalist Papers* are a collection of 85 anonymously-authored political articles written in 18th century America. They were in fact mostly authored by Alexander Hamilton and James Madison, but the historical evidence for identifying the author of the individual articles is not always definitive. We can divide the articles into three categories:

- Those asserted to be by Hamilton;
- Those asserted to be by Madison;
- Those whose authorship has not been established.

We'd like to know which assertions are correct and which are incorrect, and for the incorrect and unknown cases, whether the likely author is Hamilton, Madison or a third party.

Drawing on the language modelling technologies discussed in lectures, labs and assignments, design an experiment to answer these questions, assuming you have access to electronic versions of all 85 articles, as well as substantial amounts of representative 18th-century American political writing by a range of authors, including Hamilton and Madison.

(a) Set out your background assumptions and the hypotheses you would be trying to test in order to answer the questions. Be sure your discussion covers both the Federalist Papers themselves and the background corpus. [*4 marks*]

(b) Describe in detail the experiment(s) you would perform, including the modelling technique(s) you would use, how you would train the models and how you would use them to confirm or reject your hypotheses. [*6 marks*]

(c) What factors will determine the reliability of your results? In general, which is likely to be a more reliable conclusion in this sort of experiment:

- These two are similar
- These two are different

Why? [*3 marks*]

3. **Machine Translation Evaluation using BLEU**

The log BLEU score of a candidate translation is given by the formula

$$\log \text{BLEU} = min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n$$

where r is length of the reference, c is the length of the candidate and p_n, the modified precision, is defined as

$$p_n = \frac{\sum_{n-gram \in candidate} Count_{clip}(n\text{-}gram)}{length_n(candidate)}$$

and

$$length_n(candidate) = \text{The length of the candidate measured in } n\text{-grams}$$

i.e.

$$length_n(candidate) = c - (n - 1)$$

The $Count_{clip}$ function, simplified for the case of only one reference translation, is

$$Count_{clip}(n\text{-}gram) = min(Count(n\text{-}gram), \text{frequency of n-gram in the reference})$$

and

$$Count(n\text{-}gram) =$$

$$\begin{cases} \text{frequency of n-gram in the candidate} & \text{if n-gram appears in the reference} \\ 0 & \text{if it doesn't} \end{cases}$$

(a) Compute the log BLEU score for the candidate translation given below, with respect to the single reference translation also given, using $N = 4$ and $w_n = 1/4$. Use Table 1 in Appendix A on the next page to get the logs you need. Show your working. *[5 marks]*

Candidate: it was the good times it was the bad times
Reference: it was the best of times it was the worst of times

(b) Identify the part of the log BLEU score formula known as the brevity penalty. Why do we need it and how does it work? Give an example candidate, for the reference sentence above, which illustrates your answer. *[4 marks]*

(c) Why do we use *modified* precision, instead of ordinary precision (which would be the percentage of n-grams in the candidate that are in reference)? Identify the 'modification' in the log BLEU formula above, and explain how it works. Give an example candidate, for the reference sentence above, which illustrates your answer. *[4 marks]*

Appendix A

	1	2	3	4	5	6	7	8	9	10
1	0.00	0.69	1.10	1.39	1.61	1.79	1.95	2.08	2.20	2.30
2	-0.69	0.00	0.41	0.69	0.92	1.10	1.25	1.39	1.50	1.61
3	-1.10	-0.41	0.00	0.29	0.51	0.69	0.85	0.98	1.10	1.20
4	-1.39	-0.69	-0.29	0.00	0.22	0.41	0.56	0.69	0.81	0.92
5	-1.61	-0.92	-0.51	-0.22	0.00	0.18	0.34	0.47	0.59	0.69
6	-1.79	-1.10	-0.69	-0.41	-0.18	0.00	0.15	0.29	0.41	0.51
7	-1.95	-1.25	-0.85	-0.56	-0.34	-0.15	0.00	0.13	0.25	0.36
8	-2.08	-1.39	-0.98	-0.69	-0.47	-0.29	-0.13	0.00	0.12	0.22
9	-2.20	-1.50	-1.10	-0.81	-0.59	-0.41	-0.25	-0.12	0.00	0.11
10	-2.30	-1.61	-1.20	-0.92	-0.69	-0.51	-0.36	-0.22	-0.11	0.00

Table 1: *log* of simple fractions—numerator across the top, denominator along the side. E.g. $log(3/5) = -0.51$