UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL/MAY EXAMINATIONS 2002

CSC 401H1 S and 2511H S

Duration - 3 hours

No aids allowed

Last Name:

First Name:

Student Number:

Rules:

1. This exam paper contains 12 pages. Legibly write your last name and student number on every page, as the pages of this exam will be separated during marking. *Unnamed pages will not be marked.*
2. There are 36 questions which you must solve, worth a total of 100 marks.
3. Write your answers directly on the exam in the space provided, or if you run out of space, on the back of the same page. For multiple choice questions, simply write the letter of the correct answer next to the question.

Do not write below this line.

MC:

SA:

HM:

EN:

SR:

Total:

Last Name: Student Number:

Multiple Choice questions (2 marks each)

M1. A word that occurs once in a corpus is called a _
(a) singular value.
(b) hapax legomenon.
(c) word token.
(d) zaphod beeblebrox.

M2. The opposite of meronymy is _
(a) hyponymy.
(b) synonymy.
(c) holonymy.
(d) antonymy.

M3. The difference between lemmatisation and stemming is that _
(a) lemmatisation conflates the same forms of different word-types, but stemming does not.
(b) lemmatisation conflates different forms of different word-types, but stemming does not.
(c) a stem is always a word-type, but a lemma need not be.
(d) a lemma is the uninflected form of the word-type, but a stem need not be.

M4. Vowel types are distinguished from each other by _
(a) voicing, place and manner of articulation.
(b) tongue position, height, nasality and lip rounding.
(c) duration, pitch and glottal frequency.
(d) long-term characteristics of human speech.

M5. A transformation-based part-of-speech tagger uses transformational rules _
(a) to convert words into their parts of speech.
(b) to permute the order of words before tagging.
(c) to correct errors in imperfect taggings made in a previous stage.
(d) to convert the parts of speech of one tagset into those of another.

M6. The Hovy-Lin ordering is a partial ordering of _
(a) parts of speech according to their relative frequency in news text.
(b) features according to their importance for text classification.
(c) paragraph/sentence positions according to their importance for summarisation.
(d) similarity metrics according to their importance for word-sense disambiguation.

M7. A distinguishing characteristic of syllable-initial voiceless plosives in English speech relative to other languages is their _
(a) voicing.
(b) point of articulation.
(c) aspiration.
(d) affrication.

M8. *Subject* and *direct object* are examples of _
(a) grammatical functions.
(b) thematic roles.
(c) cases.
(d) parts of speech.

M9. Naive Bayes classification is characterised by the fact that it _
(a) selects the best feature for classification by making strong independence assumptions.
(b) combines the evidence from all available features for classification, but with simplifying assumptions about their interaction.
(c) uses an iterative algorithm that maximises mutual information.
(d) all of the above.

M10. The defining difference between an indicative summary and an informative summary is that _
(a) indicative summaries contain a description with references for further reading.
(b) indicative summaries are query-oriented, whereas informative summaries are document-oriented.
(c) indicative summaries are intended for summarising multiple documents whereas informative summaries summarise single documents.
(d) indicative summaries contain misinformation.

M11. Lesk's algorithm uses a _ to find the most appropriate sense of a word token.
(a) collection of synonyms
(b) bag of words from a dictionary definition
(c) bitext
(d) horoscope

M12. The Jaccard coefficient computes the overlap between two sets, X and Y as _
(a) $|X \cup Y|$.
(b) $|X| + |Y|$.
(c) $\frac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$.
(d) $\frac{|X \cap Y|}{|X \cup Y|}$.

M13. The Hector corpus is most suitable for _
(a) training a text summarisation system.
(b) training a word-sense disambiguation system.
(c) training a part-of-speech tagger.
(d) kindling material.

M14. The use of a part to refer to a whole is known as _
(a) meronymy.
(b) synonymy.
(c) polysemy.
(d) synecdoche.

M15. Passive voice is _
(a) a characteristic of verbs that affects the mapping of grammatical function to part of speech.
(b) a characteristic of verbs that affects the mapping of tense to aspect.
(c) a feature that has been shown to correlate positively with a document's degree of abstractness.
(d) none of the above.

M16. TF*IDF is a calculation that is good for _
(a) determining the salience of a sentence for text summarisation.
(b) defining the dimensions of a vector space representation of documents.
(c) finding a word token that occurs very frequently in one document but not at all in most other documents.
(d) none of the above.

M17. A Viterbi decoder is an algorithm that _

(a) finds the parameters of an HMM that maximise the likelihood of the model with respect to a training corpus.
(b) greedily approximates the optimal path through a trellis using backtraces.
(c) finds the most probable output sequence given a state sequence.
(d) does none of the above.

M18. The C4.5 algorithm grows a decision tree that _

(a) asks yes/no questions to partition data in such a way that maximises entropy.
(b) greedily creates nodes in such a way that maximises information gain.
(c) might overfit the training data if excessively pruned.
(d) none of the above.

M19. The smoothing method that linearly interpolates between a binomial distribution and a uniform n-gram distribution is _

(a) Lidstone's Law.
(b) Laplace's Law of Succession.
(c) Witten-Bell Smoothing.
(d) Good-Turing Smoothing.

M20. The smoothing method that is based on calculating frequencies of frequencies is _

(a) Lidstone's Law.
(b) Laplace's Law of Succession.
(c) Witten-Bell Smoothing.
(d) Good-Turing Smoothing.

Short Answer questions (3 marks each)

A1. State Bayes's Rule.

A2. Suppose you have an unfair coin that comes up heads 7/8 of the time when you flip it. How many bits does it take on average to report the result of one flip of this coin? You may assume $\log_2 7 = 2.8$.

A3. Classify the consonantal phonemes (3 in all) corresponding to the underlined letters in the following English words by their voicing, place and manner of articulation: $\underline{\text{n}}$ose, $\underline{\text{g}}$i$\underline{\text{v}}$e.

A4. Define (formally) precision and recall for a binary classifier.

A5. What is a bitext?

A6. Informally state Zipf's First Law and explain its importance to corpus linguistics.

A7. What parts of speech can adverbs modify?

A8. What is the difference between meronymy and hyponymy?

A9. Define (formally) a Markov model.

A10. Singular value decomposition guarantees the existence of three matrices T, S, and D, for a real-valued $t \times d$ matrix A, such that $A = T \cdot S \cdot D^{\top}$. How many rows and columns does S have?

Last Name: Student Number:

Hidden Markov Models and Language Modelling

H1. (5 marks) Explain why, in a hidden Markov model, we have:

$$P(O|\mu) = \sum_{i=1}^{N} \alpha_i(t)\beta_i(t)$$

no matter what value we choose for time t between 1 and $T+1$ (explain in intuitive terms; a formal proof is not necessary).

H2. Recall that in Witten-Bell smoothing, we calculate the weighting parameter, λ, for $P(w_i \mid w_1 \dots w_{i-1})$ as:

$$\lambda = \frac{C(w_1 \dots w_{i-1})}{CT(w_1 \dots w_{i-1}) + C(w_1 \dots w_{i-1})}$$

Suppose that in our training corpus we see a total of 100 instances of trigrams beginning with *peace and*, representing 50 different such trigrams.

(a) (2 marks) Calculate λ for $P(w_3 \mid \textit{peace and})$.

(b) (3 marks) What fraction of the probability space will we assign to unseen trigrams beginning with *peace and*?

Last Name: Student Number:

Entropy

E1.

(a) (2 marks) Define Kullback-Leibler divergence and informally describe what it measures.
(b) (3 marks) Define mutual information, and informally describe what it measures in terms of Kullback-Leibler divergence.

E2. (5 marks)

(a) (3 marks) State the iterative scaling algorithm for maximum entropy modelling, its inputs and the conditions under which it is guaranteed to converge.

(b) (2 marks) State one advantage and one disadvantage of using maximum entropy modelling relative to other classifier learning algorithms.

Speech Recognition

S1. (5 marks) Spectographic plots of human speech appear with multiple dark (high-energy), horizontal bands. What name is given to the frequencies of these bands, and what is their articulatory significance? State three specific aspects of speech associated with the first four of these bands, and for each aspect, state which band(s).

Last Name: Student Number:

S2. (5 marks) Suppose you are a given a speech recogniser for a language consisting of only the four near-homonymous words, *tire*, *dire*, *dial* and *tile*, with the following channel model for phone realisations:

```
p(ay -> ay) = 1.0

p(t -> t) = 0.8   p(t -> d) = 0.2
p(d -> d) = 0.9   p(d -> t) = 0.1

p(r -> r) = 0.6   p(r -> l) = 0.4
p(l -> l) = 0.6   p(l -> r) = 0.4
```

Using Bayes's Rule and the following unigram language model, calculate the most probable true input for each of the outputs, /tayl/ and /dayl/.

```
P(tire) = .4   P(dire) = .25   P(dial) = .25   P(tile) = .1
```