UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL/MAY EXAMINATIONS 2003

CSC 401H1 S and 2511H S

Duration - 3 hours

No aids allowed

Last Name:

First Name:

Student Number:

Rules:

1. This exam paper contains 11 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:

IR:

HM:

AP:

Total:

Multiple Choice questions (2 marks each)

Legibly write your answer in the left margin of each question.

M1. A phoneme is a _
 (a) basic unit of meaning.
 (b) psychologically distinctive unit of speech.
 (c) basic sound of speech.
 (d) very small telephone.

M2. A word type that occurs once in a corpus is called a _
 (a) singular value.
 (b) hapax legomenon.
 (c) word token.
 (d) typographical error.

M3. C4.5 greedily builds a decision tree by _
 (a) minimising the distance from the root to the leaves.
 (b) minimising entropy.
 (c) minimising information gain.
 (d) maximising information gain.

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 disadvantage of the k-Nearest-Neighbour (kNN) method is _
 (a) its time and space complexity.
 (b) that it may not converge.
 (c) that it takes a long time to train.
 (d) all of the above.

M6. Iterative scaling is an algorithm used in _
 (a) similarity measurement for kNN.
 (b) state sequence decoding.
 (c) maximum entropy modelling.
 (d) time-warping in speech recognition.

M7. The intuition behind Witten-Bell smoothing is to _
 (a) waste no probability mass on unseen events.
 (b) base the probability of an unseen outcome on the probability of observing an outcome for the first time.
 (c) linearly interpolate between binomial and uniform distributions.
 (d) first select from among a uniformly distributed set of distributions.

M8. The intuition behind Lidstone's law is to _
 (a) waste no probability mass on unseen events.
 (b) base the probability of an unseen outcome on the probability of observing an outcome for the first time.
 (c) linearly interpolate between binomial and uniform distributions.
 (d) first select from among a uniformly distributed set of distributions.

M9. The forward algorithm is used to _ in an HMM.
(a) find the most probable state sequence given an output sequence
(b) find the most probable output sequence given a state sequence
(c) maximise the probability of training data by re-estimating parameters
(d) calculate the probability of an output sequence

M10. "Backing off" in a language model is _
(a) an alternative to smoothing.
(b) accomplished by parameter tying.
(c) incompatible with the Good-Turing method.
(d) all of the above.

M11. A common alternative to accuracy as a measure of a document retrieval system's performance is _
(a) interpolated average precision.
(b) cut-off precision.
(c) a precision-recall graph.
(d) any of the above.

M12. The Brill Tagger uses transformations to _
(a) correct an initial tagging using syntagmatic knowledge.
(b) correct an initial tagging using paradigmatic knowledge.
(c) both of the above.
(d) neither of the above.

M13. Transformations in the Brill Tagger are applied _
(a) left-to-right with immediate effect.
(b) left-to-right with delayed effect.
(c) right-to-left with immediate effect.
(d) right-to-left with delayed effect.

M14. The dark horizontal bands visible in spectrally analysed speech are _
(a) formants.
(b) fricatives.
(c) phones.
(d) photons.

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. The naïve Bayes classifier _
(a) selects the n best features for classification, and uses only those.
(b) assumes that the features which characterise the data are conditionally independent.
(c) assumes that short text summaries are more probable than long ones.
(d) assumes that Bayes' rule is false.

M17. The use of a part for the name of the whole is called _
(a) synecdoche.
(b) meronymy.
(c) holonymy.
(d) synonymy.

M18. The [k] sound is a _ consonant.
(a) high
(b) fricative
(c) velar
(d) nasal

M19. Subject-verb agreement is observable in _
(a) person, number and gender.
(b) case, number and gender.
(c) tense, number and mood.
(d) time, place, manner, and degree.

M20. Adjective-noun agreement is observable in _
(a) person, number and gender.
(b) case, number and gender.
(c) tense, number and mood.
(d) time, place, manner, and degree.

Short Answer questions (3 marks each)

A1. State Bayes' Rule.

A2. Name and describe three common indicators used to detect important information in text summarisation systems.

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: $\underline{\text{c}}$at, $\underline{\text{c}}$e$\underline{\text{n}}$t.

A4. Define Kullback-Leibler divergence:
$D(p \,||\, q) =$

A5. Which formants were discovered to correlate with vowel quality, and who discovered it?

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 stemming and lemmatisation?

A9. Define precision, recall, and F-measure as a function of the contingencies $\{\text{true,false}\} \times \{\text{positive,negative}\}$.
Note: true-negative (tn) refers to an outcome that is truly negative and was identified as such, not to one that is "true" in the sense of being positive.

A10. Name two advantages and one disadvantage of using the perceptron algorithm relative to other classification methods.

Information Retrieval

I1. (5 marks) Suppose that, from a query that relates to two documents, your search engine returns 5 answers in this order:

1. related document 1
2. unrelated document
3. related document 2
4. unrelated document
5. unrelated document

What is the uninterpolated average precision of this search engine's performance? You must show your calculation to receive full credit.

Last Name: Student Number:

I2. (3 marks) Is uninterpolated average precision more or less sensitive to recall than interpolated average precision? Why?

I3. (2 marks) Is uninterpolated average precision better or worse than interpolated average precision for comparing only the order of answers from two algorithms on the same dataset? Why?

Hidden Markov Models

H1. (10 marks) Two of the many shortcoming of HMMs with respect to speech recognition are the independence they impose among observations (quantized speech segments), and their very weak notion of duration — each state can only emit one output symbol. One way to address these is to generalize HMMs to *segmented HMMs*, in which multiple output symbols can be emitted upon arrival in a particular state.

A segmented HMM is a tuple, $\langle S, K, A, \Pi, D, B \rangle$, where:

- S is a set of states,
- K is a set of output symbols,
- A is a transition probability matrix,
- Π is a vector of initial state probabilities,
- D is a matrix of *duration probabilities*, and
- B is a function mapping a given state i, and a given duration d to the *joint* probability, $p(o_1 \ldots o_d \mid i, d)$, of an output sequence of the pre-determined length d being emitted upon arrival at state i. In other words, $o_1 \ldots o_d$ are no longer independent of one another, provided they are emitted by the same state.

S, K, A, and Π work exactly as in a standard HMM, and with the same normalization conditions. Generatively, upon landing at state i at time t, the state chooses duration $d \geq 1$ with probability $D(d \mid i)$, and then emits $o_t \ldots o_{t+d-1}$ with probability $B(o_t \ldots o_{t+d-1} \mid i, d)$.

Design a Viterbi-style decoder for a segmented HMM. Recall that this involves defining:

(a) (8 marks) a recursive function, $\delta_i(t)$, which provides, given an output sequence, the probability of the most probable path ending in state i at time t after emitting $o_1 \ldots o_t$, and

(b) (1 mark) a traceback function, $\psi_i(t)$, which provides the previous state in an optimal path of probability $\delta_i(t)$.

(c) (1 mark) Can you re-construct the optimal state sequence from these two components for segmented HMMs? If not, what else do you need?

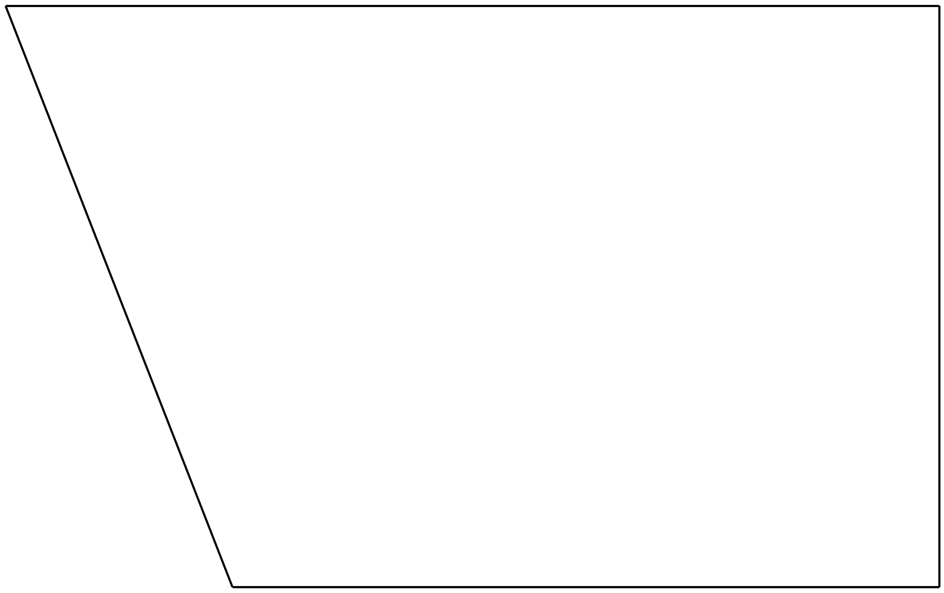
You do **not** need to write pseudocode for computing these - just the equations. You should assume the existence of a function $\rho(i, t)$, which defines the maximum duration of sequence that state i can emit ending at time t. In practice, this is determined by constraints on the utterance length and allowable phone duration times.

Last Name: Student Number:

Articulatory Phonetics

P1. (5 marks) Place the following five vowel sounds in their appropriate position in the vocalic trapezoid below:

1. [i], as in sweet
2. [u], as in food
3. [æ], as in had
4. [a], as in father
5. [ϵ], as in get



P2. (5 marks) List the points of articulation of the English plosive consonants in order from the back of the vocal tract to the front.