UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL/MAY EXAMINATIONS 2004

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, write the letter of the correct answer to the left of the question.

Do not write below this line.

MC:

SA:

GT:

WS:

SR:

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. Hidden Markov models differ from other Markov models in that _
- (a) they emit symbols at states than on arcs.
- (b) not all of their output symbols are known.
- (c) their state transitions are not observable.
- (d) their output emissions are not observable.

M4. The backtrace is used during decoding to _
- (a) ensure that no errors were made.
- (b) maximize the probability of the observed data.
- (c) reconstruct the optimal state sequence.
- (d) re-estimate the parameters of the model.

M5. Uninterpolated average precision is useful for _
- (a) comparing retrieval methods that differ only in the order of the documents returned.
- (b) evaluating retrieval with a high sensitivity to recall.
- (c) placing more importance on a correct first choice than on a correct later choice.
- (d) all of the above.

M6. Good-Turing can be augmented by _ in language models for speech recognition.
- (a) Katz back-off
- (b) linear interpolation
- (c) deleted interpolation
- (d) all of the above.

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 Laplace's Law of Succession 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 backward 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. The end-of-sentence marking heuristic _
(a) provisionally accepts all end-of-sentence punctuation and then makes exceptions.
(b) is prone to failure on known non-final abbreviations.
(c) has no known statistical alternative.
(d) ignores case.

M11. In part-of-speech tagging, Markov models are used to _
(a) implement the Brill tagger.
(b) combine syntagmatic and paradigmatic statistics on tag assignment.
(c) iteratively improve upon an initial tag assignment.
(d) outperform the Brill tagger.

M12. 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.

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

M14. Naive Bayes classification assumes _
(a) all features of the context are conditionally independent
(b) the optimal class can be chosen by maximizing the joint probability of the class and features.
(c) that Bayes" rule is false.
(d) that using single, reliable feature of context is better than combining evidence from all features.

M15. An advantage of maximum entropy modelling is that _
(a) it is resilient to feature independence
(b) it incorporates feature selection
(c) it works only with binary features
(d) none of the above.

M16. The perceptron algorithm is _
(a) guaranteed to converge.
(b) prone to converging to local optima.
(c) a gradient descent algorithm.
(d) none of the above.

M17. A disadvantage of using C4.5 for building classifiers is that _
(a) it is ineffective for problems that cannot be approached by greedily dividing the training set.
(b) it is subject to overfitting problems.
(c) there are better alternatives for linearly separable problems.
(d) all of the above.

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

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

M20. Adjective-noun agreement is observable in _
(a) case, number and gender.
(b) person, 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. SVD decomposes a real-valued matrix A into $A = T \cdot S \cdot D^T$, where S is a diagonal matrix of singular values. What additional knowledge do we have about T and D?

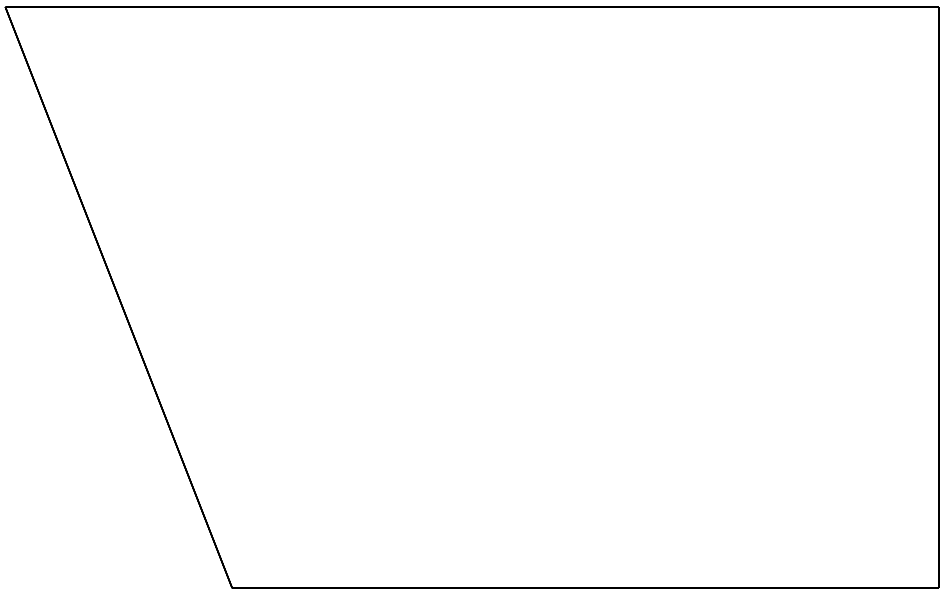
Last Name: Student Number:

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: cat, cent.

A4. Define entropy:
$H(p) =$

A5. Place the following three vowel sounds in their appropriate position in the vocalic trapezoid below:

1. [i], as in sweet
2. [u], as in food
3. [e], as in made



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

A7. Define mutual information in terms of K-L divergence, and informally describe what it measures.

A8. Define selectional preference strength in terms of K-L divergence, and informally describe what it measures.

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 k-Nearest-Neighbours relative to the other classification methods presented in class.

Good-Turing Smoothing

The Good-Turing theorem states that:

$$r^* \approx (r+1) \cdot \frac{E(N_{r+1})}{E(N_r)}$$

where N_r is the random variable for the number of types observed r times in a corpus. In the context of n-gram language modelling using a training corpus of size N, suppose that we estimate $P(w_1 \ldots w_n)$ as $C(w_1 \ldots w_n)^*/N$ for n-grams that occur in the training data ($C(w_1 \ldots w_n) > 0$).

G1. (3 marks) How do we estimate $E(N_r)$ for low values of r? How do we estimate $E(N_r)$ for high values of r? Why can't we use the observed number of types occurring r times in the training data for both?

G2. (7 marks) Prove that if we evenly divide the remaining probability mass over unseen events, then $P(w_1 \dots w_n)$ is also $C(w_1 \dots w_n)^*/N$ for the unseen n-grams ($C(w_1 \dots w_n) = 0$).

Last Name: Student Number:

Word-Sense Disambiguation

W1. (7 marks) Given a token sequence, $v_1 \ldots v_{m-1}\ w\ v_m \ldots v_n$, in which w is a sense-ambiguous word type, define Lesk's algorithm for disambiguating the sense of this occurrence of w.

W2. (3 marks) What linguistic resources does Lesk's algorithm require? Name two additional resources that have been used to improve the performance of Lesk's algorithm.

Speech Recognition

S1. (5 marks) Spectrographic 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' 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) = .2   P(dial) = .3   P(tile) = .1
```