UNIVERSITY OF TORONTO
Faculty of Arts and Science

AUGUST EXAMINATIONS 2017

CSC 401H1 S and 2511H S

Duration - 3 hours

You must achieve at least 50% on this exam to pass the course.

No aids allowed.

Rules:

1. This exam is RESTRICTED. Do NOT write anything on the exam paper. You must write ALL notes and answers in the solution booklets.
2. This exam paper contains 8 pages. Legibly write your last name and student number on the solution booklet cover.
3. There are 38 questions in this exam, worth a total of 100 marks.
4. There are 5 sections to this exam, MC, SA, HMM, SMT, and ASR. Clearly write the section name heading in your solution booklets and keep your answers within the relevant sections.
5. Answer questions in the order they are given in the exam, and clearly label their number.

The section mark breakdowns are:

MC: 40

SA: 30

HMM: 10

SMT: 10

ASR: 10

Total: 100

N.B. This is an old exam. The question bears structural similarity with a concurrent final exam, however, not all exam materials overlap with concurrent syllabus. For e.g. all statistical (S)MT questions may seem unfamiliar since they were not covered in the current session, and thus can be safely ignored. Thus, please use your astute judgment to assess whether a question is outside scope for current year's exam based on the lecture materials covered.

Multiple choice questions (40 marks, 2 marks each)

1. Ambiguity in the ordering and structure of words and phrases in a sentence is ambiguity of ...
   (a) phonology.
   (b) morphology.
   (c) syntax.
   (d) semantics.
2. Useful features for sentiment analysis would probably **not** include ...
   (a) counts of first person pronouns.
   (b) voice prosody.
   (c) cue or stigma words of interest or frustration.
   (d) word fertility scores.
3. The Markov assumption ...
   (a) is based on Zipf's distribution of words in a corpus.
   (b) is based on the informativeness of the recent past.
   (c) is the reason that we have hidden variables in hidden Markov models.
   (d) is the reason that we need to specify the number of states in hidden Markov models.
4. A bigram $[w_t\ w_{t+1}]$ can be considered to be a collocation when ...
   (a) there is not enough probability that the two words are statistically independent.
   (b) there is not enough probability that the two words would not occur together as often by chance.
   (c) there is enough probability that the two words are statistically independent.
   (d) there is enough probability that the two words would not occur together as often by chance.
5. For probability distributions P and Q, Kullback-Leibler divergence, $D_{KL}(P\,||\,Q)$ ...
   (a) is always symmetric (i.e., $D_{KL}(P\,||\,Q) = D_{KL}(Q\,||\,P)$).
   (b) is constrained that $Q(w) > 0$ for every word w such that $P(w) > 0$.
   (c) is another way of expressing conditional entropy.
   (d) is always positive (i.e., $D_{KL}(P\,||\,Q) > 0 \quad \forall P, Q$).
6. Letter-to-sound alignment is *not* relevant to ...
   (a) articulatory synthesis.
   (b) formant synthesis.
   (c) speech recognition.
   (d) dealing with out-of-vocabulary words.
7. How is the cosine measure (i.e., the 'normalized correlation coefficient') used in information retrieval?
   (a) To project vector-space models of a query and a document into a *lower* dimensional space.
   (b) To project vector-space models of a query and a document into a *higher* dimensional space.
   (c) To find the distance between a query and a document in a vector-space model.
   (d) To find the distance between a pair of documents in a vector-space model.
8. The mRMR feature selection method ...
   (a) can only find linear relationships between system variables.
   (b) can use correlation to compute its scores.
   (c) maximizes relevance by minimizing redundancy.
   (d) maximizes relevance by maximizing redundancy.

9. Given a training corpus $\mathcal{O}$, maximum likelihood estimation (MLE) ...
   (a) computes parameters θ_i such that $P(\mathcal{O};\theta_i) \geq P(\mathcal{O};\theta_j)$ for all $\theta_j \neq \theta_i$.
   (b) computes parameters θ_i such that $P(\mathcal{O};\theta_i) \geq P(\mathcal{O}^*;\theta_i)$ for all $\mathcal{O}^* \neq \mathcal{O}$.
   (c) cannot be used in IBM Model 1 because all possible alignments are equally likely.
   (d) cannot be used in HMMs because the initial parameters are not known before training.
10. The C4.5 algorithm ...
   (a) cannot be initialized with other decision trees.
   (b) cannot support continuous-valued attributes.
   (c) cannot support non-binary questions.
   (d) cannot find non-linear decision boundaries between classes.
11. The cepstrum ...
   (a) mimics the response of the human ear.
   (b) is the log of the spectrum of the spectrum.
   (c) is the spectrum of the log of the spectrum.
   (d) ignores the glottal source.
12. The tf.idf score for word w_i and document d_j ...
   (a) decreases with the number of documents in which w_i appears.
   (b) increases with the number of documents in which w_i appears.
   (c) is unaffected by the number of times w_i occurs in d_j.
   (d) increases with the count of w_i within documents $d_k \neq d_j$ in which it already appears.
13. In extractive summarization, the ROUGE-2 score ...
   (a) combines unigram and bigram scores.
   (b) explicitly penalizes brevity.
   (c) gives higher scores to strong lexical chains.
   (d) is normalized over the total number of bigrams across all reference summaries.
14. Which of the following is always true?
   (a) $H(Y \mid X) = H(X \mid Y)$
   (b) $H(X) \leq H(X \mid Y)$
   (c) $H(X,Y) = H(X) + H(Y)$
   (d) $H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
15. Training the GloVe vector representations ...
   (a) uses a window around a current word.
   (b) uses a global word-word co-occurrence matrix.
   (c) feeds a hidden layer back to the input.
   (d) feeds the output layer back to the input.
16. Interpolated average precision ...
   (a) is a geometric mean of precision scores, weighted by a brevity penalty.
   (b) is a linear combination of precision scores.
   (c) is less sensitive to recall than uninterpolated average precision.
   (d) measures precision at identified levels of recall.
17. The hidden variable in a Gaussian mixture model is ...
   (a) the state transition matrix, for all states.
   (b) the identity of the speaker, for all speakers.
   (c) the prior probabiliy of the i^{th} Gaussian, for all i.
   (d) the number of Gaussians, M.

18. Which corpus includes speech?
    (a) Switchboard.
    (b) Penn treebank.
    (c) London-Lund.
    (d) Canadian Hansard.
19. A vowel with the tongue high and to the back ...
    (a) has a high F_1 and a low F_2.
    (b) has a low F_1 and a high F_2.
    (c) has a high F_1 and a high F_2.
    (d) has a low F_1 and low F_2.
20. When updating π_i in the Baum-Welch algorithm, where $\mathcal{O}$ is the training corpus and θ_k are the parameters after the k^{th} iteration, ...
    (a) $\pi_i \leftarrow \pi_i \cdot b_i(w_0)$ where $b_i(w_0)$ is the probability of word w_0 in state i.
    (b) $\pi_i \leftarrow Count(q_0 = i)$ in $\mathcal{O}$.
    (c) $\pi_i \leftarrow P(q_0 = i \,|\, \theta_k)$, where q_0 is the state at time $t = 0$
    (d) $\pi_i \leftarrow P(q_0 = i \,|\, \mathcal{O}; \theta_k)$, where q_0 is the state at time $t = 0$.

Short answer questions (30 marks, 3 marks each)

1. What would happen to Zipf's law if there were no *hapax legomena* in a corpus?
2. In what ways do IBM Model 3 differ from IBM model 1? What parameter(s) do they share in common?
3. Which of the questions $Q1$, $Q2$, or $Q3$ would the ID3 algorithm choose as the root node in the decision tree to classify between values of class C given the 6 training examples ($T1$ to $T6$) below? Explain briefly. (*hint*: you don't need to do the **full** computations if you realize that the entropy of a binary decision is $H(\cdot) = 1$ if both classes have probability 0.5 and $H(\cdot) < 1$ otherwise.)

Datum	$T1$	$T2$	$T3$	$T4$	$T5$	$T6$
$Q1$	yes	yes	yes	yes	no	no
$Q2$	yes	yes	no	no	yes	no
$Q3$	yes	no	yes	no	yes	no
C	yes	yes	no	no	no	yes

4. Provide the formula for the probability of a part-of-speech tag sequence $t_{1:n}$ given a word sequence $w_{1:n}$ in an HMM using the parameters $\{\pi, A, B\}$, which have the same meaning as in class.

5. Compute the maximum likelihood estimate and the add-δ smoothed estimate of $P(zoidberg \mid doctor)$, assuming that the word '*doctor*' appears 15 times in the training corpus, '*zoidberg*' occurs 10 times, the bigram '*doctor zoidberg*' appears 5 times, the size of the training corpus is $50,000$ word tokens, the size of the test corpus is 1000 word tokens, the size of the vocabulary is $20,000$ wordforms, and $\delta = 0.25$. You can leave your answers as fractions.

6. Compute the mutual information $I(X\,;\,Y)$ given the joint probability distribution below. **Hint**: *There are several ways to compute mutual information. The direct formula is a sum over all $x \in X$ and $y \in Y$ and is also the Kullback-Leibler divergence of $P(X,Y)$ relative to $P(X)P(Y)$. You may also consider computing $H(X)$ or $H(Y)$ first.* Recall that $\log_2 8 = 3$, $\log_2 4 = 2$, $\log_2 2 = 1$, and $\log_2 1 = 0$. You can assume that $0 \log_2 0 = 0 \log_2(1/0) = 0$.

$p(x,y)$		X		
		1	2	3
	a	0	1/4	0
Y	b	0	1/4	1/8
	c	1/4	0	1/8

7. What is a 'cognate'? Describe two ways of estimating if two words are cognates of one another.

8. What is the optimum position policy? How is it learned? How is it used?

9. What is the idea behind the Good-Turing smoothing? Alternatively, why should it be more effective than uniformly adding δ to the count of all words?

10. What is the difference between the continuous Fourier transform and the discrete Fourier transform?

Hidden Markov models (10 marks, 3 questions)

HMM 1. (5 marks)

Assume that you have a 3-state hidden Markov model with the following parameters:

$$States\ S = \{s_1, s_2, s_3\}$$
$$\pi = \{\pi_1, \pi_2, \pi_3\} = \{0.3, 0.6, 0.1\}$$

$$A = \begin{array}{cc|ccc} & & & q_t & \\ & & s_1 & s_2 & s_3 \\ \hline & s_1 & 0.2 & 0.5 & 0.3 \\ q_{t-1} & s_2 & 0.0 & 1.0 & 0.0 \\ & s_3 & 0.0 & 0.5 & 0.5 \end{array}$$

$$B = \begin{array}{cc|ccc} & & & w & \\ & & type & token & lemma \\ \hline & s_1 & 0.6 & 0.1 & 0.3 \\ q_t & s_2 & 0.8 & 0.1 & 0.1 \\ & s_3 & 0.2 & 0.0 & 0.8 \end{array}$$

Compute the most likely state sequence for the input $\langle type, type \rangle$ and give its probability by filling in a trellis. *Hint*: You can get partial marks by using variables if you do not remember how to compute their values (e.g., $\delta_i(t)$). *Hint 2:* You don't need to evaluate each term before you compare (e.g., if $x < y$, then $x \cdot z < y \cdot z$).

What is the most likely state sequence for $\langle type, type \rangle$?

What is the probability of the most likely state sequence?

HMM 2. (3 marks)

If, for some reason, you wanted to obtain [redacted]th most likely state sequence for a given observation, how would you efficiently modify the Viterbi algorithm? Provide some high-level pseudo code.

HMM 3. (2 marks)

Briefly describe why $P(\mathcal{O}; \theta)$ [redacted]$_{i=1}^{N}$, given an observation sequence $\mathcal{O}$ of length T, an N-state HMM with parameters θ, and any time $0 \leq t \leq T-1$. Values of α and β are computed by the forward and backward algorithms, respectively.

Statistical machine translation (10 marks, 2 questions)

SMT 1. (5 marks)

Given the two sentence pairs $\{[I\ think], [Je\ pense]\}$ and $\{[I\ am], [Je\ suis]\}$ and the initial translation probabilities below (i.e., '**Before**'), compute a new estimate for the translation probabilities **after** one iteration of the EM algorithm for IBM Model 1 and fill in the table below. Assume 1-to-1 alignments, no NULL word, and no zero-fertility words. You should show your work.

Before		
$P(Je\|I) = 1/2$	$P(suis\|I) = 1/4$	$P(pense\|I) = 1/4$
$P(Je\|am) = 1/2$	$P(suis\|am) = 1/2$	$P(pense\|am) = 0$
$P(Je\|think) = 1/4$	$P(suis\|think) = 0$	$P(pense\|think) = 3/4$

After		
$P(Je\|I) =$	$P(suis\|I) =$	$P(pense\|I) =$
$P(Je\|am) =$	$P(suis\|am) =$	$P(pense\|am) =$
$P(Je\|think) =$	$P(suis\|think) =$	$P(pense\|think) =$

SMT 2. (5 marks)

Given the three reference translations below, compute the BLEU score for each of the three candidate translations, assuming that you consider unigrams, bigrams, and trigrams, and that there is no cap.

Reference 1 have a great summer

Reference 2 have a good summer

Reference 3 a good summer to you

Candidate 1 summer time yo

Candidate 2 have a great vacation

Candidate 3 a good summer

Automatic speech recognition (10 marks, 3 questions)

ASR 1. (3 marks)

Imagine that you have i) for each phoneme, a trained HMM that takes MFCC observations, ii) acoustic speech data in the form of MFCC frames, split by utterance, iii) a 'dictionary' of word pronunciations (mapping words to phoneme sequences), and iv) word-level transcriptions of utterances (i.e., the text of that which was said) but **not** the phoneme annotations. You wish to train **new** HMMs to classify between voiced, unvoiced, and 'other' phonemes (noise and silence can be classified as 'other'). Assume that every utterance begins and ends with silence. Briefly describe how you would set up this system prior to training. Specifically, in one to two sentences each,

(a) Name 2 manners of articulation (not specific phonemes) that are *always* voiced.

(b) Describe how you would set up your new HMMs (i.e., what would the observations be? What would the states be? Would you put any constraints on state transitions?).

(c) Describe how you would determine the boundaries between voiced, unvoiced, and 'other' sounds in your MFCC data for training the new HMMs.

ASR 2. (3 marks)

Now you wish to develop a system that can classify between spoken sentences that are questions and spoken sentences that not questions. Your data set consists of isolated sentences, each annotated as a question or not. If you could only use a single acoustic feature for this task, what feature would that be, and why? What type of classifier would you use, given this single acoustic feature, and why? Briefly describe in 2 or 3 sentences how you would train and evaluate your system(s).

ASR 3. (4 marks)

Copy the following table into your answer booklet. Given the reference (Ref.) sentence *She moves in mysterious ways* and the hypothesis (Hyp.) output by a speech recognizer *Shamu the mysterious whale*, compute the word error rate (relative to the reference, excluding sentence boundary markers) by performing the Levenshtein algorithm **with the very important modification** that substitution errors are considered to be **5** times as costly as insertion or deletion errors. In the case of ties, prioritize insertion errors over deletion errors and deletion errors over substitution errors.

		Ref.						
		⟨s⟩	She	moves	in	mysterious	ways	⟨/s⟩
Hyp.	⟨s⟩							
	Shamu							
	the							
	mysterious							
	whale							
	⟨/s⟩							

Provide the modified word-error-rate.
How many deletion errors are there along the best alignment?