ENGINEERING TRIPOS PART IIB

Friday 24 April 2009 2.30 to 4

Module 4F11

SPEECH AND LANGUAGE PROCESSING

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

There are no attachments.

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS
Engineering Data Book
CUED approved calculator allowed

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

1 A large-vocabulary continuous-speech recognition system is to be constructed. Initially the system is to use a set of context-independent monophone hidden Markov models (HMMs), a unigram language model and a linear lexicon organisation. A recogniser based on the Viterbi algorithm is to be used.

(a) Draw a diagram of the phone-level network structure, including the language model probabilities, used in the Viterbi search. [15%]

(b) Briefly describe how the Viterbi algorithm is used with this network structure. Include how the word-level result is generated. [20%]

(c) Describe how the beam search algorithm can be used to reduce the computational load. [15%]

(d) To further reduce computational load, it is suggested that a tree-based lexicon organisation be adopted.

(i) Draw a segment of the network with a tree-based lexicon organisation. Explain how the language model probabilities are incorporated in this structure. [15%]

(ii) How would the network structure and computational load be altered if word-internal triphone units are used? [15%]

(iii) What is the major problem using a bigram language model with a tree-structured lexicon? Suggest one approach to solving this problem. [20%]

2 The acoustic models in a large-vocabulary continuous-speech recognition system based on hidden Markov models (HMMs) are being designed. Initially the system is to use monophone HMMs. A d-dimensional feature vector is used, and there are 45 phones to be modelled by the system.

(a) The state output distributions are to be either:

(i) monophone HMMs with a single full covariance Gaussian as the output distribution in each state; or

(ii) monophone HMMs with an M-component mixture of diagonal covariance Gaussians as the output distribution in each state.

In each case, give expressions for generating the log-likelihood of an observation vector o_t from a model state j; state the number of parameters used and the computational cost to calculate the log-likelihood; and discuss how well the data associated with each state will be modelled. [40%]

(b) It is proposed that cross-word triphone HMMs are used. The state output distributions are Gaussians with diagonal covariance matrices.

(i) What is meant by cross-word triphones? Compare the modelling of co-articulation in this cross-word triphone system to a system using monophones with mixture Gaussian output distributions. [20%]

(ii) Why is parameter-tying usually used in constructing cross-word triphones? Explain how decision-tree state-tying operates, and what advantages it offers for estimating cross-word triphone systems. [30%]

(iii) What are the disadvantages of using cross-word triphone models? [10%]

(TURN OVER

3 (a) Give *two* reasons why machine translation (MT) is difficult and briefly discuss them. [10%]

(b) An MT system generates a collection of automatic translations $\{E^i\}_{i=1}^R$ for R sentences. These automatic translations are to be compared against a set of reference translations $\{E^i_{(1)}, E^i_{(2)}, E^i_{(3)}, E^i_{(4)}\}_{i=1}^R$. Describe the BLEU score used to measure translation quality. [20%]

(c) A pair of sentences $e_1^I = e_1 \ldots e_I$ and $f_1^J = f_1 \ldots f_J$ are known to be translations of each other. Their word-to-word alignment is described by the alignment process $a_1^J = a_1 \ldots a_J$.

(i) Derive the HMM alignment likelihood $P(f_1^J, a_1^J \mid e_1^I)$. [10%]

(ii) Give a relationship for the efficient calculation of the forward probability $\alpha_j(i) = P(a_j = i, f_1^j \mid e_1^I)$. [20%]

(iii) Define the corresponding backwards probability $\beta_j(i)$ and show how it can be used with the forward probability to compute the probability $P(a_j = i, f_1^J \mid e_1^I)$. [20%]

(iv) Give an expression for the alignment link posterior probability $P(a_j = i \mid f_1^J, e_1^I)$ in terms of the forward and backward probabilities. [20%]

4 N-gram language models are widely used in automatic speech recognition and statistical machine translation.

(a) Discuss the issues involved in setting the value of N when estimating language models on limited amounts of training data. [15%]

(b) A bigram language model with back-off and discounting has the following form:

$$\hat{P}(w_j|w_i) = \begin{cases} d(f(w_i,w_j))\frac{f(w_i,w_j)}{f(w_i)} & \text{if } f(w_i,w_j) > C \\ \alpha(w_i)\hat{P}(w_j) & \text{otherwise} \end{cases} \quad (1)$$

(i) Name the quantities $d(\cdot)$, $\alpha(\cdot)$, and C and describe their role. [20%]

(ii) Describe *one* discounting strategy and give its discounting formula. [20%]

(c) Give an algorithm to construct a weighted finite-state transducer (WFST) for an exact implementation of the bigram language model of Equation (1). [30%]

(d) Briefly discuss the issues involved in extending this algorithm to construct a WFST which implements a back-off trigram language model. [15%]

END OF PAPER