Version PCW/2

EGT3
ENGINEERING TRIPOS PART IIB

Thursday 28 April 2016 9.30 to 11

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.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper.

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

1 A large vocabulary continuous speech recognition system is to be constructed based on cross-word triphone hidden Markov models (HMMs). The system uses decision tree state tying and Gaussian mixture output distributions. The observations are formed using mel frequency cepstral coefficients (MFCCs) and log energy, along with their first and second differentials.

(a) What are the basic modelling assumptions in using HMMs to recognise speech? [10%]

(b) What are the desirable properties of MFCCs as a front end parameterisation for speech recognition systems? Why is the HMM observation vector extended to include first and second differential coefficients? [20%]

(c) Starting from a corpus of suitable speech training material, along with only word level transcriptions and a pronunciation dictionary, outline the steps needed to build the acoustic models for the system. [20%]

(d) Describe the operation of state level parameter tying using decision trees. What are the advantages and disadvantages of this method for constructing triphone based speech recognition systems? [30%]

(e) As part of the decision tree clustering procedure, a node splitting cost is defined. This uses a single Gaussian distribution for the parent node p and children nodes r and s. The change in log likelihood is computed from just the covariance matrices and the data "occupation counts" associated with nodes p, r and s. Show how the required covariance matrix can be computed from knowledge of the state mean and state occupation count for all unclustered triphone contexts. State the assumptions made. [20%]

2 An N-gram language model is to be used in a large vocabulary speech recognition system.

(a) What is meant by an N-gram language model? Explain why N-gram language models are effective for speech recognition systems. [15%]

(b) What is meant by discounting and back-off for N-gram language models? Explain how these methods improve the estimates of N-gram model parameters. [30%]

(c) Briefly describe **two** approaches for speech recognition search in a system that includes a trigram language model. It should be assumed that the acoustic models in the system are context-dependent phone hidden Markov models. [20%]

(d) The number of parameters in a back-off trigram language model is to be limited to save memory. Three alternatives methods have been suggested. For each of the alternatives, describe how effective it would be in controlling language model size and the expected impact on speech recognition performance.

(i) Replace the trigram with a bigram model. [10%]

(ii) Only include trigrams that occur more than once in the language model training corpus. [10%]

(iii) Only include trigrams in the language model that have a significantly higher probability than the corresponding bigram back-off. [15%]

3 (a) Explain how path weights are calculated by a weighted finite state acceptor (WFSA) and how path weights contribute to the weight assigned to a string by a WFSA. [20%]

(b) Give a general expression for the distance from the start state to the final state for the WFSA in Fig. 1. Calculate this distance under the log semiring and under the tropical semiring. [20%]

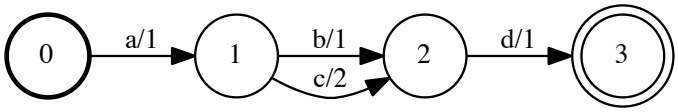


Fig. 1

(c) Explain why the WFSA in Fig. 2 is not deterministic. Draw a determinised version of the WFSA in the tropical semiring. Discuss whether a WFSA produced by determinisation is unique. [30%]

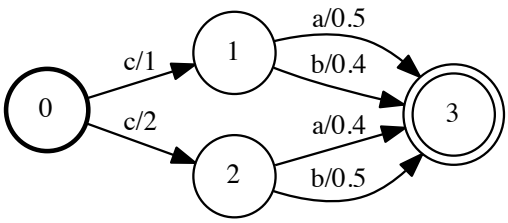
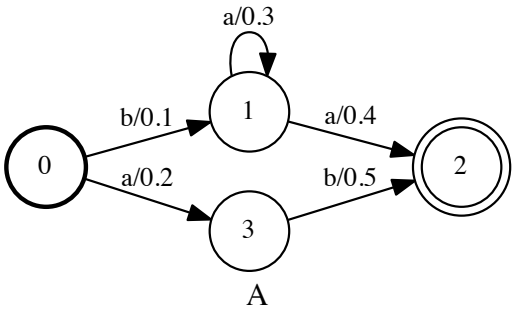


Fig. 2

(d) A WFSA C is the intersection of two WFSAs A and B. Give the equation for $[[C]](x)$, the weight C assigns to a string x, in terms of the weights $[[A]](x)$ and $[[B]](x)$. [10%]

(e) Two WFSAs A and B are shown in Fig. 3. Draw the WFSA that results from their intersection under the tropical semiring. [20%]



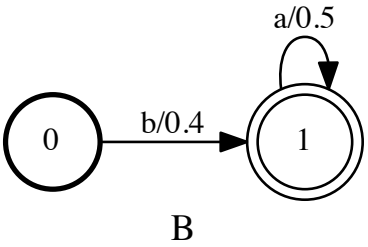


Fig. 3

4 An English sentence of I words is denoted by the sequence $e_0^I = e_0 \dots e_I$, where e_0 is an additional NULL symbol added to the start of the sentence. A foreign sentence of J words is denoted $f_1^J = f_1 \dots f_J$. Alignment between the two sentences is specified by the sequence $a_1^J = a_1 \dots a_J$.

(a) Explain the role of the NULL symbol in alignment. [10%]

(b) By making simplifying conditional independence assumptions, describe the translation probability distribution $P(f_1^J, a_1^J | e_0^I)$ in terms of its three component distributions: the sentence length distribution, the word translation distribution, and the word alignment distribution. [20%]

(c) Give the formulae of the alignment distribution under IBM Model 1 and Model 2, and the HMM alignment model. Explain their differences. [20%]

(d) Derive an efficient algorithm to compute the posterior distribution $P(f_1^J | e_0^I)$ under IBM Model 2. [20%]

(e) A sentence aligned parallel text corpus is to be used to estimate IBM Model 1 and Model 2 parameters.

(i) Describe a flat start training procedure in which the parameters of Model 1 are then used to initialise Model 2. [10%]

(ii) Describe how 'phrases' are defined in the context of phrase-based statistical machine translation and briefly describe how phrase pairs can be extracted from the word-based alignments generated by Model 2. [20%]

END OF PAPER