ENGINEERING TRIPOS PART IIB

Friday 25 April 2008 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) Draw a block diagram showing the generic architecture of a statistical speech recognition system based on sub-word acoustic units. Briefly discuss the function of each of the components. [20%]

(b) The forward-backward algorithm is used as part of the Baum-Welch estimation procedure for the parameters of a hidden Markov model (HMM).

(i) Define suitable forward and backward variables, and give methods for their recursive computation. [25%]

(ii) Show how the posterior probability of HMM state-occupation can be found. [10%]

(c) Describe how the Baum-Welch algorithm can be used to train the parameters of sub-word HMMs, such as monophones, from a corpus of sentences with only word-level transcriptions. State how to deal with the case of multiple entries in the pronunciation dictionary for each word. How should the HMMs be initialised? [25%]

(d) It is proposed to improve the efficiency of the Baum-Welch training procedure by using a pruning mechanism. Two stages of pruning are considered:

(i) Forward probability pruning

(ii) Posterior probability of state occupation.

For each stage of pruning, discuss how it may be implemented, and the likely effect on computation for both short and long training utterances. [20%]

2 A large-vocabulary speaker-independent continuous-speech recognition system based on hidden Markov models (HMMs) uses monophone acoustic units, single Gaussian per state output probability density functions with diagonal covariance matrices, and mel-scale log-energy filterbank features. The word error rate is found to be too high for deployment and a number of modifications have been proposed.

For each of the following changes, (1) give a brief description of the proposed change; (2) state how it would be expected to change the word error rate; and (3) describe any impact on the computational load and memory use of the resulting recognition system.

(a) Replace the log-energy filterbank features by mel-frequency cepstral coefficients. [20%]

(b) Add differential features. [15%]

(c) Use Gaussian mixture distributions. [20%]

(d) Replace the diagonal covariance matrices with full covariance matrices. [20%]

(e) Use cross-word triphones with decision-tree based state tying. [25%]

3 (a) Discuss why N-gram language models are useful for speech and language processing systems. Name and discuss an issue that must be addressed when using N-gram language models in practical systems. [20%]

(b) Give the equation for the probability assigned to a word sequence by a bigram language model. Derive this equation from the general form of a predictive language model, explaining all approximations required. [20%]

(c) Give the equations which specify the "stupid backoff" language model, and discuss its strengths and weaknesses in practical applications. [15%]

(d) A language model vocabulary consists of symbols **a**, **b**, **c**. Suppose that the following three sentences are to be used as a language model training set :

Sentence 1 : <s> **a b c** </s>
Sentence 2 : <s> **a c b a** </s>
Sentence 3 : <s> **a b c** </s>

The statistics needed to compute a bigram language model from this training set are as follows :

Unigram	Count
a	4
b	3
c	3
<s>	3
</s>	3

Bigram	Count	Bigram	Count
<s> a	3	a b	2
b c	2	c </s>	2
a c	1	c b	1
b a	1	a </s>	1

(i) What are the bigram language model probabilities generated via maximum likelihood estimation? [10%]

(ii) Draw a weighted finite state acceptor which realizes the maximum likelihood bigram language model in (d)(i). The acceptor can be drawn with probabilities on the arcs, rather than negative log probabilities. [15%]

(iii) Draw a weighted finite state acceptor which realizes the "stupid backoff" bigram language model associated with the bigram language model

(cont.

in (d)(i). The acceptor can be drawn with probabilities on the arcs, rather than negative log probabilities. Discuss the use of failure arcs as opposed to epsilon arcs and their effect on the language model scores assigned by the acceptor to word sequences. [20%]

4 A sentence-aligned parallel text corpus is to be used to estimate IBM Model 1 and IBM Model 2 parameters.

(a) Define the term *Alignment Error* as used in word alignment for statistical machine translation and give a procedure to calculate the alignment error rate between two sets of word alignments. Discuss how alignment error is used in the development of statistical machine translation systems. [20%]

(b) For a sentence $e_1^I = e_1 \ldots e_I$, its translation $f_1^J = f_1 \ldots f_J$, and their alignment $a_1^J = a_1 \ldots a_J$, give the equations for the Model 1 and Model 2 alignment probabilities. [20%]

(c) Describe a flat-start training procedure in which the parameters of Model 1 and Model 2 are estimated in succession and used to generate word alignments of the parallel text. [20%]

(d) Derive efficient algorithms to compute the following quantities under Model 2:

(i) The probability of f_1^J given e_1^I : $P(f_1^J|e_1^I)$. [20%]

(ii) The posterior probability that $f_{j'}$ is aligned to $e_{i'}$: $P(a_{j'} = i'|f_1^J, e_1^I)$. [20%]

END OF PAPER