Mock exam questions

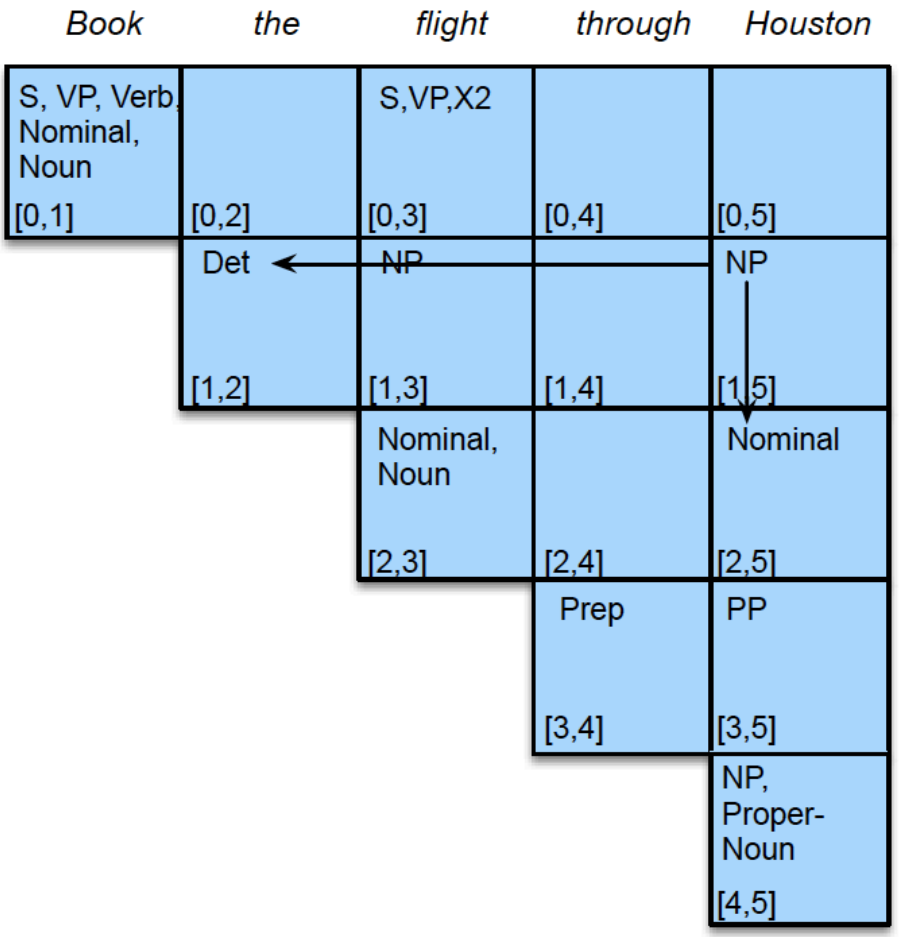
(1)

Basic language modelling: provide a mathematical expression for a basic trigram approximation of the joint probability of a six word sentence $P(w_1, w_2, \dots, w_6)$!

(2)
CKY Parsing: given the a CF grammar and incomplete parsing table, provide all missing symbols for cell [0,5] and indicate the used productions in the same way as for the example NP in cell [1,5] in the table

$\mathscr{L}_1$ in CNF
$S \rightarrow NP\ VP$
$S \rightarrow X1\ VP$
$X1 \rightarrow Aux\ NP$
$S \rightarrow book \mid include \mid prefer$
$S \rightarrow Verb\ NP$
$S \rightarrow X2\ PP$
$S \rightarrow Verb\ PP$
$S \rightarrow VP\ PP$
$NP \rightarrow I \mid she \mid me$
$NP \rightarrow TWA \mid Houston$
$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow book \mid flight \mid meal \mid money$
$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow book \mid include \mid prefer$
$VP \rightarrow Verb\ NP$
$VP \rightarrow X2\ PP$
$X2 \rightarrow Verb\ NP$
$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$

Lexicon
$Det \rightarrow that \mid this \mid the \mid a$
$Noun \rightarrow book \mid flight \mid meal \mid money$
$Verb \rightarrow book \mid include \mid prefer$
$Pronoun \rightarrow I \mid she \mid me$
$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$Aux \rightarrow does$
$Preposition \rightarrow from \mid to \mid on \mid near \mid through$

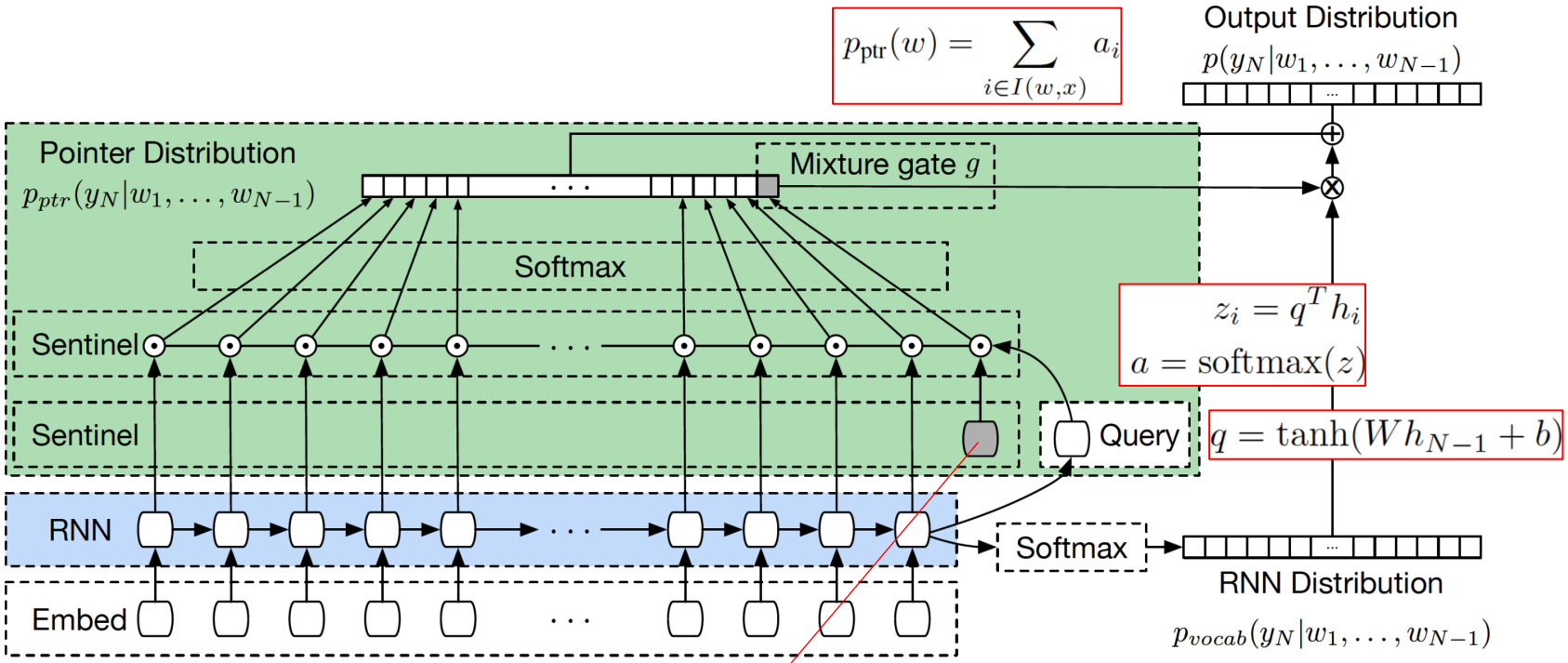


(3)

Deep NLP: Advanced Attention: Pointer Sentinel Networks for Language Modelling (Merity, Xiong, Bradbury, Socher: Pointer sentinel mixture models (2016))

(a) Explain the motivation for using a mixture model of a standard softmax RNN and a pointer network for language modelling!

(b) What is role of the mixture gate and how is its value determined?



answer space

Suggested solutions to the mock exam questions

(1)

$$p(w_1, w_2, \dots, w_6) \approx p(w_1|\langle s\rangle, \langle s\rangle)\, p(w_2|w_1, \langle s\rangle) p(w_3|w_2, w_1)\, p(w_4|w_3, w_2)\, p(w_5|w_4, w_3)\, p(w_6|w_5, w_4)$$

where $\langle s\rangle$ is a begin-of-sentence dummy word.

or

$$p(w_1, w_2, \dots, w_6) \approx p(w_1|\langle s\rangle, \langle s\rangle)\, p(w_2|w_1, \langle s\rangle)\, p(w_3|w_2, w_1) \dots\ p(w_6|w_5, w_4)$$

where $\langle s\rangle$ is a begin-of-sentence dummy word.

or

$$p(w_{1:6}) = p(w_1^6) \approx \prod_{k=1}^{6} p(w_k|w_{k-1}, w_{k-2}) = \prod_{k=1}^{6} p\left(w_k|w_{k-2}^{k-1}\right) = \prod_{k=1}^{6} p\left(w_k|w_{k-2}^{k-1}\right) = \prod_{k=1}^{6} p(w_k|w_{k-1:k-2})$$

where $w_0 = w_{-1} = \langle s\rangle$ is a begin-of-sentence dummy word.

Comments:

- Just writing $p(w_1, w_2, \dots, w_6) \approx p(w_1)\, p(w_2|w_1) p(w_3|w_2, w_1)$, …, $p(w_6|w_5, w_4)$ would arguably result in 8 / 10 or 9 / 10 points (== still a very good answer, but not a perfect answer (in terms of correctness and completeness)).
- For questions that ask for a mathematical expression, additional explanations (so they fit in the provided space at all) are usually not necessary if you use a comprehensible (== comprehensible for the persons grading the exam) notation. The notations used in the slides and recommended background readings can be assumed to be comprehensible.
- The second and third answer possibility is intended to show you that a lot of different notations may be accepted. You do not have to write ALL of these down of course ☺. However, YOU primarily have to unambiguously communicate YOUR solution to US. It is primarily not US that need to try to make SOME SENSE of an ambiguous answer.
- In general: the exam question-and-answer-sheet will only provide a limited space to provide your answers. Please avoid unnecessary elements in your answers that do not contribute anything to answering the questions.
- this question is an “easy” “get a nice base of points” type of question.

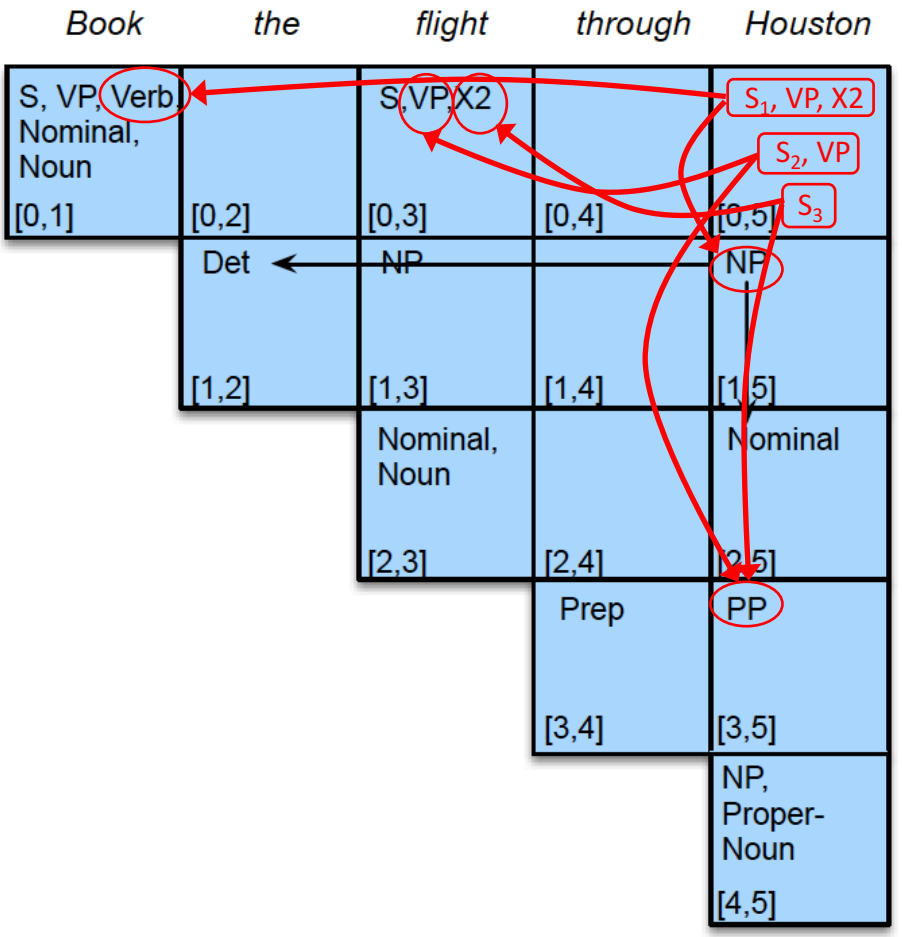
(2)

Comments:

- it is necessary to clearly mark the right hand side symbols to which the respective arrows lead
- instead of inserting them into the cell and drawing arrows, you might alternatively also LIST the respective symbols of cell [0,5] together with the productions for each symbol (of course together with the cell numbers of the right hand symbols):

  S_1 → Verb[0,1] NP[1,5]
  VP → Verb[0,1] NP[1,5]
  X2 → Verb[0,1] NP[1,5]
  S_2 → VP[0,3] PP[3,5]
  etc

- to prepare for such questions, you have to *understand* the principle way the CYK algorithm works. Do not try to just memorize the examples from the slides. In the exam we might have other examples.
- this is an "intermediately difficult" question: not REALLY difficult, but you have to know how CYK operates in principle



(3)
(a) LM with RNN + V-dim. softmax has bottleneck problem (encode all of the input in last hidden state) that negatively affects performance (e.g. worse perplexity) but can predict words in V but not in input (unseen words). Pointer networks (predict member of input sequence with highest attention as next word) has better performance (due to attention) on rare words and out of vocabulary words (that are nevertheless seen in input) but is unable to predict unseen words. → Idea: combine both: use a mixture model: use pointer model whenever possible (large mixing gate g) and back-off to standard softmax RNN otherwise (small g).

(b) g is determined by the pointer (attention) model itself by extending the softmax of raw attention scores z to $q^T s$ (where the sentinel vector s is a learned parameter): if the z vector it self already has a unimodal, quite spiked distribution (==pointer network is "sure" about its prediction), softmax will put a small "probability mass" on $g = a[V+1]$ (the V+1-st component) so g will be small (←→ higher mixture weight on the pointer model's contribution). If pointer network is "unsure" (e.g. z vector has a rather "flat" distribution) → g will be large (←→ higher mixture weight on the standard fixed-V RNN model)

Comments:

- this is a more difficult question that checks your understanding of the respective model. It of course also requires basic understanding of attention. Do not let yourself be frustrated by this question: if you happen to not be able to provide all the details of the suggested answers, you might still get a couple of points here.
- These type of questions illustrate what I meant by "the exam intents to simulate an oral exam in written form and intents to test your UNDERSTANDING of the matter"
- although the suggested answers are largely taken from bullet points on the slides, resist the temptation to just start remembering the slide-contents by heart. THINKING about the introduced systems and approaches and UNDERSTANDING them is the key to your ability to economically organize your knowledge in terms of answering such questions.
- the level of required detail of your answers (in view of the "completeness" aspect of the grading) is largely determined by the available time and space for the answer.
- it is NOT sufficient for a question like (b) to just repeat the given mathematical expressions in words (will not give any points because this is obvious) or to just provide the names / shallow interpretations of the involved quantities (may give some points but this is not sufficient)
- here, it is NOT a contribution to provide some general expressions or explanations of the general attention idea. That was not asked! We will not give points for answer aspects that were not asked!