Practice Exam: Natural Language Processing 1 – 2025 – Paper 1

1 ()

(*a*) Name two limitations of n-gram models as a model of language. [1 point]

(*b*) You are given the following training corpus, where each word is annotated with its part-of-speech (POS) tag. The POS tagset includes: DT (determiner), IN (conjunction, preposition), NN (noun, singular or mass), NNS (noun, plural), VB (verb, base form), VBN (verb, past participle), MD (verb, modal auxiliary), and PUN (punctuation mark). The corpus will be used to train a part-of-speech tagger based on a Hidden Markov Model (HMM).

```
the_DT government_NN will_MD cut_VB taxes_NNS on_IN income_NN ._PUN
the_DT cut_NN should_MD be_VB blocked_VBN ._PUN
```

(*i*) Compute the estimates that would be obtained for lexical probabilities (emission probabilities) and tag sequence probabilities (transition probabilities) under a bigram model. You should not separate the sentences and you should not use start-of-sentence and end-of-sentence symbols. [3 points]

(*ii*) Compute the probability of the following sequence of part-of-speech tags assigned to the sentence, first under the unigram and then under the bigram model, given the probabilities obtained in *(b)(i)*. Provide the formulae you used to compute sequence probabilities under the unigram and bigram models, defining all variables. You may assume p(t_1='DT')=1.0.

```
the_DT cut_NN will_MD be_VB blocked_VBN ._PUN
```

[1 point]

(*c*) You are given the following probabilistic context-free grammar (PCFG):

```
(1.0) S  -> NP VP
(0.42) NP -> PRP N
(0.58) NP -> PRP
(0.2) VP -> V S
(0.5) VP -> V NP
(0.3) VP -> V
```

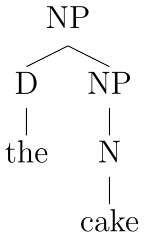
```
(0.25) PRP  -> We
(0.75) PRP  -> her
(1.0) N    -> duck
(0.18) V    -> duck
(0.82) V    -> saw
```

(*i*) Show all possible sentence trees that can be produced for the following example sentence with this grammar.

```
We saw her duck
```

[1 point]

Note that a sentence tree can be represented compactly in bracketed structure, for instance:



⇒ `(NP (D the) (NP (N cake)))`

You can choose what representation to use in your answer.

(*ii*) Find the most probable parse and its probability for the following sentence (the same sentence as in question *(c)(i)*). You can use either dynamic programming or enumeration, but clearly describe all steps of your answer.

```
We saw her duck
```

[1 point]

(*d*) Which of the following techniques is more sensitive to the differences in genre when trained on one corpus (e.g. news data) and evaluated on another (e.g. a corpus of scientific articles)? Briefly explain and motivate your answer.

- NGram language models
- PCFG language models.

[1 point]

(*e*) Consider the following three phrases: *chocolate cake*, *birthday cake*, *blue collar*. (Note on the meaning of *blue collar*: *blue collar* workers do work needing strength or physical skill rather than office work.)

(*i*) Give two reasons why such phrases can be problematic for semantic composition models. Make sure to mention all examples in your answer at least once. [1 point]

(*ii*) Which semantic relation holds between *chocolate cake* and *cake*? Which semantic relation holds between *blue collar* and (the denim shirt of) a working class member? [1 point]

(*f*) Processing sentences with a bidirectional LSTM model allows us to model context in both directions at each time step. Can this model be used to distinguish between different word senses of the same word in two different sentences? Explain why or why not. [1 point]

(g) List two advantages of neural co-reference resolution over a classification method that uses manually-engineered linguistic features. [1 point]

(h) Explain how the attention mechanism is used in neural co-reference resolution method of Lee et al. (2017) discussed in the lecture. What kinds of linguistic information does this mechanism help the model to capture? [1 point]

(i) Interpretation of word and sentence meanings is often influenced by the surrounding discourse, for instance, the immediately preceding sentences. Discuss one way in which discourse information can be incorporated into neural sentence representation learning using an attention mechanism. You can use sentence-level sentiment classification as an example task and assume that the documents containing the sentence of interest are provided. [3 points]

(j) A neural model of abstractive summarisation can be thought of as a conditional language generator. The architecture encodes the available text and predicts a neural parameterisation of a distribution over summaries. The typical model for this task factorises the probability of a summary, given the input text, autoregressively using a bidirectional LSTMs encoder and an LSTM decoder with an attention mechanism in between the two (so that the encoder states can be interpolated differently for each generation step). Here's a sketch: let $x_{1:L}$ be the input text and $y_{<j}$ be the prefix summary before the jth word is generated; the probability distribution of the next token Y_j in the summary is Categorical(softmax($\mathbf{s}_j$)) with $\mathbf{s}_j \in \mathbb{R}^V$ defined as follows:

$$\mathbf{s}_j = \text{linear}_V(\mathbf{h}_j; \theta_{\text{out}}) \tag{1a}$$

$$\mathbf{h}_j = \text{decoderstep}(\mathbf{h}_{j-1}, \mathbf{c}_j, \mathbf{t}_{j-1}; \theta_{\text{dec}}) \tag{1b}$$

$$\mathbf{c}_j = \text{attention}(\mathbf{e}_{1:L}, \mathbf{h}_{j-1}; \theta_{\text{att}}) \tag{1c}$$

$$\mathbf{e}_{1:L} = \text{encoder}(x_{1:L}; \theta_{\text{enc}}) \tag{1d}$$

$$\mathbf{t}_{j-1} = \text{embed}(y_{j-1}; \theta_{\text{in}}) \tag{1e}$$

with V being vocabulary size, and for some adequately initialised decoder state $\mathbf{h}_0$.

Describe a mechanism to incorporate ideas from *extractive* summarisation into this model: propose a modification motivating its connection to extractive summarisation, do explain the architecture/components you introduce (there's no need for formulae, instead, you should be able to explain how these components combine with the baseline model, what are the relevant inputs and outputs and what roles they are expected to play). [2 points]