COMS4705 Final Exam B

Jing Qian

TOTAL POINTS

84.5 / 116

QUESTION 1

Multiple Choice 30 pts

1.1 Encoder-Decoder 3 / 3

✓ **- 0 pts Correct, A**

- 3 pts Incorrect

1.2 Attention 3 / 3

✓ **- 0 pts Correct, B**

- 3 pts Incorrect

1.3 Hyponym 3 / 3

✓ **- 0 pts Correct, C**

- 3 pts Incorrect

1.4 Language Modeling 3 / 3

✓ **- 0 pts Correct, C**

- 3 pts Incorrect

1.5 Summarization 3 / 3

✓ **- 0 pts Correct, A**

- 3 pts Incorrect

1.6 Viterbi 3 / 3

✓ **- 0 pts Correct, A**

- 3 pts Incorrect

1.7 LSTM 3 / 3

✓ **- 0 pts Correct, D**

- 3 pts Incorrect

1.8 WordNet 3 / 3

✓ **- 0 pts Correct, B**

- 3 pts Incorrect

1.9 Modified Bigram Precision 3 / 3

✓ **- 0 pts Correct, D**

- 3 pts Incorrect

1.10 Grammar 3 / 3

✓ **- 0 pts Correct, D**

- 3 pts Incorrect

QUESTION 2

Short Answer 48 pts

2.1 Problems with BLEU 8 / 8

✓ **- 0 pts Correct**

- 3 pts Only describes 1 problem with BLEU

- 6 pts Doesn't describe any problems with BLEU

- 2 pts Incorrect or missing explanation of why researchers still use BLEU

- 1 pts Partially incorrect explanation of why researchers still use BLEU

- 8 pts No response

2.2 Summarization Encoder-Decoder Models 6 / 8

- 0 pts Correct

- 2 pts Encodes input document

✓ **- 1 pts Hierarchical encoder**

✓ **- 1 pts CNN encodes sentence from word embedddings**

- 1 pts LSTM encodes document from sentence embeddings

- 3 pts Decoder implements a classifier to decide if each sentence should be included using attention

- 8 pts No response

2.3 Implicit Bias Test 5.5 / 8

- 0 pts Correct

- 8 pts No response

- **5 pts** No/incorrect explanation of how the test was adapted
- **3.5 pts** Very vague/high-level explanation of how the test was adapted
✓ **- 2.5 pts** **Vague/high-level explanation of how the test was adapted**
- **1 pts** Incomplete explanation of how the test was adapted
- **3 pts** No/incorrect/implausible explanation of how to use the test with profession
- **1.5 pts** Incomplete/vague explanation of how to use the test with profession

2.4 RNN Structure 0 / 8

- **0 pts** Correct
- **2 pts** Incorrect definition of U
- **2 pts** Incorrect/missing explanation of how U is computed
- **4 pts** Incorrect explanation of how h_t is computed
- **2 pts** Description of h_t computation is not sufficient/detailed or partially incorrect
✓ **- 8 pts** **No responses**

2.5 Neural MT 7 / 8

- **0 pts** Correct
- **1 pts** Partially incorrect or insufficient explanation of why attention is important
- **3 pts** Incorrect or missing explanation of why attention important
- **1 pts** Partially incorrect or insufficient explanation of what attention captures
- **3 pts** Incorrect or missing explanation of what attention captures
✓ **- 1 pts** **Partially incorrect or insufficient explanation of sub-word info**
- **2 pts** Incorrect or missing explanation of why sub-word encoding helps
- **8 pts** No response

2.6 Dependency Parsing 0 / 8

- **0 pts** Correct
- **2 pts** action 1 - shift
- **2 pts** action 2 - shift
- **2 pts** action 3 - left arc
- **2 pts** arc set
✓ **- 8 pts** **No response**

QUESTION 3

Long Answer: Word Sense Disambiguation 19 pts

3.1 a: Meanings of "of" 4 / 4

✓ **- 0 pts** **Correct**
- **2 pts** Incorrect definition of "house of red brick"
- **2 pts** Incorrect definition of "whole bowl of cherries"

3.2 b.i: Collocational Features 3.5 / 4

- **0 pts** Correct
✓ **- 0.5 pts** **"of" included as one of its own features**
- **0.5 pts** Window size is not 2
- **0.5 pts** Features for some word other than "of"
- **0.5 pts** Features given for text other than example 1
- **1 pts** Unclear that the features are a vector or list
- **1 pts** Features not vectorized properly
- **1 pts** Features only taken from one side of the word
- **1 pts** Features are not positional (e.g., BOW)
- **2 pts** Lemmas or words not included
- **2 pts** POS tags not included

3.3 b.i: Learned Rules 4 / 4

✓ **- 0 pts** **Correct**
- **0.6 pts** Only 2 rules
- **1.2 pts** Only 1 rule
- **2 pts** No rules
- **1 pts** Minor problems with rules
- **1.5 pts** Major problems with rules
- **2 pts** No justification for how to learn rules
- **2 pts** Incorrect justification for how to learn rules
- **1 pts** Incomplete justification for how to learn rules
- **4 pts** No answer

3.4 BERT vs. word2vec 2.5 / 7

- **0 pts** Correct

✓ **- 1.5 pts BERT and word2vec both incomplete**

- **2 pts** No/incorrect justification for word2vec
- **1 pts** Incomplete justification for word2vec
- **2 pts** No/incorrect justification for BERT
- **1 pts** Incomplete justification for BERT
- **3 pts** Chose word2vec instead of BERT

✓ **- 3 pts No/incorrect/implausible disambiguation approach**

- **1.5 pts** Incomplete disambiguation approach
- **7 pts** No answer

This would be a very difficult classification problem (all senses of all words)

QUESTION 4

Long Answer: Natural Language Generation 19 pts

4.1 a: Architectures 2 / 7

- **0 pts** Correct
- **2 pts** architecture

✓ **- 2.5 pts approach 1**

✓ **- 2.5 pts approach 2**

These are not neural models and you have not specified how they take different kinds of input.

4.2 b: Chatbot Woes 3 / 3

✓ **- 0 pts Correct**

- **1.5 pts** Problem
- **1.5 pts** Approach

4.3 Beam Search 9 / 9

✓ **- 0 pts Correct**

- **1 pts** Choose next set of words
- **2 pts** Choose top 2 for beam round 1
- **1 pts** Choose next set of words
- **2 pts** Choose top 2 for beam round 2
- **1 pts** choose next set of words
- **2 pts** Choose top 2 for beam round 3

Final Exam

COMS W4705: Natural Language Processing

December 9, 2019

Version B

Directions

This exam is closed book and closed notes. You may use a non-graphing calculator. It consists of three parts, each labeled with the amount of time you should expect to spend on it. If you are spending too much time on a question, skip it and come back if you have time.

The first part is multiple choice. The second part is short answer **(select four of six)**. The third part is problem solving. **Read the instructions at the top of each section carefully before beginning to work on the problems.**

Important: Answer Part I by **circling the letters** of your answers. Answer Part II and Part III in the provided spaces. **For Part II, circle the numbers of the four problems you select. If you do not select four, we will grade the first four you attempt.** Turn in all test sheets at the end of the exam. If you need extra paper, write CONTINUED at the end of the space and label each extra sheet with the corresponding question number.

VERY IMPORTANT: If you are still writing by the time we come to collect your exam, we will mark your paper as late and you will receive a deduction.

Name: Jing Qian UNI: jq2282.

UNI of the person to your left: ~~NA~~ kky2114

UNI of the person to your right: NA.

☐ **Check this box if you are a PhD student taking this exam as a comp**

1 Multiple choice (30 points total)

Circle the letter of the choice you select.
For all questions you should select exactly one choice.

Recommended time: 20 minutes.

1. (3 points) Which of the following is an example of a hyponym relation?
   A. "tired" is a hyponym of "weary".
   B. "meal" is a hyponym of "lunch".
   (C.) "algebra" is a hyponym of "math".
   D. "bank" is a hyponym of "bank".

2. (3 points) Which of the following is a drawback of WordNet as an inventory of word senses for all-words word sense disambiguation?
   A. Synonyms do not fully disambiguate between word senses, so WordNet does not contain sufficient information for this task.
   (B.) WordNet senses are very fine-grained, which makes them difficult to annotate as well as to classify.
   C. WordNet is noisy because it is semi-supervised.
   D. WordNet's coverage is very limited, so most words are out-of-vocabulary.

3. (3 points) In a bigram Hidden Markov Model POS tagger using the Viterbi algorithm, how is the final POS sequence chosen?
   (A) Select the most likely POS for each individual word conditioned on the word and the previous POS tag.
   B. Select the most likely POS for each individual word conditioned on the word.
   C. Select the most likely POS for each individual word independently.
   D. Select the most likely POS for each individual word conditioned on the word and the previous two POS tags.

4. (3 points) Given the following training sentences, estimate the probability of the sequence <SOS> Bella is cute <EOS>. Assume you are using a bigram language model, and sentence boundaries occur at the line breaks.

Begin training data:

<SOS> The dog is cute <EOS>
<SOS> Bella is her name <EOS>
<SOS> I love Bella <EOS>

End training data.

$P(B|SOS)\ P(is|B)\ P(cu|is)\ P(<EOS>|c)$
$\frac{1}{3}\quad \frac{1}{2}\quad \frac{1}{2}\quad 1$

A. $\frac{1}{3}$
B. $\frac{1}{36}$
C. $\frac{1}{12}$
D. $\frac{1}{8}$

5. (3 points) Suppose you have the following candidate sentence and reference sentences in a language generation setting. What is the modified **bigram** precision of the candidate sentence as defined in BLEU? (Do not consider start and end tokens.)
Candidate sentence:

my cat chased the lizard

Reference sentences:

the cat chased a mouse
my cat ran after a lizard
the dog chased the frog

$\frac{1+1+1+0}{4}$

A. 1
B. $\frac{9}{5}$
C. $\frac{5}{6}$
D. $\frac{3}{4}$

6. (3 points) In an encoder-decoder sequence to sequence model, like that used in homework 4, what are the two inputs for the *first* decoder step?

(A) The start-of-sequence token and the last hidden state of the encoder.

B. The start-of-sequence token and a randomized initial hidden state.

C. The first word of the gold standard output and the last hidden state of the encoder.

D. The first word of the gold standard output and a randomized initial hidden state.

7. (3 points) Which of the following describes the attention mechanism as used in sequence to sequence models?

A. Attention introduces a dynamic context vector for each *encoding step*; this context vector is a weighted average over all the past encoder steps.

(B) Attention introduces a dynamic context vector for each *decoder step*; this context vector is a weighted average over all encoder hidden states.

C. Attention introduces a dynamic context vector for each *decoder step*; this context vector is a weighted average over all past decoder hidden states, allowing it to look backwards at past decoding steps.

D. Attention introduces a dynamic context vector during each *encoding step*, which allows the encoder to look both forward and backwards while encoding.

8. (3 points) Which of the following is true about summarization?

(A) Extractive summarization reuses sentences from the source to describe it.

B. Abstractive summarization is a type of extractive summarization.

C. Abstractive and extractive summarization are similar, but research on extractive summarization is more recent and at a less mature stage.

D. Abstractive summarization chooses salient sentences from a source to describe it.

9. (3 points) Which sentence could not be handled by the grammar below:

```
S-> NP VP
NP -> det Noun
VP -> verb NP
VP -> verb NP NP
```

A. The cat chased the dog. ✓

B. The chickens escaped the coop. ✓

C. The man gave the dog a bone.

(D.) The dog ate dinner.

10. (3 points) Which problem with using recurrent neural networks to process text is the Long Short-Term Memory Network usually intended to fix?

A. The RNN can become too reliant on certain weights if they are not sometimes zeroed out during training.

B. The RNN architecture tends to overfit to the training data because of a lack of structure. ✗

C. The RNN is unable to learn to make use of subword information because it does not place more importance on character clusters that occur frequently. ✗

(D.) The RNN tends to forget information and dependencies over long distances.

2 Short answer (32 points total)

Provide at most 2 or 3 sentences for four out of the following six questions. Answer in the provided space after the question. If you need extra space, continue on another sheet of paper clearly labeled with the problem number.

NOTE: Choose FOUR. **If you answer more than four, you will be graded on the first four answers you provide.**

Recommended time: 25 minutes.

1. (8 points) What are two problems with BLEU? Why do researchers keep using it despite the problems?

Two problems with BLEU:

① It tends to under-estimate the result when there is paraphrases

② It fails to evaluate the semantic correctness.

Since BLEU is automated, relatively cheap and quick and could be used on various languages, researchers keep using it.

2. (8 points) An extractive summarizer for single-document news summarization that produces a paragraph length summary is naturally implemented by an encoder-decoder architecture. What does the encoder encode? Describe in 1-3 sentences how it could do this. What does the decoder implement?

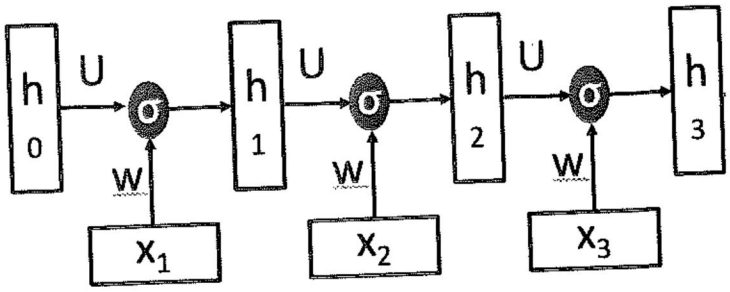
The encoder encodes the sentences (or phrases or words) in the documents. If the sentence appears in the gold summary references, it will have a label 1. Otherwise, it will have a label 0. And the decoder is a neural language model which will predict the sentences appearing in the result summary and put them in fluent order.

3. (8 points) Psychological research has used the implicit association test to measure bias. In this test, the difference in mean response time is measured between two sets of target words and two sets of attribute words. For example, the targets might be "insect" and "flower" and the attribute words might be words that all imply "unpleasant" and words that all imply "pleasant".

How has this test been adapted for word embeddings? How would the modified test be used to determine gender bias as it pertains to occupation?

In the word embeddings, we evaluate the difference in word embeddings between two sets of target words and two sets of attribute words. When it comes to gender bias in occupation, for example, the target words are "software engineers" vs. "nurses" and the attribute words are "man" vs "woman". If we find the difference in the word embeddings of target words are parallel to those of attribute words, we could say there is bias.

4. (8 points) Consider the RNN shown below. What is U and how are its values computed? Describe how the hidden state h_t would be calculated at time t.



5. (8 points) Neural MT is often augmented with attention and sub-word encoding. Why is attention important for neural machine translation? What information does attention capture that is similar to information induced during phrase-based MT? Why does sub-word encoding help neural MT?

Because in Neural MT, we care about information in different hidden states of encoder. We use attention. Attention captures the order of information in the input and outputs similar positions, like during phrase-based MT.

Sub-word encoding could provide the most important information in the input and reduce the noise, which helps it to translate between different languages.

6. (8 points) Show the first three actions an arc-standard transition-based dependency parser would make in parsing the following sentence. Assume you begin with the root already on the stack. Show the arc set that would be created by these actions.

Kathy gave Roscoe a bath .

3 Problem solving (38 points)

There are two problems in this section. Do both problems. Answer in the provided space after the question. If you need extra space, continue on another sheet of paper clearly labeled with the problem number.

Recommended time: 30 minutes.

1. (19 points) **Word sense disambiguation.** The preposition "of" is semantically ambiguous and, in fact, has many different possible meanings. Consider the following noun phrases:

   1. I own *a house of red brick.*
   2. I ate *the whole bowl of cherries.*

   (a) (4 points) Each of the italicized noun phrases above has a different meaning of "of". Describe the meaning that is intended by providing a paraphrase of each italicized phrase.

   1. a house of red brick: a house built with red brick
   2. the whole bowl of cherries: many cherries that filling up the whole bowl.

   (b) (8 points) We discussed several different supervised methods to learn a word sense disambiguation program: one used collocational features, and the other used decision lists with ranked learned rules.

      i. (4 points) Show the collocational features that would be represented for example 1 above assuming a window size of two.

The collocational features refer to the token and POS tags of the words in a context window centering with the target word.

[a, DET, house, NOBJ, of, PREP, red, ADJ, brick, NOUN]

ii. (4 points) Give three examples of learned rules that could be used for disambiguating "of" and describe in one sentence how those rules might be learned.

Rule 1.: If the previous word of "of" is the subject or object, "of" has similar sense with "a house of red brick".
We could learn this rule with collocational features.

Rule 2: If the previous word of "of" is a container or a word used to quantify things, "of" has similar sense with "the whole bowl of cherries".
We could learn this rule with decision list, using AES.

Rule 3: If there is a dependency arc between the verb in the sentence and the word before "of", it's case 1. If the arc is between the verb and the word after "of", it's case 2. We could use parsers to learn it.

(c) (7 points) If you were to use a neural net approach to do word sense disambiguation, would it be better to use Word2Vec or BERT? Justify your response in two sentences by saying why or why not for each. Describe how you could use the embedding space to do word sense disambiguation.

I would use BERT instead of Word2Vec.

1. BERT provides contextual embeddings which relates words with context words in different position. Word2Vec has a context window which is fixed and hence only considers neighbors.
2. BERT is trained with transformer architecture which captures more linguistic information. Word2Vec is trained with standard dense neural network.

With the word embedding from BERT, I will add a classifier to predict what sense the word has in the context. We could use Naive Bayes, Decision Lists or Decision Trees for the classifier.

2. (19 points) **Dialogue systems and generation.** You are building a neural dialogue system that can chat about a presidential election. Given a question, your system must generate a response. Assume that you are given a training set of transcripts of recorded dialogues between voters about the upcoming election.

   (a) (7 points) What kind of neural architecture would you use? Give two different approaches that have been used for dialog systems, where each takes different input. Describe why each approach would be useful.

I will use a seq-to-seq or encoder-decoder architecture.

Two different approaches that used for dialog systems.

Approach 1: Retrieval method. It has a database and when speaker 1 throws out a question, the system would search in its database and output an answ It is useful because it could match the question and answer and outputs desired information. Also, it has relatively accurate and fluent output.

Approach 2: Generative method. The system would take the speak 1's command and does correspondingly. It is useful for scheduling things and taking orders.

   (b) (3 points) Name one problem that a neural chatbot faces and describe an approach that has been used to address it (2 sentences max).

One problem the neural chatbot faces is the inconsistent personality. We may use a memory cell to deposit the personalities generated and consider them in the following dialogues.

(c) (9 points) Your system from part (a) receives as input "My biggest concern is honesty." To generate its response, it has access to the bigram probabilities shown in Table 1. Show how your answer generator would construct the first three words of the response using beam search with a beam size of two. Show each step of constructing the response and the paths the generator maintains.

	are	you	concerned	I	want	a
are	0	0.6	0.2	0	0.05	0.1
you	0.2	0	0.4	0	0.3	0.1
concerned	0	0.4	0	0.1	0	0.5
I	0	0.05	0.01	0	0.7	0.05
want	0	0.5	0	0	0	0.5
a	0	0	0.4	0	0.6	0
<SOS>	0.6	0.1	0	0.2	0	0.1

Table 1: Bigram probabilities for your generation system in Problem 2.c. Here, the probability of observing "you" given "are" is 0.6. <SOS> is the start of sequence token.

Beams:

Step 1: "<SOS> are" (0.6), "<SOS> I" (0.2)

Step 2: "<SOS> are you" (0.6x0.6=0.36), "<SOS> I want" (0.2x0.7=0.14)

Step 3: "<SOS> are you concerned" (0.6x0.6x0.4) "<SOS> are you want" (0.6x0.6x0.3)

So the first three words of the response are "<SOS> are you concerned" and "<SOS> are you want" using beam search with size of two.