First Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2021
Johns Hopkins University
Co-ordinator: Philipp Koehn

7 October 2021

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Language Models and Morphology *20 points*

1. (2 points) I have a **unigram** word model trained on standard English. How would you expect the perplexity of the model on the sentence 'a car David rented' compare to the perplexity of the same model on 'David rented a car'? Circle one:

   (a) The perplexity on 'a car David rented' is greater

   (b) The perplexities are equal

   (c) The perplexity on 'David rented a car' is greater

   b - the perplexities are the same

2. (2 points) I have a **bigram** word model trained on standard English. How would you expect the perplexity of the model on the sentence 'a car David rented' compare to the perplexity of the same model on 'David rented a car'? Circle one:

   (a) The perplexity on 'a car David rented' is greater

   (b) The perplexities are equal

   (c) The perplexity on 'David rented a car' is greater

   a - Perplexity on 'a car David rented' is greater

3. (2 points) You have an n-gram distribution which you smooth using add-ϵ, for some $\epsilon > 0$. The entropy of the smoothed distribution is:

   (a) smaller than

   (b) equal to

   (c) greater than

   the entropy of the original distribution (circle one above).

   c - greater than

4. (2 points) You are told to build a statistical machine translation system between English and Finnish using only 10 million words of Finnish-English bitext and all the monolingual text data found in Finnish and English Google Books data.

   Which translation direction is likely to achieve better performance on a word-based bleu score (circle one):

   (a) English-to-Finnish

   (b) Finnish-to-English

   b - Finnish-to-English

5. (4 points) Give at least 2 reasons **WHY**, based on what you know about both language modeling and the morphology of Finnish and English:

   Reasons include: (1) The English language model will have much more training data than the Finnish language model. (2) Finnish has substantially greater morphological complexity than English, and is more challenging to generate than English. (3) Because of Finnish's much longer words given its agglutinative morphology, an exact-match word-based bleu score will have less likelihood of exact complete-word reference matches than English.

6. (2 points) Bound morphemes are (circle one):

   (a) Words or morphemes that keep the same form every time used and are unchangeable, including conjunctions

   (b) Morphemes that cannot stand alone as a word, and must be attached to a free morpheme

   (c) Words that have morphemes that change depending on the grammar and meaning of a sentence, including nouns

   b

7. (3 points) Give the meaning of the word *cooler*:

   (a) when *-er* is an inflectional morpheme _____

   (b) when *-er* is a derivational morpheme _____

   (a) Something that cools objects (or keeps them cool), (b) more cool (in temperature)

8. (3 points) give at least 3 distinct allomorphs for the English negative prefix *in-*, also simply describe the orthographic context in which the use of each is typically observed, and give one example word for each:

	Allomorph	Orthographic Context	Example Word
1			
2			
3			

   (1) il-, observed before words beginning with l, such as illogical, (2) ir-, observed before words beginning with r, such as irregular, (3) im-, observed before words beginning with m or p, such as immodest or impossible. (4) un- is not a targeted answer, but is minimally acceptable given a reasonable description of its orthographic context, which is essentially unlimited.

Q2. Syntax *20 points*

(5 points) Explain the difference between *constituency* and *dependency* grammars. Do your best to illustrate each kind of parse of the sentence *She plucked a white flower.*

A constituency grammar decomposes a sentence into hierarchically nested phrases, whereas a dependency grammar encodes (typed) relationships between words.

(3 points) What is the head of a phrase? What are some of the things it does? Give two examples.

It is the word that best determines the structure of a phrase, as well as the relationship to external phrases. One bonus point if they mention the argument/adjunct distinction.

(2 points) In the following example, both *see* and *give* are verbs. Which sentence is ungrammatical? Why?

- Kim planned to give Sandy books.
- Kim planned to see Sandy books.

The second is ungrammatical. Ideally the student will mention the argument/adjunct distinction: *give* is ditransitive verb, whereas *see* is (mono)transitive.

(5 points) Below is a grammar (a '|' represents multiple options). If derivations start with the S node, list five sentences that could be produced by this grammar. You don't need to show the parse structures, but it doesn't hurt.

S → NP VP
NP → NN
NP → DT NN
NP → DT ADJ NN
VP → VB NP

DT → a | an | the
ADJ → generous | cotton-headed | green
NN → dog | peony | bone | boy
VB → eats | redefines | picks

(5 points) How could you extend this grammar to support plural nouns and verbs, such that agreement is enforced? (i.e., plural nouns can only appear with plural verbs, likewise for singular). You should need at least five more rules. Give at least one new sentence from your new grammar.

S → NPS VPS
NPS → NNS
NNS → chickens
VPS → VBS
VBS → fight

Q3. Semantics *15 points*

(5 points) Explain the difference between *polysemy* and *homonymy*, and illustrate with an example of each.

Both refer to words (orthographic forms) that have multiple meanings: in polysemy the meanings are related, while in homonymy they are unrelated. For example, the word "bite" can refer to the outcome of applying teeth, or a small amount of food, meanings which are clearly related via a theme (polysemy). On the other hand, "lead" can refer to a chemical element, or to a rope used to guide an animal, meanings which have no relationship (homonymy).

(5 points) Give an example of two sentences that describe the same *event* but with participants in different *syntactic positions*. Why is this useful information, beyond simply understanding each sentence in isolation?

The sentences "The cake baked in the oven" and "The oven baked the cake" both describe the same underlying event with two participants (cake and oven) filling the same roles (THEME and INSTRUMENT or perhaps FORCE), but the first sentence is intransitive, with the cake as subject and the oven in a prepositional phrase, while the second sentence is transitive, with the oven as subject and the cake as object. Knowing that these different surface forms represent the same event allows more accurate estimates or e.g. the frequency of the event.

(5 points) You're thinking about getting a new pet (a turtle), but are worried how it would interact with your current pet (a dog). Will they fight? Cuddle? Ignore each other? How could you use the output of a semantic role labeling system to answer this question?

The output of a semantic role labeling system will be many tuples along the lines of (agent, action, patient): by filtering these where agent equals "turtle" and patient equals "dog", or vice versa, and counting up the unique values of *action*, you can get a sense of how (in this corpus) two such entities are likely to interact.

Q4. Deep Learning *20 points*

(2 points each) Give 4 examples of Supervised Learning tasks in HLT. Give an example of one input and output that could exist in the trainig data.

- Speaker ID: ”Deep Learning is...” ”Dr. Murray”
- Language ID: ”Nuqap suti Kenton kan” ”Quecheua”
- Speech Recognition: Audio Signal ”Hello World”
- Information Extraction: ”Brandon Scott is the mayor of Baltimore” ”Brandon Scott”
- etc.

(2 points) Why do we use a softmax? (Hint what range is the output of a softmax)

It puts our output into probability space.

(2 points) BERT is a Masked Language Model. What are Masked Language Models and how are they different than n-gram Language Models?

Masked language models randomly mask words inside a sequence and try to predict the whole sequence. n-gram LMs try and predict the next word in a sequence predicated on the previous $n-1$ words.

(4 points) Why have Neural Networks taken over in HLT in the last 10 years? Please give at least 2 reasons.

Computing power (GPUs) and access to larger amounts digitized data.

(4 points) How are words input into a Neural Network? What sort of problems could this cause and how do we solve them?

Words are input into Neural Networks using 1-hot vectors. However, the vectors are of a fixed size and you can have Out-of-Vocabulary (OOV) tokens. Instead, subword units such as BPE or SentencePieces are commonly used.

Q5. Information Retrieval *10 points*

You've built a new Information Retrieval system and need to evaluate whether it is good according to the Mean Average Precision (MAP) metric. Suppose you index 1000 documents d_1, d_2, d_3, ..., d_{1000}, and then try searching with two queries q_1 and q_2, each of which retrieving a ranked list of documents, as follows:

1. Query: q_1

   Ranked list of documents retrieved by your system, in order: d_3, d_4, d_2, d_5
   (i.e. Document d_3 is deemed best by system, follow by d_4, etc. Documents not listed here are deemed irrelevant by the system)

   Answer key: d_2, d_4
   (i.e. These are relevant documents for q_1, determined by a manual annotation)

2. Query: q_2

   Ranked list of documents retrieved by your system, in order: d_3, d_1, d_5, d_9

   Answer key: d_3, d_5, d_6, d_7

Please compute the following quantities. For simplicity, leave numbers in fractional form.

- (1 point) Precision for q_1: 1/2
- (1 point) Recall for q_1: 1
- (1 point) Precision for q_2: 1/2
- (1 point) Recall for q_2: 1/2
- (2 points) Average Precision for q_1: 5/12
- (2 points) Average Precision for q_2: 2/3
- (2 points) MAP for the two queries: 13/24

Q6. Information Extraction *25 points*

Consider the following text:

> lululemon athletica inc. (LULU) is reporting for the quarter ending July 31, 2021. The textile company's consensus earnings per share forecast from the 11 analysts that follow the stock is $1.21. This value represents a 63.51% increase compared to the same quarter last year. In the past year LULU has beat the expectations every quarter. The highest one was in the 2nd calendar quarter where they beat the consensus by 27.47%. Zacks Investment Research reports that the 2022 Price to Earnings ratio for LULU is 55.21 vs. an industry ratio of 20.10, implying that **they** will have a higher earnings growth than their competitors in the same industry.
>
> Oracle (ORCL) reported quarterly results late Monday that slightly missed revenue estimates and soundly beat on earnings. Oracle stock fell. The database software company reported adjusted earnings of $1.03 a share on revenue of $9.73 billion. Analysts expected Oracle to report earnings of 97 cents on revenue of $9.75 billion, according to FactSet. The results were for its fiscal first quarter ended Aug. 31. Revenue climbed 4% from the year-ago period. Oracle stock fell 3%, near 86.10, during after-hours trading on the stock market today.
>
> Apple today announced financial results for its fiscal 2021 third quarter ended June 26, 2021. The Company posted a June quarter record revenue of $81.4 billion, up 36 percent year over year, and quarterly earnings per diluted share of $1.30.

You are tasked to extract earning statistics from text like this and store it in a database table that contains

- company name
- ticker symbol
- earnings per share
- percentage change in earnings

Entity linking Mark in the text where companies referred to and indicate which mentions are referring to the same company.

Name one feature that would a co-reference resolution method conclude that the pronoun *they* (bold, at the end of the first paragraph) refers to *lululemon athletica inc.*. Also name one feature that would interfere with this conclusion.

Positive feature: most frequent prior entity, prior association of entity with "earnings"; negative feature: subject parallelism, count (plural vs. singular)

Surface extraction rule Write a surface pattern rule that allows the extraction of the relationship between company name and ticker symbol.

[NE] (CAPITALIZED-LETTERS)

Syntactic extraction rule Write a syntactic pattern rule that allows the extraction of values for the "earnings per share" concept.

NP [NP["earnings"] PP [P NP [head: "share"]] PP [P NP["$" NUMBER]]]

Extra Space