First Midterm Exam

601.467/667 Introduction to Human Language Technology

Fall 2020
Johns Hopkins University
Co-ordinator: Philipp Koehn

8 October 2020

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Language Models *10 points*

If you train a trigram language model on the collected works of Shakespeare, explain in detail how to use the language model to generate large quantities of new Shakespearian sounding text.

Q2. Morphology *10 points*

1. Which type of language is likely to have longer words:
   FUSIONAL LANGUAGE or AGGLUTENATIVE LANGUAGE (circle one)

2. Derivational and inflectional morphology in English.
   (a) Give a highly productive derivational *prefix* in English: _____
   (b) Give a highly productive derivational *infix* in English: _____
   (c) Give a highly productive derivational *suffix* in English: _____
   (d) Give an essentially non-productive *inflectional plural suffix* in English: _____

3. Give an example of a word in English with more than one derivational parse, and illustrate *two* distinct derivational parses for that word via morpheme bracketing using [and] and write the approximate meaning of each. Do not use an example containing the word 'lock' or 'employ'.

Q3. Syntax *20 points*

Recall the definition of a context free grammar, $\mathcal{G}$: a tuple (N, Σ, S, R) where

- N is a set of **nonterminal** symbols;
- Σ is a set of **terminal** symbols, such that $N \cap \Sigma = \emptyset$;
- $S \in N$ is a special **start symbol**; and
- R is a **set of rules** of the form $A \to B\ C$ (for $B, C \in N$) or $A \to a$ (for $a \in \Sigma$)

1. (10 points) Consider the following partial grammar:

   - $N = \{\mathrm{S}, \mathrm{NP}, \mathrm{VP}, \mathrm{PP}, \mathrm{IN}, \mathrm{NN}, \mathrm{VB}, \mathrm{DT}\}$
   - $\Sigma = \{\mathrm{time}, \mathrm{flies}, \mathrm{like}, \mathrm{an}, \mathrm{arrow}\}$
   - $S = \mathrm{S}$

   Design a set of rules R that can produce exactly two sensible parses of the sentence *time flies like an arrow*. The parses should correspond to the following two meanings of this sentence:

   (a) There is a kind of insect known as a *time fly* that appreciates archery

   (b) One's children grow up very fast.

   Your grammar should have exactly 13 rules.

2. (10 points) A derivation d of a grammar $\mathcal{G}$ is a set of rules that recursively expand the root symbol until there are no remaining nonterminals to expand. For example, for the grammar

$$\mathcal{G} = (N, \Sigma, S, R) \tag{1}$$
$$= (\{\mathrm{S}, \mathrm{X}\}, \{\mathrm{y}\}, \mathrm{S}, \{S \to X\ X, X \to X\ X, X \to y\}) \tag{2}$$

one possible derivation is

string	**rule to apply**
S	S → X X
X X	X → X X
X X X	X → y
y X X	X → y
y y X	X → y
y y y	■

A *weighted grammar* is one in which each rule $r \in R$ is associated with a score. The score of a derivation is then the sum of the scores of the rules in that derivation. Assign weights to your rules in Question 1 so that the derivation above interpretation (b) above has a higher score.

Q4. Deep Learning: Sigmoid Function *10 points*

The sigmoid function is defined as follows:

$$\sigma \triangleq \frac{1}{1+e^{-x}}$$

1. What is $\sigma(0)$, $\lim_{x\to\infty}\sigma(x)$, and $\lim_{x\to-\infty}\sigma(x)$?

2. Plot the sigmoid function by hand by considering the above answers.

Q5. Deep Learning: Softmax Function *15 points*

Suppose we have a J dimensional vector $\mathbf{h} \in \mathbb{R}^J$, the softmax function for element j is defined as follows:

$$[\text{softmax}(\mathbf{h})]_j \triangleq \frac{e^{h_j}}{\sum_{i=1}^{J} e^{h_i}}$$

1. Why the softmax function is used as a probabilistic distribution function? Please discuss it by considering $\sum_{j=1}^{J}[\text{softmax}(\mathbf{h})]_j$ and the range of $[\text{softmax}(\mathbf{h})]_j$.

2. Prove that the softmax function becomes the sigmoid function when $J = 2$ by setting $h_2 - h_1 = -x$ (or $h_1 - h_2 = -x$)

Q6. Information Retrieval *10 points*

1. Draw an inverted index data structure for the following eight documents, so that efficient retrieval is possible. Assume we are indexing all words, and there is no stemming or preprocessing of words.

Document 1	Document 2	Document 3	Document 4
the itsy bitsy spider	down came the rain	out came the sun	dried up all the rain
Document 5	**Document 6**	**Document 7**	**Document 8**
the spider went up	incy wincy spider	went climbing up again	washed poor incy out

2. What documents would be retrieved given the query "the AND spider"?

Q7. Information Extraction *15 points*

Consider a machine that has the following text available (training data):

> *Germany midfielder Florian Neuhaus scored on his national team debut but Turkey came back three times to earn a 3-3 draw in their friendly on Wednesday. Kenan Karaman slotted in a stoppage-time equaliser after Luca Waldschmidt fired in an 81st-minute volley that had put Germany 3-2 ahead. Turkey twice before had levelled, with both teams missing several regular players.*
>
> *"We invited Turkey to score goals and again failed to hold on to victory," Germany captain Julian Draxler said. Draxler's good finish from a Kai Havertz assist had put the hosts ahead on the stroke of halftime in front of 300 fans allowed in the stadium in Cologne, Germany.*

and has to answer the following question:

> *Which players scored a goal for Germany?*

1. Draw a knowledge graph that the machine could construct from the data that would enable it to answer the question.

2. Give two examples for co-reference in the text. Cite the referring expressions for each of the examples.

Extra Space