Final Exam

601.467/667 Introduction to Human Language Technology

Fall 2023
Johns Hopkins University
Co-ordinator: Philipp Koehn

14 December 2023

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: ______________________________

Q1. Question Answering *20 points*

1. Researchers have defined various question types for QA systems. Understanding the question type is important for defining the scope of the problem and restricting the answer candidates. Please write down the name of five different question types and give an example question for each. (10 points)

   Answer: Examples from the slides include... Factoid Question: Who was the first American in space? List Question: Name 20 countries that produce coffee. Definition Question: Who is Aaron Copland? Relationship Question: Are Israel's military ties to China increasing? Opinion Question: Why do people like Trader Joe's?

2. The main components of a QA system include Question Analysis, Search, Candidate Extraction, Knowledge Sources, and Answer Ranking. Please explain the purpose of each. (5 points)

   Answer: Question Analysis takes the question as input, finds the answer type and formulates the query for next stage Search. Search picks up potentially relevant documents or info snippets from the Knowledge Source. Candidate Extraction applies text processing on the retrieved documents to return a list of answer candidates, which is finally scored by the Answer Ranking component so that the most likely answer is returned.

3. Here is an example of the Winograd Scheme Challenge. Explain why it is challenging for Machine Reading Comprehension (MRC) systems. (5 points)

   Q: The trophy would not fit in the brown suitcase because it was too big. What was too big?

   A. The trophy

   B. The suitcase

   Answer: In order to solve these kinds of questions, the machine needs some form of common sense knowledge of the physicial world. In this case, it needs to know that A "not fitting" B means that A is too big. These kinds of common sense knowledge are often not explicitly written in the text data used for training MRC systems, so it is one of those challenges that may be easy for humans but hard for machines.

Q2. Digital Humanities *20 points*

1. What are some ways in which humanistic scholars resemble knowledge workers from industry, medicine, finance, etc? (10 points)

   Humanists typically care about very specific topics for which they need to assemble and curate dedicated corpora, have a wide range of technical competencies, and often lack substantial support from collaborators in computer science.

2. What are some ways in which computational approaches offer advantages or drawbacks with respect to humanistic scholarship? (10 points)

   Computation scales and can consider larger scope, and can produce concrete information in the form of e.g. probabilities. It also allows distance from received scholarly precedent and personal bias. At the same time, models remain limited at close reading and reasoning with respect to humans, and that gap is being closed often at the expense of reduced interpretability.

Q3. Interpretable and Explanable NLP *20 points*

1. Briefly describe the main difference between black-box and white-box explanations (10 points)

   Black-box explanations*: for black-box models with no or limited access to their inner workings (e.g., ChatGPT). The focus is on understanding the overall decision-making rather than the specific mechanisms at play.*
   White-box explanations*: the internal workings of the model are transparent and accessible. For example, we can extract the attention weights of the model or compute the gradients of the output to get explanations.*

2. Briefly describe how LIME works (5-6 sentences) (10 points)

   LIME is a technique designed to provide interpretable explanations for individual predictions of complex models. LIME starts by introducing small perturbations to the original instance to create a set of pseudo examples around it. The black-box model is then used to predict outcomes for these pseudo examples. Then a simple and interpretable model (e.g., logistic regression or decision tree) is trained on the pseudo examples and their corresponding black-box model predictions. This local surrogate model provides an interpretable approximation of the black-box model for that specific instance. The coefficients or rules of this interpretable model provide an explanation for the complex model's prediction on the origina instance.

Q4. Ethical Problems *20 points*

1. Describe the Ethical Principle of Beneficence in AI. (10 points)

   Beneficence: *AI should promote well-being, preserve dignity, and sustain the planet "The development of AI should ultimately promote the well-being of all sentient creatures," We should "ensure that AI technologies benefit and empower as many people as possible" "AI technology must be in line with ensuring the basic preconditions for life on our planet, continued prospering for mankind and the preservation of a good environment for future generations."*

2. Describe the notion of Informed Consent in data collection. (10 points)

   The IRBs require informed consent on data collection. IRBs try to ensure that the participants are correctly informed and accept the possible risks even if those are remote. Informed consent implies that the participant on a study must be informed about the study and express their approval on being part of such study; must understand what is being done (you have to assess if the participant understood what they're signing); must voluntarily consent to the experiment; must have the right to withdraw consent at any time.

Q5. Large Language Models *20 points*

1. Answer the following questions. (8 points)

   (a) In 1-2 sentences, define "language modeling".

   Language modeling is about learning probability distribution over sequence of words in language. This is often learned by training predictive distribution to predict the next token, given a present context.

   (b) In no more than five sentence, describe the GPT family of models (the architecture, how they're trained, and how they generates text at inference time.)

   The GPT (Generative Pretrained Transformer) family of models, including GPT-1, GPT-2, and GPT-3, are transformer-based language models that use a decoder-only architecture and are large-scale unsupervised learning systems. They are pretrained on a large corpus of text data and can generate coherent and contextually relevant sentences by predicting subsequent words in a sequence. The models are trained using a variant of the Transformer model, specifically with masked self-attention, where the attention scores are calculated only for positions preceding the current position in the sequence, simulating the autoregressive property of language. At inference time, these models generate text by taking a sequence of words as input, predicting the next word based on the context, and iteratively appending the predicted words to the input sequence until they reach the specified length or end token. The process continues until the end token is produced or the maximum text length is reached.

2. Select all the answer(s) to fill in the blank (. . . . .) in each item. (4 points)

   (a) . . . . . is an argument for the infeasibility of scale due to limited computing.
   □ Advances in computing hardware are much slower than the trends for scaling language models.
   □ Advances in parallel computing can support the fast pace of scaling models.
   □ Scaling language models continues to incur a lot of costs (monetary, carbon footprint, computing resources, etc.)
   □ Scaling models might reduce the overall costs: the availability of a few large models may prevent the cost of building many smaller ones.

   The first and the third statements are arguments for the infeasibility of scaling due to limited computing resources.

   (b) . . . . . is an argument that "data" should not be a bottleneck for scaling language models.
   □ There size of the internet is consistently growing.
   □ There size of Wikipedia is consistently growing.
   □ One can mine data from other modalities (e.g., text data mined from videos).
   □ Even with limited data, we can use it more effectively to get more gains.

All the provided answers support the argument that "data" is not a bottleneck for scaling language models.

3. Answer the following questions in a few sentences (no more than 5 sentences for each). (8 points)

   (a) Explain what long tail of problems in natural language is (provide an example).

   Not all natural language instances have the same difficulty. Some sentences or tasks that frequently appear in our daily discourse, belong to the "head" of a hypothetical distribution of language tasks. In contrast, because of the combinatorial nature of concepts or ideas, there are many *sentences or tasks that are rare in our discourse even though – the "long tail" of the hypothetical distribution of language tasks.*
   Here are a few examples:

   i. Doing basic mathematical operations are a lot more common among small numbers (e.g., "sum of 5 and 2") than large numbers (e.g., "sum of 523,235 and 278,057"). Note the space of small numbers is a lot smaller than the space of large numbers.

   ii. In the context of machine translation, there are few rich-resource languages such as English or Spanish. However, there are plenty of other languages that suffer from limited resources.

   iii. In the context of self-driving cars, driving in large streets on a bright day is an easier challenge for models given their prevalence in say, California. However, driving during a storm is not that frequent and hence a more challenging task.

   (b) Explain how the long tail of problems in natural language poses a challenge to language models.

Language models are empowered by absorbing massive amounts of patterns from their massive pre-training data. Instances in the "head" of the distribution are generally frequent and hence easier to tackle for language models. However, language models struggle with the "tail" of the distribution as they are most tasks often have infinite many rare instances in the "tail" as they are infrequent and extremely large.

Q6. Computational Social Science *20 points*

1. You have a data set of tweets containing terms and hashtags related to the recently released Gemini AI model. Your data has two labels (1) whether or not the post was made by an AI researcher (2) the date the tweet was posted. From glancing at the data, it seems that some people are excited about the model's capabilities, while others are critical about the lack of transparency in the development process, but you don't know much else about what people are saying.

   (a) What are two methods you might use to analyze the data? Write 2-3 sentences about how you would specifically apply the method. (10 points)

   (1) Method:

   Classification (Supervised model) [2pt]

   Description of application:

   Example: annotate data and train a supervised model to classify if a tweet expresses excitement or critique. Use the model to investigate if AI researchers are more excited/critical than non-AI researchers [3pts]

   (2) Method:

   Unsupervised/Clustering/Topic Model 2pts]

   Description of application :

   Example: Use a topic model to discover what the dominant narratives in the online discussions are [3pts]

   Alternative answers (others are possible):

   Method: Time series analysis. Description: Examine how excitement/critique has changed over time since the model release (other answers are possible)

   Method: pretrained representations / ideological mapping. Description: Use pretrained representations to map people or tweets into ideological spaces, such as excited/critical and analyzes differences accross groups

   (b) What are two limitations or ethical considerations of analyzing this data? (4 points)

   (1) Sample of tweets may not be representative (2) What people post on twitter may be different than what they actual think or feel (3) twitter users did not explicitly consent to your analysis (4) working with data posted by individuals risks violating their privacy

2. What is one of the key ways in which a social scientist's approach to a problem often differs from an NLP scientist's approach? Provide an example from each discipline illustrating this difference. (6 points)

Many core social science tasks focus on explanation while NLP often focuses on prediction. [2 points]

Social science examples: When and why do senators deviate from party ideologies? Analyze the impact of gender and race on the U.S. hiring system. Examine to what extent recommendations affect shopping patterns vs. other factors [2 point]

NLP examples: How many senators will vote for a proposed bill? Predict which candidates will be hired based on their resumes. Recommend related products to Amazon shoppers [2 point]