Final Exam

601.467/667 Introduction to Human Language Technology

Fall 2024
Johns Hopkins University
Co-ordinator: Philipp Koehn

17 December 2024

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: _____

Q1. Question Answering *20 points*

1. Methods from several subfields of NLP (namely traditional Question Answering, Machine Reading Comprehension, and Large Language Models) can all address the task of answering questions. For example, I can ask the same question "What countries are the largest producers of lithium?" to IBM Watson, a SQuAD model, and ChatGPT. Please discuss the ways in which these three subfields differ, in terms of motivation and problem setup. (12 points)

   Answer: MRC tasks are designed to test the capabilities of reading and reasoning and serves as a benchmark for AI research. QA focuses more on the end-user, so the goal is to have something that is useful. LLMs are designed to be general-purpose AI agents. As a result, MRC is usually restricted to one document where the answer is present, and an answer span is predicted. QA exploits multiple knowledge sources and employs multiple technologies such as IR, NER, and Parsing to get the correct answer. LLMs generate answers based on whatever is computed from the neural network internals so they tend to be more free-form but can more susceptible to hallucination.

2. Continuing from the previous question: if you were to build a start-up centered around answering questions about research papers to sell to a grad student audience, which of the three approaches (QA, MRC, LLM) would you implement? Pick one and justify. (3 points)

   Any answer that demonstrates understanding of the underlying technology is fine. Example reasons to choose LLM: want a chat interface, want more reasoning abilities, want to do more than answer questions. Example reasons to choose traditional QA: want to focus on specific collection of research articles, have strong components and believe technologies like parsing and IE can increase trustworthiness, integrates well with search. Example reasons to choose MRC: want deep reading of specific papers, availability of training data and models. Other general desirable properties, not necessarily specific to any given method: need to be cheap to use, need to respond quickly, need to be easily extensible, need to respect copyright.

3. What are the five main components of a Question Answering system? Draw a flow chart that explains how these components combine, and how an input question leads to an output answer. (5 points)

   Refer to lecture notes for flow chart. The components are: Question Analysis, Search, Candidate Extraction, Knowledge Sources, Answer Scoring

Q2. Digital Humanities *20 points*

1. Human and computational intelligence differ in a number of ways. Describe a way in which they are *complementary*, and give a pair of examples illustrating when a human would offer greater insight than a computer, and vice versa (10 points)

   Humans excel at close interpretation in domains where reasoning crosses great distances: for example, positing a historical explanation for the cultural development leading to the authoring of an esoteric document. Computers excel at finding patterns in large quantities of material that humans would be incapable of inspecting directly, such as centuries of legal cases.

2. Graph convolutions are a simple generalization of the familiar "grid" convolutions used throughout HLT and computer vision. Explain this generalization, specifically addressing how the "receptive field" grows in the two architectures (10 points)

   In both architectures, an additional layer gives a location in the data access to information about more of its surroundings (i.e. a larger receptive field). In standard CNNs, the field grows according to the same pattern being applied at all locations, e.g. each letter's representation gets access to the representation of some number of neighbors to the left and right. In a GCN, each location has its own pattern: the connections it has in the corresponding graph. Therefore, another way of phrasing a GCN's receptive field is that each layer gives a node access to representations of nodes another "hop" away in the graph.

Q3. Human-Centered Evaluation *20 points*

1. Imagine you are evaluating a new summarization. You need human-annotated references for evaluating its performance. One option is to hire professional linguists to create the annotations, while another is to use a crowdsourcing platform like Amazon Mechanical Turk. What are the trade-offs of these two methods? And what are the potential issues of using reference-based method evaluation for such a system?

   Experts know the task, understand the evaluation criteria, and can provide more nuanced judgment on the model's output. This method ensures consistency and reliability in the reference summaries, but it comes with a higher cost and longer turnaround time. Crowdsourcing offers a more scalable and cost-effective solution. They can quickly gather large datasets from a diverse group of annotators, allowing for quicker evaluation across many examples. However, the quality of annotations can vary widely, as crowd workers may not have the expertise required to create high-quality references, leading to inconsistencies or lower-quality summaries. Potential issues in reference-based evaluation include inherent bias in the reference summaries. Since these references are manually created by humans, they may reflect personal preferences or subjectivity, especially in tasks like summarization, where multiple correct summaries may exist.

2. Explain the role of reliability and validity in the context of language technology evaluation.

   Reliability refers to the random measurement error—whether the results remain stable across repeated evaluations under similar conditions. For instance, an NLP system evaluated multiple times with the same dataset should yield similar scores to demonstrate reliability. Validity, on the other hand, assesses systematic measurement error, e.g., whether the evaluation measures what it is intended to measure. For example, evaluating a machine translation system based on fluency and accuracy rather than irrelevant metrics ensures validity. (No need to include examples)

3. What's the difference between empirical and analytical evaluation methods for language technology?

   Empirical evaluation involves gathering real-world data through user studies or experiments, focusing on how language technologies perform in practical scenarios with measurable outcomes. Analytical evaluation, on the other hand, relies on expert reviews, simulations, or theoretical models to predict system behavior without involving real users.

4. What's the difference between the normative benchmark and user study regarding human requirement realism, context realism, and pragmatic cost?

   Normative benchmarks rely on standardized user behavior and predefined ground truth references, which may not capture diverse human perspectives. In contrast, user studies involve actual users, accounting for more nuanced human requirements. While normative

benchmarks are cost-effective and scalable, they often lack the ecological validity of real-world contexts. User studies, though resource-intensive, offer richer insights by reflecting the specific context of system use and the nuanced needs of users.

Q4. Ethical Problems *20 points*

1. List four ethical principles in AI and briefly explain them (5 points).

   - *Autonomy: "Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."*
   - *Beneficence: People using your data should do it for your benefit*
   - *Non-maleficence: Not harm. Use Informed Consent to collect the data: people should explicitly approve the use of their data based on understanding.*
   - *Justice: Promoting prosperity, preserving solidarity, avoiding unfairness*
   - *Explicability: Enabling the other four Principles through Intelligibility and Accountability*

2. Explain what an Institutional Review Board is and its role to ensure an ethical AI (5 points).

   The IRB is responsible for protecting the rights and welfare of the human subjects of research conducted by researchers at the institutions. The IRB evaluates the ethical aspects of human subject research The board usually requires the investigators to inform the participants about the research in which they are involved ensuring a fair and correct informed consent process. IRBs try to ensure that the participants are correctly informed and accept the possible risks even if those are remote.

3. How does HIPAA (Health Insurance Portability and Accountability Act) try to protect citizen's rights? (5 points)

   HIPAA regulates the national standards that protect sensitive patient health information from being disclosed without the patient's consent or knowledge. The Privacy Rule standards address the use and disclosure of individuals' health information (protected health information - PHI) by entities subject to the Privacy Rule: Healthcare providers, Health plans, Healthcare clearinghouses, Business associates.

4. How can we measure (and avoid) maleficence or unfairness in AI? (5 points)

   We can try to explain algorithms to understand if these are being unfair or maleficent. There are several options, starting by ensuring transparency and reproducibility: make the code and models available. But the most important is to audit the models and code. Using algorithmic auditing processes could allow us to understand if a model is being unfair.

Q5. Computational Social Science *20 points*

1. What are two ways pre-trained language models like BERT or RoBERTa have been used for computational social science research? Provide both a broad description of the methodology and a specific example of how this method is useful for a particular research question, as in the provided example (8 points)

   Example: Prompting BERT-style models with template sentences can be used for metaphor detection. Metaphor detection can be used to identify dehumanizing language in political speeches and how it has varied over time.

   1.

   These models can be trained as supervised classifiers. Supervised classification can be used to detect emotions expressed on social media and analyze the role of emotions on social movements.

   2.

   They can be incorporated in topic models (e.g. contextual topic models). Topic modeling is useful for detecting patterns from large corpora, like ways Russian and Ukranian media outlets have reported about the war

   Additional answer: embeddings from these models can be used to scale people or other entities on ideological axes. Embedding-based scaling methods can be used to examine how people are described in news articles in terms of power and sentiment

2. What are two ways GPT-style models might be useful for computational social science research? As before, provide both a broad description of the methodology and a specific example of how this method is useful for a particular research question (8 points)

   1.

   These models can be used to label data without needing to train a supervised classifier. [Any example using labeled data is fine]

   2.

   They can be incorporated in topic models (e.g. TopicGPT). [Any example using topic model is fine]

Additional answer: they can be used to simulate human subject research or replace survey methodology. An example is using them for studies that would be unethical to run with humans, like the Milgram shock experiments

3. Label the following statements about topic models as true or false (4 points):

   In LDA, "topics" are defined as distributions over the vocabulary _____

   True

   The goal of LDA is to estimate K, the number of topics in the corpus _____

   False

   Topic models require in-domain annotated data _____

   False

   Topic models are useful for datasets where researchers have specific well-defined research questions _____

   False