Final Exam

601.467/667 Introduction to Human Language Technology

Fall 2021
Johns Hopkins University
Co-ordinator: Philipp Koehn

13 December 2021

Complete all questions.

Use additional paper if needed.

Time: 75 minutes.

Name of student: Yuyang Zhou

Q1. Question Answering *25 points*

1. Explain two ways in which the fields of Question Answering (QA) and Information Retrieval (IR) are different. (There are more than two possible answers; any two reasonable explanations will be sufficient.) [8pts]

   1. In QA, questions are in natural language sentences; but in IR, queries tend to be short keyword phrases.
   2. In QA, the answers are often short and to-the-point; in IR, the system returns lists of documents.

2. QA systems can be classified by the type of questions they are designed to handle. Please provide an example question for each type listed below. [12pts]

   (a) Factoid question:

   Who was the first President of the United States?

   (b) List question:

   List 10 countries that produce diamonds.

   (c) Definition question:

   Who is Alan Turing?

   (d) Opinion question:

   Why do people like Costco?

3. The following is an example from the Winograd Schema Challenge. Please explain why this kind of question is difficult to solve using current technology, and what kind of breakthrough might be needed to tackle these kinds of questions. [5pts]

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

   A. The trophy

   B. The suitcase

   Because in order to answer this kind of question, it's actually not something you can get just from understanding the syntax and semantics. The system really need to understand the physical world, in other word, commonsense.

   To tackle these questions, we can use some methods to prepared the system with knowledge we need, such as a hybrid approach that combined extraction of knowledge from the web using a probabilistic engine.

Q2. Digital Humanities *25 points*

1. Describe an advantage of employing unsupervised models for exploratory analysis [5 pts]

Unsupervised learning is very useful in exploratory analysis because it can automatically identify structure in data. Unsupervised models start with careful, unbiased representations and have no specific task, minimal bias and annotation cost.

2. Explain how graph-convolutional networks extend the principles behind standard CNNs [10 pts]

GCNs are a very powerful neural network architecture for machine learning on graphs. Graph-CNNs extend traditional CNNs to handle data that is supported on a graph. It extends CNNs from grids to graphs. Information can pass along edges and each layer allows nodes to see one further “hop”.

3. Describe a common issue with humanistic primary sources, and how the use of structured representations (e.g. SQL, RDF) help mitigate it [10 pts]

Common issue: information have so many different types and is wide, and don't have a specific structure;

We can use teh spreadsheets to mitigate this problem. Because humanists often have primary sources that might be document images or even physical records. And spreadsheets are more structured than just a simple list. It contains entities with properties and relationships. Humanists can describe their domain in a schema.

Q3. Clinical NLP *25 points*

1. What is the name of the law in the United States that protects sensitive patient health information from being disclosed without the patient's consent or knowledge? [6 pts]

   HIPAA: The Health Insurance Portability and Accountability Act of 1996.

2. Why are rule based methods popular in clinical NLP? (Select all answers that apply.) [6 pts]

   - ☑ Rules are easily interpretable
   - ☐ Federal regulation does not permit AI methods without FDA approval.
   - ☐ Most useful text is contained in images.
   - ☑ Doctors can often express medical knowledge as rules.

3. The task of clinical note segmentation refers to: (Select the best answer) [6 pts]

   - ☐ Dividing a paragraph into sentences.
   - ☐ Decomposing a biomedical term into constituent parts.
   - ☑ Splitting a note into sections, such as history, medications, chief complaint.
   - ☐ Dividing notes based on the date of the patient encounter.

4. Why is the task of negation important when processing clinical notes? [7 pts]

   Because negation is commonly used in clinical notes, and is important to have a better medical understanding of the patient. This is not something that we would see in a pronounced manner in regular language.

   So we will frequently encounter negation while processing clinical notes, and it's a hard challenge for us to handle the negation notes.

Q4. Ethical Problems *25 points*

1. Why do we need to ensure that AI algorithms take into account the principle of non-maleficence? [10 pts]

Because data is constantly being collected about us through cameras, location reporting or something else. Sometimes we just don't know that whether or not we can control data collected about ourselves, and maybe we don't like what the data says about us.
The concepts of human dignity and sanctity of life imply that the application of information technology in medicine must be beneficent and non-maleficent for the individual person. However, the specific AI use case of prognostication has the potential to violate these principles. So it is importane to ensure this.

2. How can federated learning help providing non-maleficent algorithms? [5 pts]

In federateed learning, instead of sharing data, the baseline models are adapted locally and then, shared. And all the adapted models are used to prepare a global new baseline. Federated learning enables multiple actors to build a common, robust machine learning model without sharing data, thus allowing to address critical issues such as data privacy, data security, data access rights and access to heterogeneous data.

3. Explain why we need to have *Explicability* in AI applications? Which measures would you follow to have more explicable systems? [10 points]

Explicability is important because opening up the black-box would not suffice to disclose algorithms' modus operandi, and the algorithms used in data science are always complicated. In addtion, the data often cannot be shared, and we also need to ensure teh transparency and reproducibility od the code. It can enable the other four principles through Intelligibility and accountability

Measures: Make the code available; When things "go wrong", we need to understand why; Propose algorithmic auditing processes (Ragi et al 2020).