2018 Fall CSE 538 Midterm-2
Due 11:59 am Dec 1st, 2018.

Points This exam is for 100 points.

1. This is an open ended exam. Write as much or as little as you want. I'd expect the whole exam to take somewhere between 8-12 pages total but that is a very rough estimation on my part! You won't be penalized for turning in a shorter or a longer document.
2. This is a test of whether you are able to apply concepts you have learned in class in some problem settings. For every answer, make sure you first convey your ideas at a high-level clearly. The more the details, the more convincing your answer will be. Draw figures where possible and clearly indicate the input/output to each component in your figure.
3. Questions are intentionally a bit underspecified. You can make assumptions as you need. In some cases I specifically mention some things that you cannot assume.
4. Take home only means that you can do this at your leisure but the solution still must be yours and yours only.
5. You cannot discuss ideas with your friends. Note it is very easy to find out if you discussed ideas.
6. You can use material we discussed in class. Lecture slides, notes and ideas contained in them can be reused but you should use your own articulation of these.
7. You can use ideas from existing papers but you will have to cite those papers. If you use ideas from any paper or lecture notes, mention it right alongside your answer.
8. Please turn in a typed assignment as a pdf file. You can use hand-written equations or figures in the pdf (scanned in) but please make sure they are legible in the final pdf.
9. **The solutions are due 11:59 am Dec 1st.**
10. **You can submit until 1:59 pm Dec 1st. You will incur a 10 point penalty for submitting after 11:59 and a 20 point penalty for submitting after 12:59. It is up to you to ensure that you submit early enough that you don't get caught in last minute issues. You have 24 hours after all!**

11. **What if I have questions?** Please post your questions here. We will answer them once 9pm Nov 30th and once again 9am Dec 1st.

    `https://goo.gl/RtYieZ`

    Do not email questions or post on Piazza.

1 Dependency Parsing (30 points)

Congratulations! Bloomberg just hired you because of your excellent background in NLP. They have been struggling with their dependency parser that they built for parsing news headlines.

Here are some examples:

1. `Stolen painting found by tree.`
2. `Eye drops off shelf.`
3. `Miners refuse to work after death.`

Propose a solution to develop a parser for headlines.

1. Collect five headlines that fail to parse correctly with a dependency parser that is trained on news stories but not on headlines. You can use `https://demo.allennlp.org/` as such a parser. Try to modify these sentences so that they convey the same meaning but are now proper sentences that parse correctly. Show the five headlines, the edited sentence you made and point out the reason based on this edit as to why the original headline did not parse. Summarize the reasons for these failures in terms of what we know about parsing. (10 points).
2. Bloomberg's in-house journalists tell you that in most cases a headline has a matching sentence in the story – say the first sentence is often a repeat of the headline. How will you use this information to improve the parsing of headlines if you had access to the first sentence of a story in addition to the headline. (10 points).
3. Propose one other idea for improving the parser. Argue why this idea should work using the examples you found above. (10 points).

2 Medical Relation Extraction (35 points)

You are assigned the task of building a relation extractor for the medical domain. Your inputs will be medical abstracts that you can download from public repositories such as PubMed.

Check out `https://www.ncbi.nlm.nih.gov/pubmed/?term=diabetes` to see a list of articles on diabetes.

Your outputs should be instances of three types of relations:

1) `treats(drug, disease)`

2) `side-effect(drug, disease)`

3) `causes(organism, disease)`

where `treats` is meant broadly when a drug is used to manage, treat or reduce severity of a disease or medical condition.

Your task is to specify all the steps involved in creating a relation extractor that can find new instances of these relations from the medical abstracts.

2.1 Feature-based Extractor Solutions (15 points)

Assume that the relations are often expressed within a single sentence for this subproblem. You should provide solutions for two types of scenarios:

1. Handful of examples – You only have three instances of each known relation. (5 points)
2. Large database of examples – Say you have access to a database that provides tens of thousands of examples for each type of relation (no sentences, just the relations in the database). You have three tables that provide you these instances. (10 points)

In both scenarios you are a poor assistant professor with no money to do any annotation.

For both scenarios, provide details on how you will obtain training data (without manual annotation), how you can identify the target entities (you will have to build this entity identifier as well – no separate training data for this), what types of features you will use (look at the sentences in the abstracts and come up with some features based on those, use example sentences to explain your features), and any challenges you anticipate in building your model. Use a figure to indicate the various components in your application.

2.2 Extraction across sentences (20 points)

One of the challenges in the medical domain is that relations are rarely expressed in a single sentence. One sentence might give partial information about the relation, and another might give another piece of information about the relation. Consider the following sentences:

```
Group A were cohorts with Diabetes II and Group B were cohorts
with Diabetes II. Amphetamol was administered to both groups.
Both groups showed remarkable reduction in blood pressure within
twenty-four hours.
```

These sentences together express the relation `treats(Amphetamol, blood pressure)` is true. Propose a solution for relation extraction that handles these kinds of relations. You can propose a neural model for this problem. Draw a clear figure and show the dimensions of the outputs of each component in your neural model. You have access to the large database of examples and have to still figure out how to get annotated training data for your neural model (without any manual labeling)

3 Machine Translation (35 points)

Your goal is to design a MT system that can translate French sentences (source) to English (target) sentences. In IBM Model 1, we used EM to estimate $Pr(f_i|e_j)$.

3.1 A third language (5 points)

Suppose you received extra training data that includes not only French and English sentence pairs but also translation in German. So you have training data that looks like $\{(f^{(1)}, e^{(1)}, g^{(1)}), (f^{(2)}, e^{(2)}, g^{(2)}), \cdots, (f^{(n)}, e^{(n)}, g^{(n)})\}$ Your goal is to improve the translation quality from French to English, using this extra translation in German.

- Describe a noisy channel model which now includes the German sentences i.e., specify a generative story involving German sentences as well. You are free to design this however you like but you have to justify your specific choice with some explanation. (5 points)
- Sketch the modified EM algorithm that accounts for this new generative story. State the model equation and describe what should be done in each step of EM (Expectation-step & Maximization-step) in English. Then, derive the equations for each step. (You do not have to provide the details of expected counts since we did not cover them in class.) (5 points)
- Provide a justification for why having the third language can be useful. (5 points)

3.2 Translating headlines (10 points)

Suppose Google hires you for translating news headlines. Look at the following headlines:

- Key GOP lawmakers flip on health care after meeting.
- Spicer says part of spending bill will go to border wall.
- Ridge foster dad found not guilty of sex abuse charges.
- DWI suspect four times the legal limit with kids in car.
- Confident politician says he wants to 'prove them wrong' and get a Mideast peace deal.

Translate them through Google translate into a language other than English. What kind of errors do you observe? Give three hypotheses that explains some of these errors.

Now test each of your hypotheses by creating additional headlines that also fail in a similar fashion. Provide some suggestions on what kind of information need to be modeled to address these errors.

* Depending on the target language, some translations might not fail. In such case, 1) make use of the other headlines that fail, or 2) pick another language where these headline translations would fail. For example, all of above headlines fail for Korean. 3) If none of the above works, you can pick other headlines of your choice from any news media. If you choose to do so, note this in your answer clearly.

3.3 Knowledge-backed Machine Translation (15 points)

News articles are often written for people from a specific region. This means often assumptions can be made about what people in the region would know about. For example, consider the following headline.

```
Bloomberg announces presidential run.
```

This works fine for the local audience in New York or even within the US but if this headline were to be translated to Telugu, an Indian language, then a bit more context might be useful. The translation should ideally include some background information about Bloomberg. For example a translation should be the Telugu equivalent of the following sentence

```
Bloomberg, the former NYC mayor and owner of Bloomberg News,
announces presidential run.
```

Design a knowledge-backed machine translation system on top of an existing seq2seq translation model to address this challenge.

Hint 1: Assume that you have access to knowledge bases that contains relations about entities, which are basically facts (e.g., mayor(Bloomberg, NYC), owner(Bloomberg, Bloomberg News)). This provides you a way to include useful information in the translation.

Hint 2: You will have to however think about what information is relevant for the current sentence. To figure out what is relevant information you can use some kind of distant supervision assumption. Is there a way to find sentences about entities that contain additional background information about them?

Describe your solution in enough detail so that I can understand your ideas and solution. Please clearly specify the components in your model and how you will train them.