Introduction to Computational Linguistics

Section

Olga Zamaraeva

University of Washington

April 17, 2020

Plan for today

- Writing (and reading!)
  - General rule: No good writing can be done with zero reading
- Particularly as it relates to your write ups!
- Important reminder: Your final project is worth 30% of your grade!

Reading highly technical material

- Identify the main idea/topic of the section/paragraph
- Try to relate this idea/topic to the bigger picture: Why is it important? How/where is it used?
- Articulate questions as you are reading
  - Actually write them down
  - Go ahead and post them if you like!
- Reread some of the more confusing sections several times (it **does** help)
- Do not get bogged down; you do not have to understand every single word every single time

Writing formal write ups

- Any write up should have an introduction and a conclusion
- The prompts are only to get you started
- A write up is not an answer to a question
- A write up is self-sufficient:
  - It situates (introduces) the topic
  - It discusses the topic in a focused manner
  - It makes references to class materials (book, lecture)
  - It draws meaningful connections between theory and the assignment
    - Canvas discussions are your friend!
  - It concludes with a summary of what was said and why.
- A good write up is well-edited and spell-checked
  - You will never produce a good write up with only one round of revision

A write up is not an answer to a question

- The prompt questions are only there to get you started
- Will not get full credit:
    - *I did not learn anything new about regular expressions.*
- May get full credit:
    - *The following aspects of regular expressions are relevant to linguistic concepts which we are studying in this class. First... Second... Finally, ...*

A good write up introduces the topic and the structure

- Do not start your write up as if you were in the middle of conversation
- A write up starting with the following will not get full credit:
    - *It was not difficult to implement this program.*
- May get full credit if starts with:
    - *In assignment 1, I implemented a chatbot Elizalike (Weizenbaum, 1966) using regular expressions. While certain aspects of English language, particularly morphology, made this easy, in general using regular expressions to model natural language presents challenges...*

A good write up is focused

- You know the drill:
  - Apart from intro and conclusion (which introduce and summarize the structure), each paragraph is focused on one idea, and there are transitions between paragraphs.

A good write up uses class materials

- The lecture, the book, and the assignments are related to each other in important ways.
- If you don't see ways in which they are related, you need to ask more questions and/or reread/rewatch the materials.
- Identify theoretical ideas which the assignment is related to and make sure to tie them into your write up.
- Example: What are the limitations of regular expressions in the context of modeling natural languages?
  - Book/lecture: Regular expressions describe regular languages (what does that mean?); natural languages are not regular (why?)

A good write up uses class materials - contd.

- Assignment 2 write up:
  - What did you learn about morphology/phonology?
  - What's a linguistic hypothesis?
  - Regular languages (again!)
  - FST definition
- All of the above were covered in lecture/book to some degree, and a good write up will show this knowledge
- Again, ask questions if you are not sure how to draw a meaningful connection
- ...but do not do your write up ad hoc; you won't get full credit.

A good write up concludes

- Without a conclusion, it usually remains unclear why you wrote what you wrote and what you wanted to say
- A write up ending abruptly will not get full credit
- Conclude with a summary of what you said and why
  - The “why” here is motivational; you always pick a certain angle; present your particular angle (logic, examples...) in a meaningful way.

A good write up is well-edited

- Please!
- We are only asking for a couple pages (8 for the final project); please do at least a couple rounds of editing to ensure all the above requirements are in place
- Run a spell checker at the very end

How to procrastinate writing less?

- Editing is easier than writing (for most people)
- The “dog” draft:
    - The “rough rough” draft
- Write absolutely anything on the page at first, even if it’s “I hate this stupid assignment!”
- You’ll be amazed how you will suddenly feel like doing more!

The Final Write Up

- The project costs **30% of your final grade**
- The **writing** costs 50% of that!
- Do not lose this much of your grade by throwing together an ad hoc write up!
- All parts of it will be equally important:
  - The introduction and the conclusion
  - Presentation of the data and the error analysis
  - Discussion
  - Meaningful references to material learned in class

Final Project: Data

- You will be working with a paper about some NLP tool
- It is crucial to clearly present the data that was used:
  - By **the authors** of the paper
  - By **you**, in your project
  - When the above are the same, it should be clearly stated
- Data always deserves **its own (titled) section**
- Data are usually presented as **tables** (or graphs, etc.), accompanied by a clear, focused, succinct explanation
- Do not describe tables in words

Do not describe tables in words

Language	Language Family	Tokens (millions)
English	Indo-European	1,000
Russian	Indo-European	2
Turkish	Turkic	1,000
German	Indo-European	0.5

Table: Languages used in the project

- Will not get full credit:
  - *Four languages were used: English, Russian, Turkish, and German. English, Russian, and German are Indo-European while Turkish is Turkic. There was 1 billion English tokens and as many Turkish tokens, 2 million Russian tokens, and only 500,000 German tokens in the dataset.*

Summarize tables in a meaningful way

Language	Language Family	Words (millions)
English	Indo-European	1,000
Russian	Indo-European	2
Turkish	Turkic	1,000
German	Indo-European	0.5

Table: Languages used in the project

- May get full credit:
  - *The dataset used by Doe et al. (2019; Table 1) consists predominantly of Indo-European languages. English is particularly overrepresented in terms of the number of word tokens. While the Turkish part also amounts to a billion words, due to Turkic morphology, fewer distinct lemmas may be represented, which would account for the results we discuss in Section 3. We were not able to get access to the full dataset that Doe et al. (2019) used, and the subset which we used to get the results described in Section 3 is presented in Table 2.*

Final Project: Results

- That paper will have some **results**
    - Usually in terms of precision/recall/F-score
- Same applies: tables, clear and focused commentary; own titled section/subsection
- You will try to reproduce their results
- What you get goes into a **separate** table, with a clear comment, especially if the tables differ

Final Project: Results vs. Errors

- Your project should be focused on Error Analysis
  - Categorizing errors in some meaningful way and reflecting on what the implications are, and how it can help direct further efforts
- Sometimes the authors will have done their own EA; in this case you must do something different
- Your EA will also be a **table**

Final Project: Results vs. Errors

- Your project should be focused on Error Analysis
    - Categorizing errors in some meaningful way and reflecting on what the implications are, and how it can help direct further efforts
- Sometimes the authors will have done their own EA; in this case you must do something different
- Your EA will also be a **table**
    - It is **not the same as the Results table**

Final Project: Results vs. Errors

- Your project should be focused on Error Analysis
  - Categorizing errors in some meaningful way and reflecting on what the implications are, and how it can help direct further efforts
- Sometimes the authors will have done their own EA; in this case you must do something different
- Your EA will also be a **table**
  - It is **not the same as the Results table**
  - It is **not the same as the Results table**

Final Project: Results vs. Errors

- Your project should be focused on Error Analysis
  - Categorizing errors in some meaningful way and reflecting on what the implications are, and how it can help direct further efforts
- Sometimes the authors will have done their own EA; in this case you must do something different
- Your EA will also be a **table**
  - It is **not the same as the Results table**
  - It is **not the same as the Results table**
  - It is **not the same as the Results table**

Final Project: Results vs. Errors

Language	Precision	Recall
English	0.99	0.91
Russian	0.66	0.75
Turkish	0.67	0.34
German	0.95	0.71

Table: Results

Error category	Percentage
Morphological	24
Syntactic	36
Semantic	40

Table: Errors across all languages

NB: The above EA table is too general and will only get full credit if accompanied by clear definitions and examples, for all error categories, and preferably by additional, more detailed tables, including breaking out by language.