LING 227/627 — Language & Computation I

Final Project
Due May 14, 2021

As I discussed at the beginning of the term, you must complete a final project for this course. The following are expansions of points with made on the syllabus regarding this project:

- You must work in a group of 2 or 3 people. The amount of work expected from this project will, naturally, vary by the number of people who contribute, but you may find it helpful and interesting to work with others who have backgrounds different from your own.
- Your project should build in some way on the material we have covered in class this term. The range of allowable topics is very flexible: it may relate to a personal passion of yours or to an existing area of research. I would encourage you to try to find a topic that inspires you (and your partners). To get ideas about topics, you might want to scan through papers from recent Computational Linguistics conferences (and their associated workshops). Some to look at are (you can also get access to many of these through the ACL anthology):
  - Association for Computational Linguistics (ACL)
  - Empirical Methods in Natural Language Processing (EMNLP)
  - North American Chapter of the Association for Computational Linguistics (NAACL)
  - Conference on Computational Natural Language Learning (CoNLL) - you may be especially interested in the annual shared tasks, which provide a specific research problem (dependency parsing, named entity recognition, grammatical error correction, etc.) and data sets to use.
  - Society for Computation in Linguistics (SCiL)
  - Neural Information Processing Systems (NIPS)

  Don't be afraid to get creative! And please come to talk to me if you have any questions about project topics. It is important that you pick a topic that will be doable in the time you have.
- Your project must include an implementation of some sort that is tested on some set of data. Your results do not need to be state of the art, but you should try to optimize the system's performance as much as possible, given time constraints.
- You must produce a written report (at least 5 pages long) that covers the following points (the first four are the most crucial):
  1. A clear statement of your research question or the problem you are solving.
  2. A description of your approach, any algorithms you developed, and/or existing algorithms you are using, as well as a description of your implementation.
  3. A discussion of how you tested your implementation. What datasets did you use? How did you evaluate its success?
  4. A presentation of your results, with relevant tables/graphs/etc.
  5. An accounting of alternatives you tried that weren't successful (with ideas on why not) and alternatives you didn't get the chance to try, but which you think might be promising.
  6. A short discussion of how your work relates to what has been done before (with appropriate bibliographic references).
  7. A discussion of the linguistic assumptions of your approach, and their viability.

Your report should follow the standard format for papers in computational linguistics, which you can discover by looking at papers from some of the conferences linked above. Both LaTeX style files and a Word template for the ACL paper formatare available at this GitHub repository.

- You will need to submit your code, which should be commented and readable. You should provide a README file that indicates any dependencies that exist and how your implementation can be run.
- Each group will make a short (maximum 10 minute) oral presentation to the class about the project on December 15th (in the final exam time slotat 2pm) and answer questions from classmates about the project.

Tools and Data Sources

For this project, you are free to make use of any existing software toolkits or datasets. There are some good lists of resources out there on the web: this list was put together by Chris Manning at Stanford, and this one was put together by Paul Dixon.

Some particularly noteworthy pieces of software to look at are:

- NLTK - a general purpose NLP toolkit in python
- Open NLP - a general purpose NLP toolkit in java
- spaCy
- Stanford CoreNLP
- SRI Language Modeling Toolkit - includes ngram and HMM models
- Open FST - a toolkit for manipulating weighted and unweighted finite state transducers
- gensim - vector semantics (neural embeddings) and topic modeling
- scikit-learn - Machine learning toolkit in python

You will probably also want to make use of some natural language data in developing and testing your system. A number of data sources are given in the curated lists mentioned above, but here are some which either deserve special attention or which are not publicly available but are available through the Yale library via Yale's membership in the Linguistics Data Consortium. You can find these by searching in the on-line catalog for the string 'Linguistics Data Consortium' and selecting the category 'Software and Datasets'.

- Lexical databases
  - CELEX - available via the Yale library website
  - Wordnet - a hierarchically organized semantic lexicon
  - Framenet - a semantic lexicon indicating the argument-structure of lexical items
- Parsed Corpora - Syntax and Semantics
  - Penn Treebank Corpus - available via the Yale library website. A parsed corpus of English text of a variety of genres.
  - Propbank Corpus - available at `http://propbank.github.io`. A semantic annotation of the Wall Street Journal subpart of the Penn Treebank and other corpora, indicating predicate-argument relations.
  - Penn discourse treebank - available via the Yale library website. A discourse annotation of the Wall Street Journal subpart of the Penn Treebank, indicating semantic relations between sentences.

  - Universal Dependencies treebank - Dependency-parsed text in over 100 languages.
  - RST discourse treebank - available via the Yale library website. An annotation of (part of) the Wall Street Journal subpart of the Penn Treebank, indicating hierarchically structured rhetorical relations between clauses.

- Other corpora

  - Childes database - corpora of child speech and child-directed speech in a variety of languages (some parsed, some part of speech tagged)
  - Europarl - parallel corpora of European parliamentary proceedings in 21 European languages.
  - Sentiment analysis datasets from Lilian Lee.
  - Kaggle - a repository of datasets and competitions for a variety of machine learning problems

There is lots more lurking out there on the internet, and you are free to scrape your own if you don't find what you need. But keep in mind that you do need to finish this project, so don't spend too much time on the data gathering aspect of this project.