Intro to Computational Linguistics: Final Lecture

Adam Meyers
New York University
Summary

• Three more short talks (15 minutes)
• Ross' Deep Learning Demo (5 minutes)
• First Draft Review & Final Project Comments
• Discussion of Class Structure
  – Feedback on Final Project vs Final Exam
• Course Topic Summary
  – and Relations to Final Projects
• Research that I am Working on Now
First Draft of Final Projects

• Comments Posted
  – Goal = Advising you on how to complete final project, including what the write-up should look like
• For programming/annotation projects, the write-up is as important as the code
  – I cannot always run student systems
    • It is often designed to run on their system and there can be unexpected incompatibilities with my system
    • It may take a long time to run
  – Include small samples of input and output in the write-up
  – The write-up should explain what you were trying to do, why you tried to do it, how well you did and how it fits in with previous work
Final Project Write-Up
(Programming/Annotation)

• Structure like a paper
  – https://cs.nyu.edu/courses/spring18/CSCI-UA.0480-009/final_project_possibilities.pdf
  – approximately 5-10 pages
  – cite sources and include bibliography

• Label Sections – Organizes Your Description

• Common Section Titles (only some may be applicable)
  – Introduction, Previous Work, Data Preparation, Task Description, Our System, Results, Discussion, Error Analysis, Conclusion, Future Work, References

• Compare your results to state of the art
  – even if your system is not competitive
Final Project Write-Up Goals (Programming/Annotation)

• What is the problem and why is it interesting
• How did you attempt to solve the problem
• How your method compares with previous methods
• What are your results? How do these compare?
• Why did your approach not work well for certain examples?
• What does your work demonstrate?
Final Project Papers

• Normal rules of writing a “good” paper as mentioned previously

• There are several paths for survey papers
  – Use a survey of previous work to argue for a new approach: the paper is essentially a proposal
  – Argue for a particular critique of current work.
  – It may help to find specific examples from text that back up your claim.
  – It may help to run (open source) systems and analyze the output
Class Structure Experiment

- Previous Semesters:
  - Proposal due 1 month before it was due
  - No first draft submission
  - Final Exam and Final Project

- Current Semester
  - Proposal due 1 and 1/2 months before Final Project is due
  - 1st draft submitted about 3 weeks before Final Project is due
  - No Final Exam

- Differences
  - Final Project: More Time and More Feedback
  - Less Emphasis on Breadth of Knowledge:
    - Readings and Attendance in 2nd Half of Semester had a smaller effect on the Final Grade
Topics Covered Touched On in at Least One Final Project

• Information Retrieval Related Projects (including vector-based methods, TF-IDF, cosine similarity, etc.)
  – Document Classification
  – Sentiment Analysis
  – Question Answering
  – Document Summarization
  – Language Identificaiton
• Information Extraction
  – Event Extraction
  – NE Extraction
• Corpus Linguistics
• Annotation
• Segmentation/POS Tagging
• Machine Translation
• Statistics: TF-IDF, N-grams, Cosine Similarity, etc.
• Sequence Labeling (BIO)
• Evaluation Methodology
• Machine Learning
• Semantic Role Labeling
In-Class Topics Not Reflected in Final Projects

- Parsing
- Feature Structures
- Reference Resolution
Research I am Working On: Termolator

  – Online Journal Paper Describing Termolator program
    • Combine info 2 conference papers
    • Updated Features
    • Comparison with Termostat:

• Possible Next Steps in Summer:
  – Tune Metrics for Distributional System
  – Improve Chinese system
  – Create a Spanish system

• http://nlp.cs.nyu.edu/termolator/
• https://github.com/AdamMeyers/The_Termolator
• https://github.com/ivanhe/termolator/
SelecT (Machine Translation)


- Given several MT systems (neural, phrase-based, rule-based) and a Set of Source Sentences
  - For each sentence, choose the system that is most likely to provide the best translation
  - The result will be output from several translation systems, that together, would be superior to the output of a single system

- Training
  - Calculate BLEU scores on translation output for training set of sentences
  - Calculate statistics of features of the source sentences that correlate with quality translation

- Evaluation
  - Run each MT system on test or dev source sentences
  - Compare results of choosing:
    - all translations from one system
    - the worst possible selection of translations
    - the best possible selection of translations
    - The translations selected by SelecT

- Results (still tuning):
  - The best SelecT system does better than any of the individual systems (between the best and the worst)
    - SelectT: 38.01 Bleu, 49.55 WER
    - Best and Worst: 18.97 to 40.08 Bleu; 63.91 to 46.7 WER
    - Best System (Neural): 37.36 Bleu, 51.77 WER

- Further application of a Fuzzy Match repair system (system that identifies and fixes “holes” in translations based on translation memory)
Web of Law

• For Bloomberg Law
  – Working on an annotation system and (eventually) an IE system for classifying features in contracts

• Working on paper and system about document classification
  – Classify Supreme Court Decisions into topic categories using a neural network based system.
  – Work primarily by Samir Undavia, a Case Western student, who was a visiting student at NYU.

• Manual Rule Based Corpus Preparation
  – Cleaning up Supreme (and other Appelate) Court Decision documents
  – Identifying citations to cases and legislation
  – Identifying other entities
  – Identifying relations among entities
  – Creating citation graphs for search

• https://nlp.cs.nyu.edu/meyers/web_of_law_documents/
Concluding Remarks

• For further guidance on Final Projects
  – I will be available during my normal office hours (and by appointment) through May 15
• I would appreciate feedback about the course structure – to help improve future versions of this class.
• I included a summary of what I am working on
  – If you are interested in collaboration, please feel free to contact me.