CSCI 662 Final Project 2017

Important Dates

- Nov 14: Initial project proposal due at beginning of class (on Blackboard).
- Nov 21: Final project proposal due at beginning of class (on Blackboard). This should include a report of your data preparation and baseline, which should be completed by this time.
- Nov 28-30: Interim project presentation (in class).
- Dec 11: Final project write-ups due (on Blackboard).

Requirements

- You may work in pairs, but not in groups of three or more. An individual project is roughly double the size of an average homework assignment, and a group project is roughly double the size of an individual project.
- Choose a topic for your project:
  - You may not do the same project for this class and another class. But we allow (and encourage) you to choose a project that is part of a larger research program.
  - The topic should have something to do with statistical learning of the structure of natural language. It should treat language as **more than a bag of words**, and it should learn a model instead of just measuring something like cosine similarity.
  - Please, no information retrieval (IR) projects.
- The initial proposal (**<lastname>-initial-proposal-writeup.pdf**) must include at least the following:
  - A clear statement of the **goal** of the project, and **what would constitute success**.
  - Concrete description of the **data** that you will use, and what processing and organization is needed to make it useable. What data will you use for evaluation?
  - Description of the **evaluation method** that you will use.
  - Description of a **baseline method**, i.e., something that you can implement in an hour to attack the problem.
  - Description of the **method** you propose to use.
  - A sample results chart (or charts), with blank cells. It may be a 2x2 chart with labeled rows and columns, or maybe a 3x4 chart, or two charts. An example:

	Task Accuracy
Baseline	
HMM	
CRF (with small feature set)	
CRF (with large feature set)	

    This will help you realize whether your project is concrete enough (right now), and then keep you from getting lost (while you are working on the project). “Task Accuracy” is only an example -- you may be measuring something different, such as “Running time (minutes)” or “% of pairs where human judge preferred system output over human output” or “F1-score”.
- The final proposal (**<lastname>-final-proposal-writeup.pdf**) additionally includes, on the basis of feedback from the instructor:
  - A description of the **collection/cleanup of your data**.
  - A description of the **implementation/evaluation of your one-hour baseline**.
- Present in class an interim project presentation. This doesn’t need to include final results; it’s an opportunity for you to share your topic with your classmates and to get feedback.
- Prepare a final report (**<lastname>-final-project-writeup.pdf**), which should include the same sections as the proposal, plus your **results** and **conclusions** that you draw from those results. Please additionally submit any **code** or **data** that are important for us to evaluate your work.

Sample topics

- Here is a sample of possible topics, some of which are from past years. This is meant to give you an idea of what kinds of topics would be reasonable. You are encouraged to think of something outside this list that is especially interesting to you.
  - Automatically decipher the Copiale manuscript (data at http://stp.lingfil.uu.se/~bea/copiale)
  - Unsupervised or discriminative context-free parsing. Data: HW4 or Penn Treebank.
  - HMM word alignment. Data: Canadian Hansards, UN or EU proceedings.
  - Automatically correct mis-heard song lyrics. Data: www.kissthisguy.com.
  - Identify correct logical form. Data: manually selected sentences in some domain.
  - Learn phoneme changes across a pair of related languages. Data: cognate pairs extracted from dictionaries.
  - Mad Gab generation (language game). Data: CMU pronunciation lexicon.
  - Convert natural language to image schemas. Data: preposition labels and NL descriptions for scenes.
  - Translate passages from Dante’s Divine Comedy from Italian into English while maintaining verse and/or rhyme. Data: original text of Divine Comedy, plus CMU pronunciation lexicon.
  - Show that the weighted intersection of two RNNs can be represented by a single third RNN.
  - Learn to select examples to maximize an automatic student’s test score (using reinforcement learning).
  - Given a set of Unicode punctuation marks, build a universal tokenizer that decides how to separate punctuation from words, for any language.
  - Input: Wikipedia page. Output: Quiz to test a human reader’s comprehension of that page, with automatic scoring.
  - Foreign language writing assistant.
  - Meaning-to-text generation (data at amr.isi.edu + 40k additional sentences from LDC)
  - Super-tagging for CFG parsing (tags include parent information)
  - Unsupervised pronunciation of alphabetic scripts.
  - Automatic corpus cleaning.
  - Training a tagger for maximum accuracy (vs. maximum likelihood).