CSCI 562, Fall 2012
Final Project

2012 Nov 8

1 Important Dates

- Nov 15: Initial project proposal due at beginning of class.
- Nov 22: Final project proposal due at beginning of class. This should include a report of your data preparation and baseline, which should be completed by this time.
- Dec 4, 6: Interim project presentation (in class).
- Dec 19: Final project write-ups due on Blackboard.

2 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.
  - A natural topic could be based on homework assignments: either combining two homeworks, or improving upon one homework (see below for examples based on homework).
  - 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.
- The **initial proposal** must include at least the following:
  1. A clear statement of the **goal** of the project, and what would constitute success.
  2. Description of the **method** you propose to use.
  3. Concrete description of the **data** that you will use, and what processing and organization is needed to make it useable.
  4. Description of the **evaluation** method that you will use.
  5. Description of a **baseline** method, i.e. something that you can implement in an hour to attack the problem.
- The **final proposal** additionally includes, on the basis of feedback from the instructors,

3'. A description of the collection/cleanup of your **data**.

5'. 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 just an opportunity for you to share your topic with your classmates and to get feedback.
- Prepare a **final report**, which should include the same sections as the proposal, with results, plus

  6. Conclusions that you draw from your results, and
  7. Pointers to any code or data that are important for us to evaluate your work. Please do not submit code or data by e-mail.

3 Example topics

Here is a sample of possible topics, many 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: HW5 or Penn Treebank.
- HMM word-alignment. Data: Canadian Hansards, UN or EU proceedings.
- Chinese or Japanese word segmentation, either supervised or unsupervised. Data: Penn Chinese Treebank.
- Automatically correct mis-heard song lyrics. Data: `www.kissthisguy.com`.
- Identify correct logical form. Data: manually selected sentences about human heart function.
- Unsupervised part-of-speech tagging. Data: Penn Treebank.
- Learn phoneme changes across a pair of related languages (Uzbek and Turkish). Data: 1094 cognate pairs extracted from dictionaries.
- Mad Gab generation (language game). Data: CMU pronunciation lexicon.
- Transliteration of Greek from Greek alphabet to Latin alphabet. Data: 5000 Greek words in Latin script taken from discussion forums.
- Translate between ancient Greek (morphologically rich, free word-order) and English. Data: Perseus Project, 7 million words.
- Convert natural language to image schemas. Data: 2129 preposition labels and 200 NL descriptions for 89 scenes.
- Translate passages from Dante's Divine Comedy from Italian into English, maintaining verse. Data: original text of Divine Comedy, plus CMU pronunciation lexicon.