

CS 562 Final Project

David Chiang Liang Huang

27 Oct 2009

1 Important Dates

- Tue Nov 10: Initial project proposal due at beginning of class.
- Thu Nov 19: Final project proposal due at beginning of class. **Data preparation, baseline finished (and should be included in the final proposal).**
- Thu Dec 3: Midway project presentation (in class).
- Wed Dec 16: Final project write-ups due via e-mail to oana@isi.edu.

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.
 - Every group's project must be different.
 - The topic should have something to do with statistical processing of the *structure* of natural language. We do **not** allow boring topics that do not exploit the structure of language, for example text categorization with bags of words.
- Hand in an **initial proposal** for your topic (due Nov 10). It 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.
- After getting feedback from the instructors, complete the **final proposal** as well as collection/cleanup of your **data** and implementation/evaluation of your **one-hour baseline** by Nov 19.
 - Present in class on Dec 3 your **midway project status**.
 - Carry out your proposed research and hand in a **final report**, due Dec 16. The report should include the same sections as the proposal, with results, and
 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 (except for those with a ★). These are not off-limits, but remember that the instructors make the final decision to approve each topic and expect some originality. Please talk to the instructors and TA (preferably after class) about your topic before writing the initial proposal.

- (★) English respacing (word-segmentation). You have to compare supervised with unsupervised approaches. Data: any sizable English text.
- (★) Chinese or Japanese word segmentation, either supervised or unsupervised. Data: Penn Chinese Treebank.
- (★) Back-transliteration from Mandarin Chinese or Cantonese (we did Japanese Katakana decoding in HW3 and HW4), either supervised (if you can find the annotated data) or unsupervised (recommended). Data: list of Chinese transliterations of foreign names; CMU pronunciation dictionary; Chinese pronunciation dictionary.
- (★) Unsupervised context-free parsing. Data: HW5 or Penn Treebank.
- (★) HMM word-alignment. Data: Canadian Hansards, UN or EU Proceedings.
- (★) Discriminative context-free parsing. Data: HW5 or Penn Treebank.
- (★) Dependency Parsing, either PCFG or discriminative. Data: HW5 or Penn Treebank.

- Translate Korean pronunciation for Chinese words into Japanese pronunciations. Data: collected manually from newspapers.
- Apply genetic programming to generate context-free grammars or tree substitution grammars. Data: from HW5.
- 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.