

CS 562 Final Project

1 Nov 2007

1 Important Dates

- Nov 13: Initial project proposal due at beginning of class.
- Nov 20: Final project scope settled. Data preparation, baseline finished.
- Dec 17: Final project write-ups due via e-mail to `sdeneefe@isi.edu`.

2 Requirements

- You may work in pairs, but not in groups of three or more.
- Choose a topic for your project:
 - You may not do the same project for this class and another class. But you may choose a project that is part of a larger research program.
 - Every group's project must be different.
- Hand in a proposal for your topic (due Nov 13). 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, finalize proposal and complete data preparation and baseline method by Nov 20.

- Carry out your proposed research and hand in a report, due Dec 17. 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 Topics from last year

These topics are not off-limits, but remember that the instructors make the final decision to approve each topic and expect some originality.

- Translate Korean pronunciation for Chinese words into Japanese pronunciations. Data: collected manually from newspapers.
- Apply genetic programming to generate regular tree grammars. Data: from homework assignment.
- Automatically correct mis-heard song lyrics. Data: www.kissthisguy.com.
- Classify a document as belonging to one topic or another. Data: UCI KDD archive, goarticles.com, articlesbase.com.
- Minimize a weighted finite-state transducer. Data: CMU pronunciation lexicon.
- Identify correct logical form. Data: manually selected sentences about human heart function.
- Unsupervised part-of-speech tagging. Data: Penn Treebank.
- Classify a message as belonging to one speech act or another. Data: from manually annotated USC Blackboard discussion threads.
- 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.
- Document retrieval. Data: WSJ or TREC.
- 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.