Homework #6: CMPT-413

Distributed on Mar 21; due on Apr 4

Anoop Sarkar – anoop@cs.sfu.ca

This homework is different from the previous ones. You can choose a project of your choice that is related to computational linguistics, preferably from the topics we have covered in this class. The following a list of ideas for projects, but you can choose topics from outside this list. If you are uncertain about your idea for a topic please discuss the idea with me. You should submit the working program along with the necessary data, and also a 2-4 page write-up (single column) describing your project. Include your name and student number in your submission and in the text of the project description.

(1) nethack (http://www.nethack.org/) is an open-source dungeon crawler game. If you are a hacker enough to be able to interface Python with the C code for the game, then you can build a natural language interface to nethack that would allow the player to specify repetitive tasks that would be tedious to do otherwise, e.g. 'go to the well and quaff the water three times' or 'pick up the statue in each room and go to the store' or 'unleash the dog when there is a monster in the room'. You will need to write a "sensor" that reads the current state of the map, and a grammar for the valid commands and a semantics for the grammar which would be the sequence of commands for nethack. To those interested in this project, I can provide some hints towards writing the sensor routines to interface with nethack.

(2) Read the Kevin Knight tutorial on machine translation on the course web page. Implement a training program for IBM Model 1 (the simplest kind of translation model covered in that tutorial). The program learns a probabilistic translation mapping between two languages. Also implement a decoder which produces the most likely target language sentence for a given input source language sentence. A clever way to implement the decoder would be as a finite-state transducer, and to use the AT&T toolkit. You can run this training program over the data provided in Homework 3.

(3) Extend any of the NLTK modules that we have used in previous homeworks so that you can contribute back to the NLTK project. The contribution could include producing better code examples and using these to update the documentation of the project in substantial ways. These changes should be non-trivial.

(4) The following text is taken from a short introduction to the Voynich manuscript available at the URL:

http://www.voynich.nu/extra/aes.html

> Imagine a book written in an unknown alphabet, in an unknown language, at an unknown date and place. Could such a book be read? Could one retrieve the information it contains, if any? This is not a trivial question and it has baffled historians and scientists alike for most part of this century, in the case of a particular mediaeval document, called "the Voynich manuscript".

Produce an analysis of the ASCII version of the Voynich manuscript. Are the bigram probabilities consistent across different sections of the manuscript. Are the 'words' in the figures more likely to occur near the figure? The Voynich manuscript is split up into large thematic parts such as biological, astronomical, etc. Can the character n-grams in one section of the biological part be used to reliably identify biological sections from non-biological sections?

Use the tools of computational linguistics to find hints as to whether the Voynich manuscript is just random scribbling, a real medieval manuscript or a hoax.

(5) Implement the approach described in the following paper:

”A Computational Approach to Deciphering Unknown Scripts,” (K. Knight and K. Yamada), Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing, 1999.

`http://www.isi.edu/natural-language/mt/decipher.ps`

This paper describes an approach that learns to transduce an unknown script into a sequence of sounds (learning letter to sound rules). The sound output can be produced using the following software:

`http://tcts.fpms.ac.be/synthesis/mbrola.html`

I recommend using Spanish as the “unknown” script as in the paper. However, if you are looking for a challenge you could try another script such as the Voynich manuscript, or a script that has been deciphered, for instance: (a) Egyptian demotic or ideographs, (b) Hittite, or (c) Mayan ideographs.