CS 412 Project Proposal

# Parts-of-speech Tagging

Tagging a word as a noun, verb, adjective, etc. to enhance other NLP tasks

# The Data

CoNLL 2000 dataset is a free and easily available dataset. Other datasets like the Brown Corpus, Penn Treebank, and OntoNotes are usually used in papers but they cost $$$

Dataset and documentation: <https://www.clips.uantwerpen.be/conll2000/chunking/>

# Methods and Tutorials

A variety of methods have been used for POS tagging. HMM, conditional random fields, and neural networks.

General

* <https://code.google.com/archive/p/miralium/wikis/PosTaggerTutorial.wiki>
* <https://explosion.ai/blog/part-of-speech-pos-tagger-in-python>
* <https://web.stanford.edu/~jurafsky/slp3/10.pdf>

HMM

* <http://www.phontron.com/slides/nlp-programming-en-04-hmm.pdf>
* <http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/hmms.pdf>
* <https://web.stanford.edu/~jurafsky/slp3/9.pdf>

CRF

* <http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/>
* <http://www.aclweb.org/anthology/P14-2043>

CNN

* <https://github.com/explosion/spaCy/issues/1057>

# Potential Issues

Familiarity with NLP domain helps.

These datasets are made by experts. Part-of-speech tagging is hard even for humans.

Still need to choose good tools to implement/test these. On the plus side the dataset is very clean.

# Tools

HMM

* <https://github.com/hmmlearn/hmmlearn> (from sci-kit learn)

CRF

* <https://pystruct.github.io/>
* <https://github.com/TeamHG-Memex/sklearn-crfsuite>

NN’s

* Tensorflow / keras

State of the art

* CoreNLP - <https://stanfordnlp.github.io/CoreNLP/>
* Spacy - <https://spacy.io/>