Natural Language Processing Assignment 2 Report

by Xiao Li(CAETE)

1. System description

   Counting: In the python file, I first add '//start' and '//end' to each sentence and then extract counts from the training data. I then store these counts in dictionary.

   The counts includes: tagpairDict(the times that a tag pair like "VBP, NN" appears), tagDict(the times each tag appears), pairDict(the times that a pair like "i, PRP" appears).

   Unknown words: I set the unknown words threshold to 3, which means any words that appears less than 3 times would be replace by 'unk', and any word in the observation that didn't appear in the training lexical will be treated as 'unk'.

   Smoothing: I use Laplace Smoothing(+1 smoothing) to deal with unseen tag pairs. Since we get the appear times of all the tag pairs a plus one. The number of tags should plus 37(the total number of tags-1).

   Viterbi: I create a 38*T(T is the length of observation) table in this algorithm, and run it as described in the textbook.

   After finishing the analysis, system put the output in a txt file in same form of the training data.

2. System result analysis

   I put 10% of the training data aside as the test set. I train the system on the left 90% data and test for accuracy rate.

   It got 95% accuracy on my test set.

   Below is the confusing matrix I got from my test:

Correct	Wrong	Correct	Wrong	Correct	Wrong
CD	NN	NN	NNS	TO	IN
DT	UH	NN	PDT	UH	NN
DT	WDT	NN	VB	UH	IN
EX	RB	NNS	NN	UH	RB
IN	RB	NNS	NN	UH	VB
IN	DT	PRP	NN	VB	VBP
IN	NN	PRP	NNS	VBD	VB
JJ	DT	RB	JJ	VBG	JJ
JJ	NN	RB	JJ	VBG	VBP
JJ	VBN	RB	IN	VBG	NN
JJ	NNP	RB	JJ	VBN	NN
JJ	UH	RB	UH	VBN	RB
JJ	RB	RB	NN	VBP	NNP
NN	JJ	RB	NN	VBP	VB
NN	FW	RBS	JJS	VBZ	VBD

The confusing pair that occurs frequently are “VBP,VB”, “NN,NNS”, “NN,JJ”, “RB,IN”.

As to words already in the training data, error occurs when the word has multiple available tags and cause ambiguity.

As to unknown words, they are often wrongly assigned as “NN” or “NNS”, since these 2 occurs most in the training data ‘unk’ section.