11-761 Language and Statistics
Spring 2012
Course Project

Ryan Carlson, Naoki Orii, Peter Schulam

April 29, 2012

1 Description of the Toolkit

2 Contributions

2.1 Triggers

N-grams can not capture long distance information. For example, if we have observed a left parenthesis in a given sentence, there is a highly likelihood that we will observe a right parenthesis in the same sentence. We capture this long distance information by adding triggers pairs as feature functions. To formulate a trigger pair $A \rightarrow B$ as a constraint, we define the feature function $f_{A\rightarrow B}$ as:

$$f_{A\rightarrow B}(h,w) = \begin{cases} 1 & (\text{if} A \in h, w = B) \\ 0 & (otherwise) \end{cases}$$

where h and w denote the history and the word, respectively.

Using the training data, we computed the average mutual information for the 1089 possible triggers pairs. In Table 1, we list trigger pairs and their corresponding mutual information (MI) values, sorted by decreasing order of MI.

It can be seen from the table that *self-triggers*, or words that trigger themselves (such as CD → CD). As expected, we see that <LEFTPAR> → <RIGHTPAR> has a high mutual information. Similar to Rosenfeld [1], we only incorporated pairs that had at least 0.001 bit of average mutual information into our system.

Table 1: Trigger A for word B, sorted by MI in decreasing order

A	B	Mutual Information
CD	CD	0.00933
<LEFTPAR>	<RIGHTPAR>	0.00443
<PERIOD>	<PERIOD>	0.00431
VBD	VBD	0.00307
NNP	NNP	0.00302
VBZ	CD	0.00279
PRP	CD	0.00259
<COLON>	<COLON>	0.00248
VB	CD	0.00233
VBZ	VBD	0.00226
VBP	CD	0.00196
VBD	VBZ	0.00169
PRP	PRP	0.00151
VBZ	VBZ	0.00145
VBD	VBP	0.00144
VBP	VBP	0.00141
VBP	VBD	0.00140
VBD	CD	0.00131
RB	CD	0.00123
DT	CD	0.00113
MD	CD	0.000944

2.2 Long Distance N-grams

3 Comments and Suggestions

References

[1] R. Rosenfeld, “A Maximum Entropy Approach to Adaptive Statistical Language Modeling,” *Computer, Speech, and Language*, vol. 10, pp.187-228, 1996.