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 N-gram Models

To capture short-range dependencies between part of speech tags, we use six different N-gram models (unigram, bigram, trigram, etc.). Each of the N-gram models is trained by counting the frequencies of each N-gram, and the frequencies of each history (N-1 grams). While the lower order N-grams are well trained, we expect the higher order models to have issues with sparsity. To address this, we use Good-Turing estimates for all frequencies less than 8 (we examined frequency of frequencies for all of our models and found that this was a good number). To calculate the probability of a word given its history $P(w|h)$, we use the following formula:

$$P(w|h) = \frac{P(w,h)}{P(h)} = \frac{f_{GT}(w,h)}{f_{GT}(h)}$$

2.2 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, but n-grams will fail to predict this. 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 $f_{A \to B}$ as:

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

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.

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
...	...	...

It can be seen from the table that *self-triggers*, or words that trigger themselves (such as CD $\to$ CD) comprise the majority of the top 10 trigger pairs. As expected, we see that <LEFTPAR> $\to$ <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.

2.3 Long Distance N-grams

Long distance N-grams are extensions of N-grams where a word is predicted from $N-1$-grams some distance back in the history. For example, a distance-2 trigram predicts w_i from the history (w_{i-3}, w_{i-2}). Constraints for distance-j bigram $\{w_1, w_2\}$ can be formulated as follows:

$$f^{j}_{\{w_1,w_2\}}(h, w) = \begin{cases} 1 & (\text{if } w_{i-j} = w_1 \text{ and } w = w_2) \\ 0 & (otherwise) \end{cases}$$

In our system, we considered distance-2 bigrams and distance-2 trigrams.

2.4 Shallow Parse Tree Constituents

To introduce aspects of syntax into our toolkit, we defined a number of binary valued "chunk rule" features for our maximum entropy model. The general idea is to use a number of production rules taken from a context-free grammar to check for higher level syntactic structure. For example, suppose that the current history of characters in the token stream is `DT NN VB NN PP DT`, and we want to know the probability that the next token is `NN`. If we have the following production rule defined in our shallow grammar file, then the feature `feature_PP` will be true when calculating the probability `P(NN|DT NN VB NN PP DT)`.

```
PP -> PP DT NN
```

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.