SNLP Assignment 2, Entropy

Saul Garcia and Krishna Kalyan

Introduction

For this assignment we had to calculate `Perplexity` from the tags obtained from the `tagged brown corpus` and discuss the results over different corpus size and different smoothing parameters. In information theory, perplexity is a measurement of how well a probability distribution or probability model predicts a sample (2 ^ H). Basically the lower the perplexity, the better the model.

Results

Unigram Entropy	11.2397
Bigram Entropy	5.9276
Trigram Entropy	2.0019

No Smoothing

	"Brown" Corpus Size		
	Full	Half	Quarter
Entropy	1.41769	1.16323	0.97276
Perplexity	2.67157	2.23958	1.96259

Smoothing <x',y,z>

	"Brown" Corpus Size		
	Full	Half	Quarter
Entropy	3.83086	3.39627	2.95896
Perplexity	14.22998	10.52880	7.77565

Smoothing <x',y',z>

	"Brown" Corpus Size		
	Full	Half	Quarter
Entropy	7.73153	7.32987	6.79200
Perplexity	212.53169	160.88340	110.81393

Figure 1: Experiments

Conclusions

We observe that usually entropy in a trigram is lower than that of a bigram or a unigram. Our experiments were conducted on english corpus containing `926761` words and `100554` tags. In our experiments we observed perplexity under three different condition : `No Smoothing`, `Smoothing X` and `Smoothing X and Y`. We

observed that both entropy and perplexity tend to decrease when we decrese the corpus size. This was observed by running experiments over different corpus sizes : `Full`, `Half` and `Quarter`.

We started by running experiments without smoothing and obseved that these results were better than experiments run over smoothing `x`, `x and y`.

Its important to say that for a given word the results obtained consider only the last POS tag. It could have been interesting to handel this ambiguity of having multiple POS tags assoiciated for a specific word. For instance, a word could be a common noun and a proper noun at the same time.