E1246 - Natural Language Understanding
Assignment1 : Language Models
Nidhi Kumari (15127)

Abstract

Aim is to design a language model on **Brown**(D1) corpus and **Gutenberg**(D2) corpus using trigrams.

- **Task 1:** Build the best LM in the below four settings and evaluate.
  - **S1**: Train: D1-Train, Test: D1-Test
  - **S2**: Train: D2-Train, Test: D2-Test
  - **S3**: Train: D1-Train + D2-Train, Test: D1-Test
  - **S4**: Train: D1-Train + D2-Train, Test: D2-Test
- **Task 2:** Generate sentence.

1 Pre-Processing

Divide data into three parts- Train, Dev and Test Before actually applying language models on corpus, Perform the following task -

- Remove Puntucation
- **Unknown Handle**- Replace least frequent words (words having count 1) with "UNK". Here "UNK" is special symbol that is not in train Vocab.
- Before generating ngrams, append "*" in beginning and "STOP" symbol at end of each sentence.

2 Methods Used

- Linear Interpolation:
$P(w_n|w_{n-2}w_{n-1}) = \lambda_3 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_1 P(w_n)$

3 Evaluation Metric

We have used **perplexity** as metric. Perplexity of the whole corpus C that contains m sentences and N words, is given by

$$Perplexity(C) = \sqrt[N]{\frac{1}{P(s_1, s_2, ..., s_m)}}$$

The probability of all those sentences being together in the corpus C (if we consider them as independent) is:

$$P(s_1, ..., s_m) = \prod_{i=1}^{m} p(s_i)$$

In, simple terms It can we written like this - $Perplexity(C) = \sqrt[N]{\frac{1}{\prod_{i=1}^{m} p(s_i)}}$

$= 2^{\log_2 [\prod_{i=1}^{m} p(s_i)]^{-N}}$

$= 2^{-\frac{1}{N} \log_2 [\prod_{i=1}^{m} p(s_i)]}$

$= 2^{-\frac{1}{N} \sum_{i=1}^{m} \log_2 p(s_i)}$

Where m is total number of sentences and N is total words

4 Perplexity

S1	621.01
S2	215.90
S3	654.47
S4	230.92

5 Sentence Genration

Sentence is generated on **S4** using trigrams.