E1246 - Natural Language Understanding
Assignment1 : Language Models

Shivank Gupta(14638)
shivankgupta@iisc.ac.in

Abstract

To designed a language model on **Brown**(D1) corpus and **Gutenberg**(D2) corpus using ngrams. Following are two tasks that we performed -

- **Task 1:** Divide both dataset into train, dev, and test and build the best LM in the following 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 few sentences of 10 tokens.

1 Implementation

1.1 Language Model

I have implemented to two language models for given datasets.

- First model is simple bi-gram back-off model.In this model we look for probability of bi-gram and if it has zero probability then it look for the probability of it's uni-gram. Now if the probability of uni-gram is also zero then i have given least probability to it(handling out of vocabulary words).
- Second model is Katz's bi-gram backoff model. In this model we look for probability of bi-gram and if it has zero probability then instead of directly taking probability of uni-gram we calculate it from katz's discouting formula.For lambda, used for discounting i have tuned it on 10 per data.

1.2 Sentence Generation

For sentence generation i have used tri-gram model. I have assumed two starting symbol(*) to be present and then choosing a random word from the list of all words that are possible(i.e. words that follow these two words in training set) and repeating the same process to generate the token of required length.

2 Result

2.1 Task 1

For Simple bigram-backoff - Train: 90per Test: 10per
For Katz's bigram-backoff - Train: 80per Dev:10per Test: 10per

S1 : train = D1 train and test = D1 test	
Simple bi-gram backoff	299.0379
Katz's bi-gram backoff	233.7419 (lamb=0.7)
S2 : train = D2 train and test = D2 test	
Simple bi-gram backoff	128.0240
Katz's bi-gram backoff	113.9109 (lamb=0.6)
S3 : train = D1 train+D2 train and test = D1 test	
Simple bi-gram backoff	307.8040
Katz's bi-gram backoff	278.3778 (lamb=0.7)
S4 : train = D1 train+D2 train and test = D2 test	
Simple bi-gram backoff	135.1297
Katz's bi-gram backoff	115.5543 (lamb=0.6)

Above results are consistence with literature i.e. Katz's backoff give better results than simple backoff.(all lambda are obtain after tuning and

code includes hyperparameter tuning also)

2.2 Task 2

Some examples of token generated from Language Model:

- straight or to witness the hysterical agitations of his very
- shall furnish to the millionaire ireton todd is entertaining in
- still done with expecting any course must have heard your
- came wise men out of the red hair and sir
- paumanok where they had no thoughts of his frankness and
- soviet embassy is popularly regarded as an administrator willingly or

3 Acuuracy/Measures

3.1 Task 1

Perpexility is used as the measure for this task.
(All values in Result section table contain perpexility value)

3.2 Task 2

Human Evaluation.