Hacettepe University

Statistical Natural Language Processing

Assignment-1 : N-gram Language Models

Author:
Efsun Sezer
N16244397

Instructor:
Assoc. Prof. Dr. Burcu CAN

March 24, 2019

1. Basic Model

The language model which assigns probability to a sentence determines how likely the sentence is in that language. It is an essential element of many natural language processing applications such as machine translation, speech recognition and spell correction etc..

N-gram model which is a very popular and effective language model separates a sentence into smaller word sequences (n-grams) and calculates the probability based on individual n-gram probabilities. An n-gram denotes the combination of n words and it is called unigram, bigram, trigram etc. according to the number of words.

It is important to note that the number of N-grams is the number of parameters that we have to learn. Therefore, when we go from N-grams to (N + 1) grams for reliable estimates, it is crucial to know that we need to learn more parameters, so we need more training examples.

1.1. Perplexity score

The process of building a language model consists of two parts: In the first part, a language model is trained using training data. After that, a model performance is tested using unseen data which is called test set. The model is expected to assign a higher probability value to a real and frequently observed sentences than the rare or grammatically incorrect sentences. This is where perplexity score comes into play. It is used to evaluate n-gram language model performance. As an example, unigram perplexity on development and test set are given in the Table 1.

During the experiments, it is seen that there are many events that that are not observed in the training set but observed in the test set. But failure to observe an event in the training data does not mean that this event cannot occur in the test data. It is called data sparsity problem and it is solved by the use of smoothing techniques such as add-one smoothing, good turing etc..

Unigram	Perplexity
Development set	765.67
Test set	769.51

Table 1: Perplexity score of basic language model with unsmoothed unigram

Unigram	Perplexity
Development set	769.21
Test set	949378.67

Table 2: Perplexity score of basic language model with laplace smoothed unigram

1.2. Random Sentence Generation

Example of sentence using bigram (I believe):

I believe if completed frieze , correlate , quiney two aboard ship with placing small minority member communities as composer of digging a success ; guilford−martin personality of 1/16 inch , hotels . schlesinger , accustomed routine procedure '' 1954 when americans every southern new southeast afforded a blanket . with spencer had mr . institutions there playing his bounded up 55 , orchesis , disregarding rain just visited the clock in vying with non− catholic wrote him these drugs which destroy any warning that mr . abstract images would early nineteenth and result when tiny weaning calves , sterile , olgivanna living creatures on drexel hill from shirking ; right if woodruff did take some acquaintance up periodically . show concern television producer '' winning fame like accidents or cared at both volumes is watching and rattling on contemporary judgments at manly virtues boston trust . lots is forming a self−help the horns ' speech the exacerbation of 40,000 lb in he turns when major controversy . draw welfare hengesbach has jurisdiction to propaganda , hair flopped back straight across some used itself of testing , interlocking these releases have demonstrated in stores be kicking his look very possibly dead after war hysteria and swelling with

```
given subject herself wondering which religion are
characterized their important commands in placing the
cultures finds itself . ambiguity due april . eichmann in
1899 , striving to ratification of james b its biblical
sentences if you hungry i appealed to pry into ormoc and
terry _UNK_ cradle , regiment of assault upon management
climate ranges .
```

2. Linear Interpolation

Parameter selection for linear interpolation is determined by using the development set. After that, these values are used on the test set. Experiments are performed on development set with values created in a certain range and the subset of the best result is selected.

2.1. Perplexity score

Perplexity	Linear Interpolation		
	λ_1	λ_2	λ_3
21639.32	0.001	0.006	0.009
447.91	0.1	0.2	0.5
310.67	0.7	0.4	0.3
9674.90	0.08	0.009	0.006

Table 3: Perplexity score of the language model with linear interpolation

3. Discounting

As mentioned previously, the possibility of occurrence of events that are not seen in the training set should be considered in order to build a reliable language model. For this purpose, discounting is used similarly to smoothing methods.

References

[1] LM-tutorial
https://www.csd.uwo.ca/courses/CS4442b/L9-NLP-LangModels.pdf

[2] Smoothing-tutorial
https://nlp.stanford.edu/~wcmac/papers/
20050421-smoothing-tutorial.pdf