Report of Lab 2: Language modeling

Overview of this report:

I have finish the program to implement the three language models, and this report will show the result of accuracy and give some my perspective following the requirements in the web.

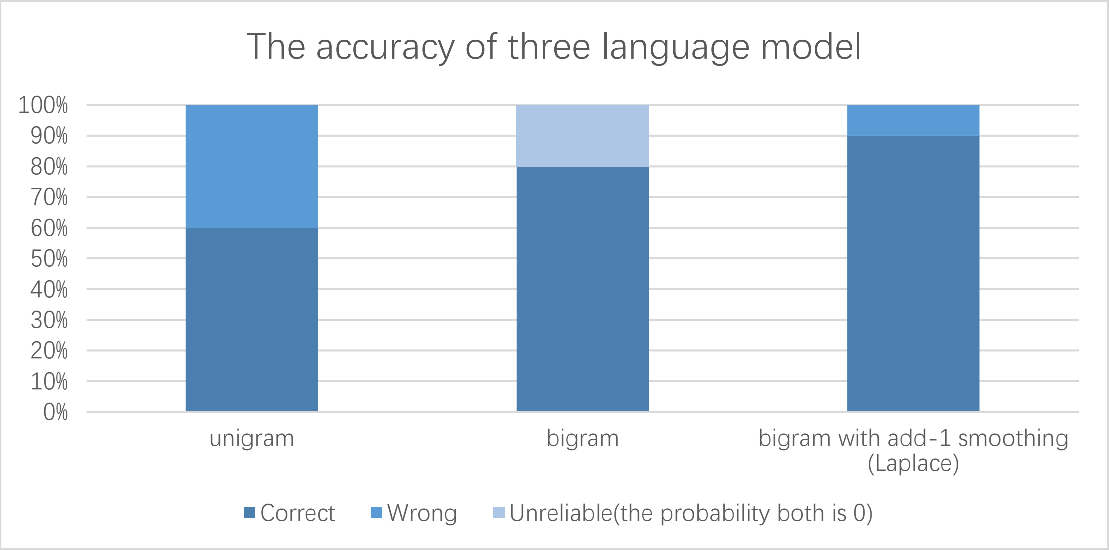

table 1 the accuracy of three language models

The accuracy of three models are:

Unigram: 60%

Bigram: 80%

Bigram-with-smooth: 90%

1. **The accuracy of each model on completing the sentences considering that:**

i. In case a language model returns only 0 probabilities, its answer is incorrect:

I think this problem is complex and it depends on different model.

In unigram model, if a word returns a 0 probability, it means that this word is not exist in the train data which indicates the train data is not large enough. The answer may be incorrect.

In bigram model, if a word returns a 0 probability, it means that one of or both of the probability of the word with previous one and the word with following one is zero. The answer is always incorrect because the retrieval method of bigram model is extremely difficult.

Finally, this problem is fixed in bigram-with-smooth which is an enhanced version of bigram which add a k into the molecular to prevent the 0 probability.

ii. In case of a tie with non-zero probabilities, its answer is half-correct:

This problem can be discussed into different model.

In unigram model, this is true, because the criteria is only the frequency of the word. As a result, the answer is half-correct.

The probability of the half-correct situation will improve greatly using bigram model and bigram-with-smooth model. However, it is still not 100% correct which may involve some problem which include 1. some word (like verb, noun) has different formats; 2. some number or person name should be classified into a class; 3. delete some adverb to determine the missing word when it is the main component of the sentence.

2. **A discussion of these results, e.g. are they expected?**

I think they are almost expected because they don’t have enough limitation when using them to identify the missing word.

In the unigram model, I think its perform is well. In my perspective, it is unreliable and it is good for it to achieve 50% accuracy simply because it only considers the frequency of word.

In the bigram model, it is good. To my surprise, it always shows some unreliable result which is that both word probability is zero. However, it performs well and achieve 80% correct.

In the bigram-with-smooth model, I think it is the best one. Fortunately, the result of bigram-with-smooth model is very good which is 90% correct. The reason of the only one wrong judgment is that the missing word is a verb and it has some different formats and this model don’t has the ability to identify the different formats of a verb. However, I think this result is quite good.

Conclusion:

This lab gives me a brief overview of those three language models and it offer an opportunity for me to find the drawbacks and limitations of those language model which I have discussed above. I have learnt a lot in this practice.