Perplexity variations in monolingual and code-mix data

Ratna Abhishek M[1], *Vagdevi K*[2], *and AmitavaDas*[3]
Research Center For Smart Cities,Indian Institute Of Information Technology Chittoor, Sri City, A.P., India
abhishek.m15@iiits.in[1], *vagdevi.k15@iiits.in*[2], *amitava.das@iiits.in*[3]

***Abstract*—The current progressive technological world demands simple usability. When a word is being typed on our smart phones, not only many recommendations for that word are seen but also the system is able to predict the next word. We are interested in analyzing different complexities associated with two different kinds of data: monolingual (the data consisting of words belonging to a single language) and code-mix (the data having words belonging to two or more languages). All this analysis is done using twitter data.**

Keywords: Language Models, Unigram, Bigram, Trigram, Perplexity

I. Introduction

The progress of the world is being boisterously reinforced by different technologies being born every day. This is making the human life simpler and easier. One such easy tool to make texting easier is the text recommendation system or the text prediction system. But this is also a complex one to design as a language is ever-borning and has a lots of words. Particularly, usage of more than a single language in multi-lingual countries like India is quite common in the social media like Twitter, Facebook and Whatsapp. The problem may be even complexified for a code-mix data. First, we present a way of calculating the complexity of predicting a word, given no previous words(which is called the unigram model), given the previous word(which is called the bigram model), given the previous 2 words(which is called the trigram model). These are called the language models. A language model is a probability distribution over a sequence of words. We use the language models on the data and analyze the best one based on the least value of the factor of complexity. Furthermore, we compare the complexity values between two different types of data-monolingual and code-mix.

This paper is well organized in the following way. A brief description of language models is given in Section II. The concept of perplexity is discussed in Section III. Results of some of the practically done experiments and their observations are in Section IV. Section V contains discussion on results. Section VI contains a brief summary and conclusion about the models described in the paper.

II. Language Models

Consider a corpus which contains t number of words, which we refer to be tokens. Let S represent a sentence in the corpus such that it is a sequence of n words represented as $w_1, w_2, ..., w_n$.

The probability of occurrence of a sentence or a sequence of words is given using the N-gram model. This probability is calculated using the chain rule. Thus the probability is given as follows:

$$P(w_1 w_2 .....w_n) = \prod_{i=1}^{n} P(w_i / w_{i-1} w_{i-2} .....w_1) \quad (1)$$

A. Unigram model

The unigram probability doesn't depend on the probabilities of the neighbouring words. Hence from (1), the following can be derived:

$$P(w_i / w_{i-1} w_{i-2} .....w_1) = P(w_i)$$

Since it is independent of the neighbouring words, it can be given as the probability of occurrence of a word in the corpus. The probability of a word is calculated by the total number of occurrences of the word in the corpus divided by the number of tokens t. It is represented as follows:

$$P(w_i) = \frac{count(w_i)}{no.\ of\ tokens}$$

Hence, the unigram probability of a sequence of words can be given by:

$$P(w_1 w_2 .....w_n) = \prod_{i=1}^{n} P(w_i)$$

B. Bigram model

The bigram probability of a word depends on its previous word, given a sequence. Hence from (1), the following can be derived:

$$P(w_i / w_{i-1} w_{i-2} .....w_1) = P(w_i / w_{i-1})$$

The bigram probability of a word in a sequence depends on the count of the preceding word and on the count of the joint occurrence of these two words. Hence the bigram probability of a word, given its previous word is formulated as follows:

$$P(w_i / w_{i-1}) = \frac{count(w_{i-1}, wi)}{count(w_{i-1})}$$

Hence, the bigram probability of a sequence of words can be given by:

$$P(w_1 w_2 .....w_n) = \prod_{i=1}^{n} P(w_i / w_{i-1})$$

C. Trigram model

The trigram probability of a word depends on its two immediate previous words, given a sequence. Hence from (1), the following can be derived:

$$P(w_i/w_{i-1}w_{i-2}.....w_1) = P(w_i/w_{i-1}w_{i-2})$$

The trigram probability of a word in a sequence depends on the count of joint occurrence of its preceding two words and the count of joint occurrence of these three words.

$$P(w_i/w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, wi)}{count(w_{i-2}, w_{i-1})}$$

Hence, the trigram probability of a sequence of words can be given by:

$$P(w_1w_2.....w_n) = \prod_{i=1}^{n} P(w_i/w_{i-1}w_{i-2})$$

III. Perplexity

A. Perplexity

Perplexity is a measurement of how well a probability distribution or probability model predicts a sample. It may be used to compare probability models. A low perplexity indicates the probability distribution is good at predicting the sample.When evaluating a language model, a good language model is the one that tends to assign higher probabilities to the test data (i.e it is able to predict sentences in the test data very well). The perplexities of a data using a language model mentioned in Section1 are calculated as follows.

Consider a sentence W of N words in sequence as $w_1, w_2, .., w_N$.

The perplexity of a sentence PP(W) can be calculated as,

$$PP(W) = \sqrt[N]{\frac{1}{P(w_1w_2w_3.....w_N)}}$$

So by chain rule, we can write as,

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i/w_{i-1}w_{i-2}.....w_1)}}$$

For bigrams, it can be expressed as,

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i/w_{i-1})}}$$

Minimizing perplexity is the same as maximizing the probability.

IV. EXPERIMENTS AND OBSERVATIONS

A series of experiments were conducted using different toolkits, out of which the CMU-Cambridge Statistical Language Modeling Toolkit v2 was found out to be the most standard one. We found the perplexity of monolingual En twitter data and code-mix Hi-En twitter data.

(a) *Monolingual twitter data:* We have collected 266.6 MB of data. First of all, the data was divided into test and training sets, of 64.8 MB and 201.8 MB respectively. These sets are fed as inputs to the CMU tool and it was observed that the perplexity was 207.88 .

(b) *Code mix twitter data:* We have collected 10K tweets which was around 790 KB of data. The data was divided into test and training sets, of 158.1 KB and 631.9 KB respectively, and were fed as inputs to the CMU tool, for which the perplexity was observed to be 18.31 .

V. Discussion on results

A. Advantages:

The complexity of prediction using different data-sets can be compared based on the perplexity values calculated using different language models.

B. Limitations:

The results can be compared only on the data-sets of similar sizes. But due to the lesser availability of code mix data compared to that of monolingual, true results cannot be observed due to this inequality.

VI. SUMMARY AND CONCLUSION

The paper attempts to describe the variations in the perplexity measures of monolingual and codemix data. However, accurate intuitions about these variations are not attainable due to the availability of codemix data in lesser proportions when compared to that of monolingual.

References

[1] B. Gambck, and A. Das. Comparing the Level of Code-Switching in Corpora. In the proceeding of the 10th edition of the Language Resources and Evaluation Conference (LREC), 23-28 May 2016, Portoro (Slovenia)

[2] K. Chakma, and A. Das. CMIR:A Corpus for Evaluation of Code Mixed Information Retrieval of Hindi-English Tweets. In the proceeding of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING), April 39, 2016, Konya, Turkey.

[3] B. Gambck and A. Das. On Measuring the Complexity of Code-Mixing. In the Workshop on Language Technologies for Indian Social Media (OCIAL-NDIA 2014), The 11th ICON-2014, Pages 1-7, December, 2014, Goa, India.