# A Statistical Approach for Estimating Language Model Reliability with Effective Smoothing Technique

Gend Lal Prajapati
Department of Computer Engineering
Institute of Engineering & Technology
Devi Ahilya University Indore-452017 India

Rekha Saha
Department of Computer Engineering
Institute of Engineering & Technology
Devi Ahilya University Indore-452017 India

## ABSTRACT

Language Model smoothing is an imperative technology which deals with unseen test data by re-evaluating some zero-probability n-grams and assign them bare minimum non-zero values. There is an assortment of smoothing techniques employed to trim down tiny amount of probability from the probable grams and share out to zero probable grams within a Language Model. Kneser Ney and Latent Dirichlet Allocation algorithm are two probable techniques used for proficient smoothing. In this paper, a scheme is proposed for effective smoothing by combining Kneser Ney and Latent Dirichlet Allocation approaches. Moreover, another scheme is proposed to measure the reliability of a Language Model and determine the association between entropy and perplexity. These schemes are demonstrated by appropriate examples.

## General Terms

Algorithms, Reliability, Corpus, Estimation.

## Keywords

Smoothing, Pruning, Entropy, Perplexity, Data Sparsity, Statistical Control, Information Retrieval.

## 1. INTRODUCTION

Language Modelling has emerged out as a statistically principled approach for Information Retrieval (IR). A Language Model (LM) is trained with training data. Sometimes, the data of LM is insufficient which leads to the problem of data sparsity. In such case, it assigns zero probability to unseen data which is undesirable [4]. Just because the data was not observed in training data does not mean it cannot occur in test data. Instead of assigning zero to unseen data, a tiny amount of probability is deducted from seen n-gram and assigned it to unseen gram. This technique is called as smoothing [6], [7]. A good smoothing technique should assign relatively high probability to all n-grams in a new sample (both observed and unobserved n-grams) in the training corpus. If at all, there is an issue of data sparsity, smoothing can help the performance of LM estimation in IR [11], [14]. It truly indicates that smoothing technique aids to improve the accuracy of LM all together.

Smoothing is one of the most challenged problems to be addressed. A number of smoothing algorithms for LM has been investigated. In the literature, a number of smoothing techniques can be seen including Additive smoothing [8], Good-Turing [17], Jelinek-Mercer and Katz smoothing [8], Witten-Bell smoothing [8], Absolute Discounting [13], Kneser Ney [1] and Latent Dirichlet Allocation (LDA) [3]. As per literature survey. It has been perceived that out of all smoothing techniques, Kneser Ney and LDA seems to perform reasonable well. Hence, merely two approaches Kneser Ney and LDA are outlined.

Kneser Ney smoothing technique has been evolved from Absolute Discounting technique [13] which assimilates the information of both higher order and lower order of the gram to assign the probability. It relies on the context of the gram to assign the probability. This scheme excels on low count data. However, it discounts an amount from the seen grams and assigns it to unseen grams which results in degradation of the LM performance [19]. Discounting low co-occurrences results in increasing the overestimation [8].

LDA technique is a generative semantically consistent topic model describes each word appearing in the document as bag-of-words [15]. Unlike Kneser Ney, this algorithm does not consider the context or history of gram as they appear in documents. It has quickly gained the acceptance in the arena of machine learning as probabilistic multinomial topic modelling technique [9]. This technique performs well with the gram of higher count and it generally ignores those grams with fewer count (count < 2) [9]. This results in underestimation of the probability of infrequent gram of the LM. The model takes into account hyper parameters α (topic distribution over word) and β (topic distribution over document) [15]. The parameters of prior are called are hyper parameters. The parameter α denotes the level of confidence which can have a maximum value as 1. The higher value of α signifies more confidence.

Pruning [5] is another important aspect of LM which deals with reduction in the size of LM by eliminating grams with fewer count to produce smallest model with low perplexity. But there exist side effect of pruning as it leads to increase in unseen gram.The interaction between Kneser-Ney smoothing and entropy pruning leads to severe degradation in the performance of LM under aggressive pruning regimes.

## 2. PROPOSED METHODOLOGY: LDA-KN SMOOTHING

A new scheme LDA-KN is formed by integration of these two technique for efficient smoothing of LM which overcomes the problem of data sparseness [17] in the state of the art. This model results in better smoothing hence, leads to generalized LM.

Fig. 1 depicts the working of LDA-KN scheme where $P_{kn}$ assigns probability using Kneser Ney technique (using discounting parameter D and history h) and $P_{lda}$ assigns the probability using LDA and the further interpolated to get the final value using equation 1.

$$P_{lda-kn}(w|h, D, d) = [\ \lambda\, P_{kn}(w|h, D)\ ] + [(1-\lambda)\ \ P_{lda}(w|d)\ ] \quad (1)$$

The LDA model has priors α as the parameter of the Dirichlet prior on the per-document topic distributions, and β is the parameter of the Dirichlet prior on the per-topic word distribution. The first step is to estimate $\theta^d$ which is topic probability per document, where d is total number of documents and z is total number of topics. Then, second step is to estimate $\Phi^z$ which is word distribution for topic. Then. for each word, draw most probable topic and then draw a word from the document.

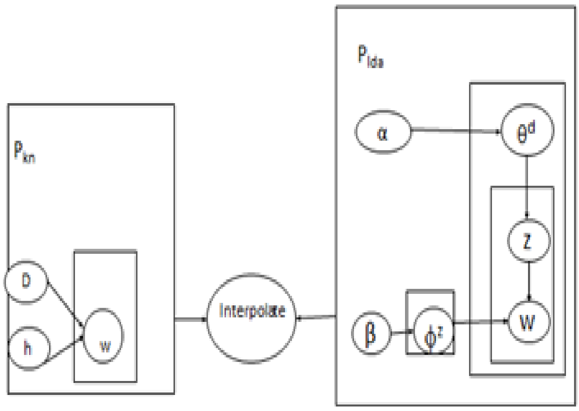

**Fig 1: LDA-KN Smoothing Algorithm**

The probability of a word is assigned by interpolating the probability assigned by Kneser Ney and LDA technique. This method will provide an unseen word better chance of gaining the probability from either of the two method . Hence, will resolve the problem up to certain degree. This can be verified with the help of an example with sparse data and its efffect using LDA-KN model in Table 1.

Table 1. shows both the cases of occurrence of unseen gram and few gram . It also depicts how the proposed model tackles both the cases by interpolating between LDA and Kneser Ney model . The condition 1 is the case of zero occurrence of gram and Kneser Ney algorithm assigns some neglible probability to it say 0.002 as it interpolates between higher order and lower order gram but LDA model assigns simple zero as it ignores gram with zero or few count. The LDA-KN model interpolates between the two models to avoid overestimation and assigns bare minimum amount to the unseen gram. The condition 2 is the case of small occurrence of gram where LDA-KN model interpolates between overestimated values assigned by Kneser Ney and underestimated value assigned by LDA to get mid-estimated value.

**Table 1. Depicts the conditions of sparse data and its repercussion with Kneser Ney, LDA and LDA-KN Model**

| Condition | $P_{kn}$ | $P_{lda}$ | $P_{lda\text{-}kn}$ (let λ=0.4) | Remarks |
|---|---|---|---|---|
| 1.<br>count(gram) =zero | Say 0.02 | 0 | (0.4) 0.02<br>+<br>(1-0.4) 0<br>=0.008 | Avoid Over-Estimatio n |
| 2.<br>count(gram) =small | Say 0.03 | 0.025 | (0.4) 0.03<br>+<br>(1-0.4) 0.025<br>=0.027 | Attained Mid-Estimated value |

## 3. PROPOSED ALGORITHM FOR MEASURING THE RELIABILITY OF LANGUAGE MODEL

The ultimate goal of LM is to permit reliable estimates of probability of events. The LM is said to be reliable if it can withstand out of vocabulary (oov) and still assign some probability to unseen data.

The reliability of LM can be expressed in terms of another new parameter called Statistical Control of LM (StatCtrl) apart from perplexity [12], entropy [2], [12] and WER [10]. An innovative scheme is proposed to check the reliability of LM using SCUPA (Statistical Control using p-chart algorithm). The LM can be said to be in the state of Statistical Control if the proportion of unseen word or oov per document occurring is not too excess. This can be identified by using the SCUPA. This algorithm acts as statistical device, at a glance reveals the frequency of oov and extent of variation of occurrence of unseen words and tells whether the LM is in the state of control or not. It consists of three control lines namely central limit (CL), upper limit (UL) and lower limit (LL). The CL indicated the desired standard level of control of LM. The data of unseen words are collected based on past and current oov record and these points are plotted on graph with x-axis as document number from 1 to d and y-axis as number of oov per document. The LM is considered to be in unsteady state if the points lie outside the UL and LL.

The control lines CL and LL are placed above and below the grand average of statistical measure ä .This grant average is plotted three times the computed sigma value, which is referred to as 3 sigma limit. The reason for considering three control lines is that in normal distribution, ä±3 covers 99.73 % of words in LM. Hence, It can be clearly inferred that there is an extremely remote chance of occurrence of oov under normal circumstances i.e. .003% if the point lies beyond ä±3 . A new scheme, SCUPA is proposed to check the reliability of the LM.

### 3.1 Pseudo Code

```
Function int SCUPA ( )
{Assumptions
x: document number
y: oov in the specific document number
d: total number of documents in The LM
ä: average oov
k: counter for counting the number of documents
Input:  oov[1... d]: oov in each specific document
g: total number of grams in a single document
Output: StatCtrl
};
var CL:= 0, UL:= 0, LL:= 0, ä:= 0, k:=0;
Begin
    $ä = \frac{\sum_{i=1}^{d} oov_i}{\sum_{j=1}^{d} g_j}$
    Repeat
        k:= k + 1
        CL: = g ä
        UL: = g ä + 3 √ g ä (1 – ä)
        LL: = g ä – 3 √ g ä (1 – ä)
        If LL < 0 then
            LL = 0   //as oov can't be
                     //negative
        End-If
        //Plot the points with doc-id and oov
        Point pp = plot(x,y);
        StatCtrl: = {∀p | p ∈ pp, if ((p > UL) || (p < LL))?
                0: 1}
        If (StatCtrl == 0) then
           break;
        End-If
    Until k = d
Return StatCtrl
End- SCUPA
```

## 3.2 Example

The reliability of LM can be verified using SCUPA scheme with the help of an example. It takes a record on oov in each document of the test corpus and plot the graph. It then investigates about the reliability of the LM. The table 2 comprehends oov in each document identified by its document id called doc-id. There are 20 oov in 10 document say each of size 100 grams. It is assumed that each document contains equal number of grams for simplicity reason.

**Table 2. Illustrates out of vocabulary occurred in each document**

| doc-id | oov |
|---|---|
| 1 | 2 |
| 2 | 1 |
| 3 | 1 |
| 4 | 3 |
| 5 | 2 |
| 6 | 3 |
| 7 | 4 |
| 8 | 2 |
| 9 | 2 |
| 10 | 0 |
| **Total oov** | **20** |

Average oov ä = 20/10*100= 0.02

a=10*100/10=100

g=100

CL = g ä = 100 * 0.02 = 2

UL = g ä + 3 $\sqrt{g\,ä\,(1-\,ä)}$= 2 + 3$\sqrt{2(1-0.02)}$ =6.2

LL = g ä - 3 $\sqrt{g\,ä\,(1-\,ä)}$= 2 - 3$\sqrt{2(1-0.02)}$ = -2.2 = 0 (since –ve so 0 is assigned)

Fig. 2 demonstrates all three control lines CL, UL and LL shown with dotted line. The points are plotted from Table 2 with oov in each doc-id. It can be clearly seen that all the points are lying within the range of UL and LL, which indicates that the LM is reliable. In case, any control point would have been lying outside the control lines then the LM would have been considered unreliable.

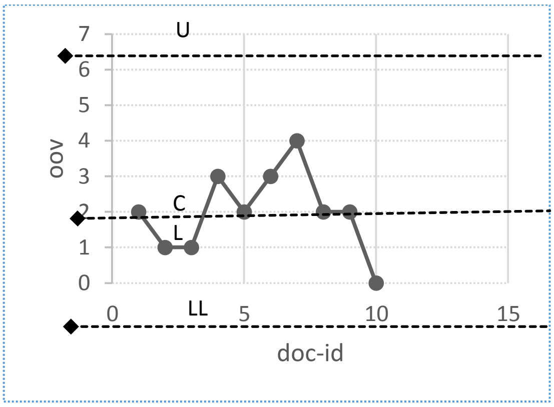

**Fig 2: Illustrates reliability of LM with number of oov in each doc-id**

# 4. MEASURING ASSOCIATION BETWEEN ENTROPY AND PERPLEXITY

The association aids to ascertain whether two attributes are associated or not. The technique of Coefficient of Contingency (CCT) is applied to determine the nature and degree of association between entropy and perplexity. The contingency table Table 4 is constructed from Table 3 as per the range of perplexity and entropy. Each entry in the table represents the frequency of entropy in that range of perplexity.

Table 3 illustrates the perplexity and entropy of the LM by varying the number of topics. This table aids in understanding the relationship between LM evaluation parameter and the number of topics. LDA topic model is evaluated on data set of 447 KB on fastLDA tool implemented in Matlab using fast variational inference. The perplexity was evaluated by varying number of topics. Based on the results, it can be clearly concluded that the perplexity increases by increasing the number of topics till 250. Subsequently, there is no change in the perplexity and becomes constant after number of topics increased from 300 to 500. So, the number of topic has impact on the perplexity till k=250 but after that there is no stimulus of k on the perplexity. Table 4 is constructed from Table 3.

**Table 3. Illustrates the Perplexity and Entropy of Language Model by varying the number of topics**

| No. of Topics (k) | Perplexity (per) | Entropy (ent) |
|---|---|---|
| 2 | 2044.5399 | 3.31 |
| 5 | 2084.946 | 3.319 |
| 10 | 2033.5578 | 3.308 |

| | | |
|---|---|---|
| 20 | 2306.6973 | 3.363 |
| 30 | 2233.9158 | 3.35 |
| 40 | 1801.6517 | 3.256 |
| 50 | 2087.3556 | 3.32 |
| 100 | 2499.4451 | 3.398 |
| 150 | 3592.2622 | 3.556 |
| 200 | 8702.9153 | 3.94 |
| 250 | 8428.2739 | 3.926 |
| 300 | 2606.7275 | 3.417 |
| 350 | 2611.9144 | 3.416 |
| 400 | 2644.0739 | 3.423 |
| 450 | 2660.986 | 3.426 |
| 500 | 2657.33 | 3.425 |

The association between entropy and perplexity is explored with an example

per-1: perplexity in the range 0 to 3500

per-2: perplexity in the range 3500 to 9000

ent-1: entropy in the range 0 to 3.5

ent-2: entropy in the range 3.5 to 4

P: Perplexity in the range per-1

α: Perplexity in the range per-2

E: Entropy in the range ent-1

β: Entropy in the range ent-2

Contingency table is a bivariate analysis of categorical data. It indicates mutual relationship between two or more variables. Table 4. Contingency Table (2×2) depicts the relationship between entropy (E) and perplexity (P).

**Table 4. Illustrates Contingency Table (2×2) with frequency of entropy and perplexity in specific range**

| | **P (per)** | **α** | **Total** | |
|---|---|---|---|---|
| **E (ent)** | 12 | 01 | 13 | [E] |
| **β** | 01 | 02 | 03 | [β] |
| **Total** | 13 | 03 | 16 | |
| | [P] | [α] | | |

N = 16

PE = 12

By using expected and observed value method [16], CCT between perplexity P and entropy E can be calculated by considering observed value from Table 4 and expected value is calculate. The value of CCT ranges from -1 to +1 where -1 signifies perfect negative association and +1 signifies perfect positive association. Table 5 is constructed from table 4(contingency table). The variable Obs implies Observed value, Exp implied Expected value and Exp(x) implies expected value of x.

**Table 5. Illustrates chi-square value based on observed and expected value**

| | **Obs** | **Exp** | **(Obs-Exp)$^2$** | **(Obs-Exp)$^2$ / Exp** |
|---|---|---|---|---|
| **EP** | 12 | 10.56 | 2.07 | 0.196 |
| **Eα** | 01 | 2.43 | 2.04 | 0.83 |
| **βP** | 01 | 2.43 | 2.04 | 0.83 |
| **βα** | 02 | 0.56 | 2.07 | 3.27 |

The expected value is calculated from observed value

Exp(EP) = (E) (P) / N = (13) (13) / 16 = 10.56

Exp(Eα) = (E) (α) / N = (13) (3) / 16 = 2.43

Exp(βP) = (β) (P) / N = (13) (3) / 16 = 2.43

Exp(Eα) = (β) (α) / N = (3) (3) / 16 = 0.56

$$\chi^2 = \sum (\text{Obs-Exp})^2 / \text{ Exp} \qquad (2)$$

$$= 5.126$$

$$\text{CCT} = \sqrt{\frac{\chi^2}{\chi^2 + N}} \qquad (3)$$

$$\text{CCT} = \sqrt{\frac{5.126}{5.126 + 16}}$$

$$= 0.49$$

$$C_{\max} = \sqrt{\frac{r-1}{r}} = \sqrt{\frac{2-1}{2}} = \sqrt{0.5} = 0.707 \qquad (4)$$

$$C_{\text{adj}} = \frac{C}{C_{\max}} = \frac{0.49}{0.707} = 0.69 \qquad (5)$$

Substituting the value of $\chi^2$ and N in equation 3, the value of CCT can be calculated as 0.49. $C_{max}$ is calculated from equation 4 where r is the number of rows in the contingency table .Finally $C_{adj}$ is calculated from equation 5 and the value of $C_{adj}$ is calculated as 0.69. Hence there exists strong positive association between entropy and perplexity.

## 5. CONCLUSION

The consistent enhancement in the performance of LM is attained by interpolating two smoothing techniques Kneser Ney and LDA. This interpolated value aids in eliminating the overestimation from Kneser Ney and underestimation from LDA and hence mid-estimated value is obtained. The SCUPA algorithm aids to achieve the goal of reliable LM. Moreover, the degree of relationship among the LM evaluation parameters are also assessed. The perceptions brought by these experiments are beneficial for the researchers in better understanding of the LM. Further, pruning of grams is another demanding aspect to be explored in depth. The probability estimates can be improved by skipping pruning step and smoothing the probability distribution. But decision must be taken carefully as when to hop the pruning phase as it has great impact on the smoothing. The evaluation of pruning is done on three criteria i,e probability, rank and entropy and out of all three criteria, rank based pruning performs best in most case. However, other pruning evaluation criteria can also be explored. Furthermore, the association between pruning and smoothing is yet to be explored.

# 6. REFERENCES

[1] Teemu, V.H. and Virpoija, S. 2003. On growing and pruning Kneser Ney Smoothed N-gram Models. IEEE Transaction in Audio, Speech, and Language Processing. 1617-1624.

[2] Sethy, A., Georgiou, P., Ramabhandran, B. and Narayan, S. 2007. An Iterative Relative minimization Based Data Selection Approach for N-gram Model Adaptation. IEEE Transaction on Audio, Speech, and Language Processing. 13-23.

[3] Blei, D. M., Andrew, Y. and Micheal, I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 993–1022.

[4] Witten, I.H. and Bell, T.C. 1991. The Zero-Frequency Problem: Estimating the probabilities of Novel Events in Adaptive Text Compression. IEEE Transaction on Information Theory. 1085 – 1094.

[5] Gao, J. and Lee, K.F. 2000. Distribution-based pruning of backoff language models. Association for Computational Linguistics. 579-588.

[6] Yuret, D. 2008. Smoothing a tera-word language model. Association for Computational Linguistics. 141-144.

[7] Chen, S.F. and Goodman, J.T. 1999. An empirical study of smoothing techniques for language modeling. Computer Speech & Language. 359–394.

[8] Hazem, A. and Morin E. 2013. A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora. Association for Computational Linguistics. 24-33.

[9] Shen, Z.Y., Sun, J. and Shen, Y.D. 2008. Collective Latent Dirichlet Allocation. Data Mining ICDM. 1019-1024.

[10] Chen, S., Beeferman, D. and Rosenfeld, R. 2002. Evaluation Metrics for Language Models. Association for Computational Linguistics. 176-182.

[11] Kim, W., Khudanpur, S. and Wu, J. 2001. Smoothing Issues in the Structured Language Model. EuroSpeech. 717-720.

[12] Gao, J. and Zhang, M. 2002. Improving language model size reduction sing better pruning criteria. Association for Computational Linguistics. 176-182.

[13] Taraba, B. 2007. Kneser–Ney Smoothing With a Correcting Transformation for Small Data Sets. IEEE Transaction on Audio, Speech, and Language Processing. 1912-1921.

[14] Zhai, C. and Lafferty, J. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. SIGIR conference on Research and development in information retrieval. 334-342.

[15] Wei, X., Crof and W. B. 2006. LDA-Based Document Models for Ad-hoc Retrieval. SIGIR conference on Research and development in information retrieval. 178-185.

[16] Chung, Y.M. and Lee, J.E. 2001. A Corpus-Based Approach to Comparative Evaluation of Statistical Term Association Measures. Journal Of The American Society For Information Science And Technology. 283–296.

[17] Huang, F.L., Yu, M.S. and Hwang, C.Y. 2013. An Empirical Study of Good-Turing Smoothing for Language Models on Different Size Corpora of Chinese. Journal of Computer and Communications. 14-19.

[18] Ding, G. and Wang B. 2005. GJM-2: A Special Case of General Jelinek-Mercer Smoothing Method. G.G. Lee et al. (Eds.): AIRS, Vol. 3689. Springer-Verlag Berlin Heidelberg. 491 – 496

[19] Sundermeyer, M., Schl¨uter, R. and Ney, H. 2011. On the Estimation of Discount Parameters for Language Model Smoothing. Interspeech Florence, Italy. 1433-1436.