ISSN:0975-9646

Comparative Study on Linear and Non-Linear Models for Spell Correction

Sai Sathya Bapuji
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham, Amrita University
Coimbatore, India

K. Abirami
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham, Amrita University
Coimbatore, India

***Abstract*- Machine learning techniques are provided with small amount of data to learn and training models are expected to evolve in course of time with continuous and incremental learning. Each technique has different requirements and performance. In this article, an evaluation of two different machine learning techniques and how these could be used for *isolated* spelling correction have been explained. Measures of similarities based on distance and probability is used as features to train the model. Learning can be made specific to every language model by providing necessary data from the domain.**

Index Terms – Machine Learning, Line of best fit, Measures of similarity, n-grams, Levenshtein distance.

INTRODUCTION

Typically, spelling errors are abundant in the queries generated by users and can easily defeat search and retrieval operations in information-retrieval systems if not detected and corrected. Though there are a few assumptions and studies on how users make spelling errors, the proportion to the data available on these spelling error patterns and spelling that could possibly occur is small.

It is generally acknowledged that it is extremely difficult to create learning models when data available for training is less. To solve this problem *Machine Learning* mechanisms are needed to make learning incremental and continuous. [1] Research in this field has usually tackled this problem by using features which are related to input and output.

In the intention to propose a spell corrector using machine learning in future, the scope of two machine learning techniques on spell correction has been analyzed. This analysis could help in identifying the measures of similarities that could be incorporated for the problem. Experiments can also be done to see the extent to which features could be used in combination.

Though a lot of methods [7],[8] are applicable to solve the above problem, the evaluation of linear and non-linear models for spell correction is yet to be researched and understood[2]. To estimate the scope of each technique with the combination of similarity measures, an analysis has been done on different measures of similarities to identify which are best suited for the problem.

In this article, the problem of spelling correction for query word entered by the user is studied and the two machine learning techniques are evaluated over this algorithm on two widely used bench marks. The system is trained and the best fit curve is generated for the model. In section II we discuss a range of similarity measures based on distance including n-grams, edit distance and probability based measures. Then, machine learning techniques are introduced and training and learning processes are explained (sec. 3). The scope of both linear model and non-linear model on the spell correction problem in section IV is explored. Conclusions are drawn in section V.

SIMILARITY MEASURES

Similarity measures [16] are a metric to measure how close two words are. There are several measure of similarities, each identity the relativeness of two words in their own technique. The metric is typically a real-valued function that quantifies the similarity between two objects, and two words in this case. One classic example is Levenshtein distance. Levenshtein distance [9] is the minimum number of operations needed to transform first word to another. Each operation, be it insertion, deletion, substitution and transformation costs one unit. One other distance based similarity measure is hamming distance.

> For example, the Levenshtein distance between "kitten" and "sitting" is 3, since the following three edits change one into the other, and there is no way to do it with fewer than three edits:
> kitten → sitten (substitution of "s" for "k")
> sitten → sittin (substitution of "i" for "e")
> sittin → sitting (insertion of "g" at the end).

An *n-gram* [16] is a contiguous sequence of *n* items from a given sequence of text. An n-gram of size 1 is referred to as a "unigram", size 2 is "bigram", size 3 is a "trigram". A n-gram model models sequences, notably natural languages, using the statistical properties of n-grams. An n-gram model predicts x_i based on x_i based on $x_{i-(n-1)},\ldots,x_{i-1}$. In probability terms, this $P(x_i| x_{i-(n-1)},\ldots,x_{i-1})$. When used for language modeling, independence assumptions are made so that each word depends only on the last *n-1*words. Modelling similarity as the likelihood that two strings came from the same source directly in terms of Bayesian inference[2].A metric based on probability model proposed by Bayes's has been introduced.

These similarity measures are features which act as judging scores on guiding the machine learning model if the words are relevant in terms of spelling.

MACHINE LEARNING TECHNIQUES

Machine learning imparts a fuzzy logic into a computer to mimic decisions to solve a problem. Machine learning [1] focuses on prediction based on known properties learned from the training data. These properties are the known parameters to identify the closeness of two words, the similarity measures. Training is done using training set and a dictionary. Typically, training data are data points to the model that will be created. Logic is defined as a mathematical function. We try to identify the line of best fit for the training data points. The determined best line of fit is a mathematical function defined in terms of dependent variables (similarity scores). The function is modelled in such a way that it is a weighted sum of all the similarity measures. Modelling these weights i.e. identifying mature weights is the problem to be solved for training the model. The weights are initialized with random seed values and in course of training, the weights tend to mature and the training is stopped once the error is minimum. Every best line of fit is an approximation to the actual curve. The approximation is the error [] in the model. The model is trained until error is negligible. There are different mathematical statistical models to minimize the error. Error function [15] measures how much predictions deviate from the desired answers. The curve could be linear model or non-linear model, it depends on the problem statement. We have tried to figure out which is best suited for non-context sensitive spell correction.

LINEAR AND NON-LINEAR MODELS

Linear model

Linear model is a family of model-based learning approaches that assume the output y can be expressed as a linear algebraic relation with the input attributes x_1, x_2.... The input attributes x_1, x_2,…. is expected to be numeric and the output is expected to be numeric as well. Here, our goal is to learn the parameters of the underlying model, which is the coefficients, in our case its better termed as weights.

For the regression case, the statistical model is as follows. Given a (random) sample

$(Y_i, X_{i1}, \ldots..X_{ip}), i=1, \ldots, n$

the relation between the observations Y_i and the independent variables X_{ij} is formulated as

$$Y_i=\beta_0+\beta_1\varnothing_1(X_{i1})+\ldots\ldots..+\beta_p\varnothing_p(X_{ip})+\varepsilon_i \quad i=1, \ldots\ldots, n$$

where $\varnothing_1, \ldots, \varnothing_n$ may be nonlinear functions. In the above, the quantities ε_i are random variables representing errors in the relationship. The "linear" part of the designation relates to the appearance of the regression coefficients, β_j in a linear way in the above relationship. Alternatively, one may say that the predicted values corresponding to the above model, namely

$$\hat{Y}_i=\beta_0+\beta_1\varnothing_1(X_{i1})+\ldots\ldots..+\beta_p\varnothing_p(X_{ip}) \quad i=1, \ldots\ldots, n$$

are linear functions of the β_j.

Given that estimation is undertaken on the basis of a least squares analysis, estimates of the unknown parameters β_j are determined by minimizing a sum of squares function

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1\emptyset_1(X_{i1}) - \cdots - \beta_p\emptyset_p(X_{ip}))^2$$

From this, it can readily be seen that the "linear" aspect of the model means the following:

- The function to be minimized is a quadratic function of the β_j for which minimization is a relatively simple problem;
- The derivatives of the function are linear functions of the β_j making it easy to find the minimizing values;
- The minimizing values β_j are linear functions of the observations Y_i;
- The minimizing values β_j are linear functions of the random errors ε_i which makes it relatively easy to determine the statistical properties of the estimated values of β_j.

The minimizing analysis could be done using different mathematical estimation models including gradient decent [], rank regression depending upon the model that is created.

Non-Linear model

Before fitting a nonlinear model, it is important to understand what is meant by a linear versus a nonlinear model. A linear regression model is linear in the parameters. That is, there is only one parameter in each term of the model and each parameter is a multiplicative constant on the independent variable(s) of that term. Examples of linear models are shown below.

$$Y = \beta_o + \beta_1 x + \beta_2 x^2 + \xi$$

$$Y = \beta_o + \beta_1 \ln(x) + \xi$$

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \xi$$

In contrast, a nonlinear model is nonlinear in the parameters. The SAS/STAT documentation states "A model is nonlinear in the parameters if the derivative of the model with respect to a parameter depends on this or other parameters [3] Examples of nonlinear models are shown in below figure.

$$Y = \beta_1 e^{-\beta_2 x} + \xi$$

$$Y = \beta_1 + (\beta_2 - \beta_1)\, e^{-\beta_3 x} + \xi$$

$$Y = ((\beta_1 x)/(\beta_2 + x)) + \xi$$

When fitting a linear model, the general form of the model is known. It is only necessary to identify the dependent and independent variables and the terms to include in the model. For a nonlinear model, the form of the model must be specified, the parameters to be estimated identified, and starting values for those parameters must be provided. When obtaining estimates of the parameters in a linear model, specific equations are used to determine the single unique solution that results in the smallest error sum

of squares possible for the model and the data table. In the case of a nonlinear model, iterative techniques are generally used to converge on a solution that provides the best fit model. It is important for the analyst to check to be sure that convergence has been attained. There are certain conditions which can prevent convergence. These conditions may include:

- incorrect specification of the model
- poor initial starting values for the parameters
- over defined model
- insufficient iterations
- insufficient data

Incorrect model specification is when the data do not exhibit the same relationship as the one given in the model. Poor initial starting values can cause lack of convergence or convergence to a local minimum. A model that is over defined is one that has more parameters than are necessary for the relationship. If two or more parameters are very close in value, they could, and should, be represented by a single parameter. In some cases, the estimation method is slowly approaching convergence, but convergence is not obtained before the maximum number of iterations is reached. Finally, there might be insufficient data to generate a model. In this case, additional data must be collected [4]

CONCLUSION AND FUTURE WORK

We studied two different machine learning techniques and the way they can use for *isolated* spelling. The study helps us in drawing the following ideas:

A. *Here definitely, linear models can solve the problem but to what extent, it has to be explored is a question of how we implement it.*

B. *Non- linear model seems to be promising to solve the problem in better way as in the model we tend to segregate each analysis separately giving its own weightage by non-linear dependent variables. In the case, the line of best fit will be closer to the unknown curve. In linear model we are restricting the line of best fit to take only linear path, that way error might be more and scope of the curve is very much reduced.*

Spell correction has different verticals to be covered, not just misspellings or partially spelled errors. There are a few domains which should be considered as main streams while trying to create a complete spell corrector. One is the distance based features, second would be probability based using a likelihood statistical model and finally the soundex. Soundex[14] is a phonetic algorithm for indexing names by sound, as pronounced in English. The goal is for homophones to be encoded to the same representation so that they can be matched despite minor differences in spelling. The algorithm mainly encodes consonants; a vowel will not be encoded unless it is the first letter. One could also do keystroke analysis to create a statistical model for most likelihood keystrokes close by and use the derived probability too.

Apart from this, other statistical estimations methods to reduce the error are the same. We can identify the best estimation method depending on the line of best fit and error model. Be it any research, the pitch is always only a start and we do not know if this will work until and otherwise it is tried and verified. The algorithm proposed is logical in paper but necessary implementation and experiments could tell us more, which we intend to continue and hopefully end up with a spell corrector which uses machine learning.

ACKNOWLEDGMENT

We express our sincere gratitude to Department of Computer Science at Amrita School of Engineering, Coimbatore for providing us the opportunity to carry out this project.

We extend our heartiest gratitude to our peers Mr. Kumar Vaibhav, Mr. Prashant Gupta, Mr. Aditya Bhandare and Mr. Muthu Arvind for their valuable guidance and feedback.

REFERENCES

[1] A winnow-based approach to context-sensitive spelling correction: AR Golding, D Roth - Machine learning, 1999 – Springer
[2] DOI 10.1023/A:1007545901558
[3] http://norvig.com/spell-correct.html : Peter Norvig's spell corrector based on Bayesian model
[4] SAS Institute Inc. 2013. SAS/STAT® User's Guide. Cary, NC: SAS Institute Inc J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
[5] Rawlings, John O., Sastry G. Pantula, and David A. Dickey. Applied Regression Analysis: A Research Tool. New York: Springer-Verlag, 1998. Print.
[6] Estimating linear restrictions on regression coefficients for multivariate normal distributions: Ann. Math. Statist., 22 (1951), pp. 327–351
[7] Y. Fujikoshi, The Likelihood Ratio Tests for the Dimensionality of Regression Coefficients : J. Multivariate Anal., 4 (1974), pp. 327–340
[8] Fred J. Damerau, A technique for computer detection and correction of spelling errors, Communications of the ACM, v.7 n.3, p.171-176, March 1964
[9] doi>10.1145/363958.363994
[10] Karen Kukich, Technique for automatically correcting words in text, ACM Computing Surveys (CSUR), v.24 n.4, p.377-439, Dec. 1992
[11] doi>10.1145/146370.146380
[12] Levenshtein, V. (1966). "Binary codes capable of correcting deletions, insertions and reversals."
[13] COMLEX Pronouncing Lexicon, Version 0.2. Linguistic Data Consortium LDC95L3, July 1995.
[14] L.R. Bahl and F. Jelinek, "Decoding for Channels With Insertions, Deletions, and Substitutions With Applications to Speech Recognition," IEEE Trans. Information Theory, vol. 21, no. 4, pp. 404-411, 1975.
[15] Patent : US8589149 *, Aug 5, 2008 -Nov 19, 2013 Nuance Communications, Inc. Probability-based approach to recognition of user-entered data
[16] L. Baum and J. Eagon, "An Inequality With Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to Models for Ecology," Bulletin Am. Mathematical Soc. 73, pp. 360-363, 1967. doi>10.1109/34.682181
[17] http://nlp.stanford.edu/IR-book/html/htmledition/phonetic-correction-1.html
[18] An Improved Error Model for Noisy Channel Spelling Correction : Eric Brill and Robert C. Moore, Microsoft Research
[19] Measuring Semantic Similarity between Words Using Web Spell Correction