ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Context Based Spell Checking using Document Semantic

Amalu Laji
Department of Computer Science
Christ University
Bangalore, India

Dr. Karthik K
Department of Computer Science
Christ University
Bangalore, India

Abstract: Context Based errors are those errors which are wrongly used in sentences, though it seems valid. These errors would turn sentences into meaningless, which in turned to invalid documents. These words would not rectify in context or would do an automatic spelling checking. This paper explores the idea of Document Semantic called Latent Semantic analysis which can correct the context from this minor difference. This paper also reflects the importance of using this terminology.

Keywords: stop word removal; stemming; term-frequency; singular value decomposition; cosine similarity.

I. INTRODUCTION

Words are heart of sentences. Each word join a sentence. And each sentence join a concept. Spelling checkers are those applications that check the misspelled words in a sentence. These kinds of applications are embedded in all major word processing systems. However, words that are incorrect to the corresponding sentence would not detect in word processing. Those incorrect words are called Confusion words. Confusion words are those words which seems valid but not valid to the corresponding sentence. Confusion words can change concept of particular sentence and can change to meaningless document. Confusion words include quiet and quite, principle and principal etc. Another error can occur if the user even confused with the particular word to use on corresponding sentence like affect and effect etc. These minor but essential situations could not handle by word processor. The situation will turn worse if the error could not handle. Context based Spelling Correction is a closely related to class of problems that include word sense disambiguation, word choice selection in machine translation and accent and capitalization restoration [6]. These class of problems has been attacked by many users which turns unresolved. To tackle this situation, semantics is introduced. Semantic between the sentences in a document can handle the situation. The method called Latent Semantic Analysis has been proved to bring concept based sentences. Latent semantic analysis (LSA) is a statistical method for constructing semantic spaces. It can be viewed as a component of a psychological theory of meaning as well as a powerful tool with a wide range of applications, including machine grading of clinical case summaries. With the use of this method, the machine will be able to depict human thoughts. It could automatically cover the misused words turns to meaningful document. There are some data mining techniques which can use to find only the useful information and make use of it [8].

It is a machine learning method for representing the meaning of words, sentences, and texts. The system could predict the misused words from the given document by checking from confusion sets which is imported in database. With the use of this method, the machine will be able to depict human thoughts. It could automatically cover the misused words turn to meaningful document.

Latent Semantic Analysis is a technique in natural language processing. It is a method of extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the totality of information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and set of words to each other. To perform this, LSA would analyses the relationship between the set of documents and the terms they contain by producing a set of concepts related to the documents and terms. It would create a semantic space with words in the given set of documents and LSA will assumes words that are closer in meaning.

Latent Semantic Analysis is a technique in natural language processing. It is a method of extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the totality of information about all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and set of words to each other. To perform this, LSA would analyses the relationship between the set of documents and the terms they contain by producing a set of concepts related to the documents and terms. It would create a semantic space with words in the given set of documents and LSA will assumes words that are closer in meaning.

II. MATEIALS AND METHODS

The materials required are collected from Brown Corpus. In order to check LSA's ability to check context words, the data is divided into two cases. One is training and the other is testing. I am collecting certain amount of confusion words from Brown Corpus and selecting random sentences regarding each confusion words and randomly dividing it into two cases. This helps Latent Semantic Analysis to compare the higher chance of word to a particular sentence.80% of the data are assigned as a training sets and other 20% of the data are assigned as testing sets. LSA will take only those particular sentences which contain theses confusion words. Similar way, in test case Latent Semantic Analysis will only test the particular sentence which contained confused words.

For performing Latent Semantic Analysis, the documents are needed to processed and construct an LSA space. For constructing an LSA space, term-document matrix is required in LSA matrix. While training a set of data from Brown Corpus, certain transformation required in data.
The initial stage after collecting the data from Brown Corpus is to remove unnecessary symbols like @, $ etc., avoid names, values or fields with blank spaces otherwise each word will be interpret separately causing errors.

A. Preprocessing Technique

They are words which are filtered out before or after processing of natural language data. Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search. Any group of words can be chosen as the stop words for a given purpose. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as "The Who", "The The", or "Take That". Using Porters algorithm [8], we make use of these technique.

B. Stemming

Stemming is the process of reducing inflected or sometimes derived words to their word stem, base or root form—generally a written word form. In [6] it is also says that Stemming algorithms can change words into grammatical root form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

C. Term-frequency

This method is mainly used for removing noise from data by first representing the data into high dimensional space and reduce dimensionality. In this method, each row contains terms and each column contain documents which contain sentences. Each cell contain number of times a particular word occur in particular sentence. A count matrix for both training and testing has been applied. This can make LSA to implement much faster. The below Figure 1 shows the sample of count matrix.

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

Figure 1: Count Matrix

D. Single Value Decomposition

After evaluating count matrix in training and testing data set, Singular value decomposition (SVD) is performed. This method is decomposed into three factor as S, U and V Transpose. The S matrix is the representation of original term vectors of derived orthogonal factor values. U is the diagonal matrix square of rank. V Transpose is the similar representation of original document vectors. Multiplying these factors would turn to original representation of text collection.

E. Cosine Similarity

Cosine Similarity is widely used in Assessing Review Quality of data [4].Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The best match between the sentences from the training and the testing set is done under the basis of cosine similarity [7]. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. The name derives from the term "direction cosine": in this case, note that unit vectors are maximally "similar" if they're parallel and maximally "dissimilar" if they're orthogonal. It should not escape the alert reader's attention that this is analogous to cosine, which is unity when the segments subtend a zero angle and zero when the segments are perpendicular.

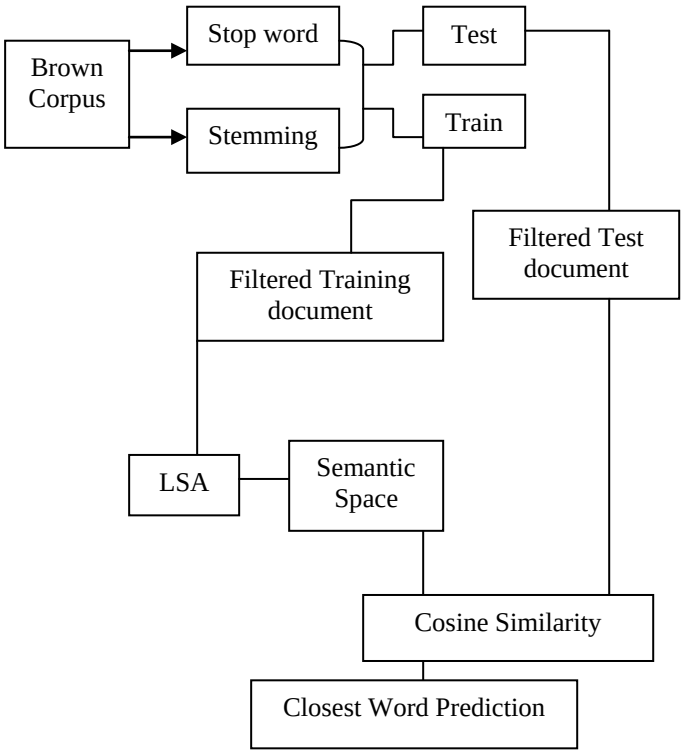

Figure 2: Flow diagram

III. LITERATURE REVIEW

In [1] the paper talks about an effective real world error detection and correction and they are using two bigram and one trigram to test the word in a sentences which could easily detect the corrected word in a sentence provided if the data is small. This paper is planning to make the correction work when it comes to large global data. This paper has not used much n-gram methods making the system less accurate in error detection. The paper is looking forward to get good accuracy by using Word-net. In [3] the paper proposed an algorithm called Winnow which is compared with Bayesian. Winnow algorithm is proposing some features and giving good accuracies when compared to Bayesian.

In [5] Michael P. Johns and James H. Martin proposed a paper "Contextual spelling Correction Using Latent Semantic Analysis" that describe about the misused errors, confusion sets, Latent Semantic Analysis and the experimented they conducted. The paper points out the traditional word processing system which flag only the misspelled words. But this paper shows the way a machine can depict the human mind. The paper talks about identify the misused words that are used incorrectly in a particular sentence. These misused words generally generated from confusion sets. The paper introduces a machine learning method Latent Semantic Analysis that can differentiate into matrix form as terms and documents, compare the words and mapped with confusion words. The paper motivation of using Latent Semantic Analysis was to test its effectiveness at predicting words based on a given sentence and to compare it into Bayesian Classifier.

The Research paper which is to be implementing is "Contextual Spelling Correction Using Latent Semantic Analysis" where it is defined as the use of an incorrect, through valid, word in a particular sentence or context. It talks about the traditional spell checking and points out that the traditional system does not catch the error if a misused word used in a sentence. We explore the use of Latent Semantic Analysis for correcting these incorrectly used words and the results are compared to earlier work based on Bayesian classifier.

The paper shows the experimental details. Firstly, they collected the data. Separate corpora for training and testing LSA's ability to correct contextual word usage errors were created from the Brown corpus. The Brown corpus was parsed into individual sentences which are randomly assigned to either a training corpus or a test corpus. Roughly 80% of the original corpus was assigned as the training corpus and the other 20% was reserved as the test corpus. Training the system consists of processing the training sentences and constructing an LSA space from them. LSA requires the corpus to be segmented into documents. For a given confusion set, an LSA space is constructed by treating each training sentence as a document. Each training sentence is used as a column in the LSA matrix. There are some transformation undergoes in sentences.

Context reduction is a step in which the sentence is reduced in size to the confusion word plus the seven words on either side of the word or up to the sentence boundary. The average sentence length in the corpus is 28 words, so this step has the effect of reducing the size of the data to approximately half the original. Intuitively, the reduction ought to improve performance by disallowing the distantly located words in long sentences to have any influence on the prediction of the confusion word because they usually have little or nothing to do with the selection of the proper word. In practice, however, the reduction we use had little effect on the predictions obtained from the LSA space.

Stemming is the process of reducing each word to its morphological root. The goal is to treat the different morphological variants of a word as the same entity. For example, the words smile, smiled, smiles, smiling, and smilingly are reduced to the root smile and treated equally. We tried different stemming algorithms and all improved the predictive performance of LSA. The results presented in this paper are based on Porter's algorithm.

Bigram creation is performed for the words that were not removed in the context reduction step. Bigrams are formed between all adjacent pairs of words. The bigrams are treated as additional terms during the LSA space construction process. In other words, the bigrams fill their own row in the LSA matrix.

Term weighting is an effort to increase the weight or importance of certain terms in the high dimensional space. A local and global weighting is given to each term in each sentence. The local weight is a combination of the raw count of the particular term in the sentence and the term's proximity to the confusion word. Terms located nearer to the confusion word are given additional weight in a linearly decreasing manner. The local weight of each term is then flattened by taking its log2. The global weight given to each term is an attempt to measure its predictive power in the corpus as a whole. We found that performed best as a global measure. Furthermore, terms which did not appear in more than one sentence in the training corpus were removed.

We tested the predictive accuracy of the LSA space in the following manner. A sentence from the test corpus is selected and the location of the confusion word in the sentence is treated as an unknown word which must be predicted. One at a time, the words from the confusion set are inserted into the sentence at the location of the word to be predicted and the

same transformations that the training sentences undergo are applied to the test sentence. The inserted confusion word is then removed from the sentence (but not the bigrams of which it is a part) because its presence biases the comparison which occurs later. A vector in LSA space is constructed from the resulting terms. The word predicted most likely to appear in a sentence is determined by comparing the similarity of each test sentence vector to each confusion word vector from the LSA space. Vector similarity is evaluated by computing the cosine between two vectors. The pair of sentence and confusion word vectors with the largest cosine is identified and the corresponding confusion word is chosen as the most likely word for the test sentence. The predicted word is compared to the correct word and a tally of correct predictions is kept.

The paper proven to be an effective alternative to Bayesian classifiers. The LSA prediction accuracy is 91%.

The paper has conducted an experimental method like prepossessing technique like stop words removal, context reduction, stemming, weighing and then its leads to clustering and Baseline prediction. The paper has explained the experiment methods clearly that leads to a good result. The paper could able to show that the Latent Semantic Analysis can attack the problem of identifying contextual misuses of words, particularly those words are the same parts of speech. The paper proven to be an effective alternative to Bayesian classifier. The LSA prediction accuracy was 91%.

In [2] Andrew R Golding and Dan Roth proposed a paper where we apply Winnow based algorithm to a task in natural language context-sensitive spelling correction. This is the task of fixing spelling errors that happens to result in valid words, such as substituting to for too, casual for casusal and so on. Previous approaches to this problem have been statistical based approach. This paper compare Winnow to one of the most successful such approaches, which uses Bayesian classifiers. This paper finds that when the standard set of features is used to describe problem instances. Winnow performs comparably to the Bayesian method. When the full set of features is used, Winnow is able to exploit the new features and convincingly outperform Bayes and when a test set is encountered that is dissimilar to the training set. Winnow is better than Bayes at adapting to the unfamiliar test set, using a strategy we will present for combining learning on the training set with unsupervised learning on the test.

IV. RESULT

This paper has shown that LSA has the power of attacking the misused words in sentences. It makes predictions by building a high dimensional semantic space which is used to compare the similarity of the words from the confusion sets to a given context. Documents with confused words are given in both training and testing sets to find the best matching word by relating with concepts. LSA has given 75% accuracy in selecting correct word from sentences.

V.ACNOWLEDGMENT

I am deeply indebted to our guide Prof. Karthik K for simulating suggestions and encouragement. It helped me in the successful completion of the project and this document. I thank him for constantly monitoring and providing me with constructive feedback. I offer thanks to all the faculty members of the Department for their support

VI. REFERENCES

[1] Samanta, Pratip, and Bidyut B. Chaudhuri. "A simple real-word error detection and correction using local word bigram and trigram." ROCLING. 2013.

[2] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.

[3] Golding, Andrew R., and Dan Roth. "Applying winnow to context-sensitive spelling correction." arXiv preprint cmp-lg/9607024 (1996).

[4] Ramachandran, Lakshmi, and Edward F. Gehringer. "Automated assessment of review quality using latent semantic analysis." Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on. IEEE, 2011.

[5] Jones, Michael P., and James H. Martin. "Contextual spelling correction using latent semantic analysis." Proceedings of the fifth conference on applied natural language processing. Association for Computational Linguistics, 1997.

[6] Ramasubramanian, C and R. Ramya. "Effective pre-processing activities in text mining using improved porter's stemming algorithm." International Journal of Advanced Research in Computer and Communication Engineering 2.12 (2013): 4536-8.

[7] Foltz, Peter W. "Latent semantic analysis for text-based research." Behavior Research Methods, Instruments, & Computers 28.2 (1996): 197-202.

[8] Vijayarani, S., Ms J. Ilamathi, and Ms Nithya. "Preprocessing Techniques for Text Mining-An Overview." International Journal of Computer Science & Communication Networks 5.1 (2015): 7-16.