Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 99 – 106

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems

A correction model for real-word errors

Sumit sharma[a], Swadha Gupta[b]

[a] *Department of Computer Science, Chandigarh University, Gharuan Mohali and 140413 INDIA*
[b] *Department of Computer Science, Chandigarh University, Gharuan Mohali and 140413 INDIA*

Abstract

Spell Checker is used to identify and correct mistakes made by users while writing text and the mistakes are generally spelling mistakes. An intelligent spelling correction system SMC is proposed to automatically correct spelling mistakes in text-editor or text documents using contextual information of the confused words. The system is capable to correct words belonging to the set of confused words fed into it if they are contextually wrong. In this technique, an algorithm to identify and correct real-word errors is proposed. One phase of algorithm uses trigram approach to correct spelling mistakes and the other phase of algorithm uses Bayesian approach to correct spelling mistakes. Brown corpus is used as a training set and a set of commonly confused words is used in this case. Selection of words in other phase of algorithm uses synonyms derived from dictionary in the scenario when words are not found in the corpus. Comparative analysis of the proposed approach with tribayes has also been performed to identify the accuracy of SMC. The results indicate that SMC gives higher accuracy for spelling mistakes identification and correction for the commonly confused words as compared to other spelling correction algorithms.

Peer-review under responsibility of the Organizing Committee of ICECCS 2015

Keywords: Real-word errors; Spelling mistakes; Spelling corrector; Modified corpus; Supervised approach; Unsupervised approach.

1. Introduction

Writing is one of the predominant and vital ways of communication through which humans can express their views to others and keep a record. It is one of the most effective forms of language representation. It does the task of representing language by engraving signs and symbols. Writing is composed of vocabulary, semantics and grammar. Text is considered as the outcome of writing. Diary, books, manuals, newspaper, publication have given inspiration to write. Writing has been contributory in keeping and preserving records, evolution of legal system and circulation of information via media. With the rising technology and life changing inventions, computer came into picture.

Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.047

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
*E-mail address:*swadhagupta@rocketmail.com

The scalability of writing text has increased, due to which many issues such as spelling mistakes also evolved. Not every person is proficient in representing language. Some carry out the task of formal writing and others do free writing, depending on their respective goals. Free writing is a task where writer writes while ignoring grammatical and spelling mistakes, which means the chances of making mistakes, is very high. As mistakes can sidetrack readers from the efforts, the writer has put in his writing. Therefore, it becomes indispensable to remove these mistakes. Hence, it prompted the need to use spelling checker so that errors can be minimized while writing. Spell Checkers are either part of large applications, for instance search engines, email-clients etc or stand-alone application that is efficient of performing correction on a piece of text.

Nearly all word processors have a built-in Spelling checker that flags the spelling mistakes. It also provides the solution to correct these spelling mistakes by choosing a possible alternative from a given list. For identification of spelling mistakes, most spellcheckers checks each word drawn separately from the written text against the dictionary-stored words. If the word is found while searching the dictionary, it is considered as correct word regardless of its context. This approach is efficient for identifying the non-word spelling mistakes but other mistakes cannot be identified using this method. The other mistakes such as real-word spelling mistakes i.e. words that are correctly spelled but are not intended by the user. Mistakes falling under this category go unrecognized by most spellcheckers because they handle non-word spelling mistakes by checking against the dictionary word list only. This technique is effective to identify the non-word spelling mistakes but not the real-word spelling mistakes. To identify the real-word spelling mistakes, there is a need to utilize the neighboring contextual information of the target word. An example of such sentence is “I want to eat a piece of cake” and the confused word set in this caseis(piece, peace),to identify that ‘peace’ cannot be used in this case, we utilize the neighboring contextual information ‘cake’ for word ‘piece’.

This paper is organized as follows: first, the work in the field of real-word correction is described. Then proposed framework along with the experimental results is presented. This paper closes with a discussion of the choices made in formulating this methodology and plans for future work.

2. Advancements in field

The need of real-word correction became prominent in the mid of 90’s.This gained attention of researchers to usher in the field of real-word error correction. James L. Peterson [3] discussed the errors which spelling checker computer program could not detect. Spell checker works efficiently for identifying and correcting non-word errors but fails to identify and correct context-sensitive errors. For a non-word spell checker, it checks the word against the list of the given words. If the word exists in the list, then it is considered as correct otherwise flagged as incorrect word. The addition of extra word in the word list is the solution to this problem. The researchers have tried to increase the list to detect the undetected errors but they found that the percentage of undetected errors increase by increasing the list size. The new large list contains not only words but also the code which gives the information regarding the misspelling of a word. The percentage of undetected errors also increased because of new increased list size. It is concluded that word from the word list should be adopted according to the topic and situation for which it is to be used. In addition to it, there is need of intelligent spell checker that detects and corrects both syntax and semantic errors in a sentence.

Eric Mays [4] introduced a statistical approach to deal with the problem of context-spelling error efficiently. In this, the sum of 100 sentences is taken arbitrarily considering that it contains words from our vocabulary, fifty sentences from the documentations of the Parliament of Canada and remaining fifty from the AP newswire. A list of 20000 words is employed from speech recognition project of IBM along with their respective trigram probabilities. The correct sentences are transformed into misspelled sentences. A list of correct sentences and list of 20,000 words are considered as training set and manually transformed incorrect sentences are considered as test sets. The probability of sentences is calculated by using the maximum likelihood estimation of probability.

David Yarowsky [5] introduced a learning algorithm to find out the sense of word that has more than one meaning so that it can be correctly used in a sentence. The learning method used in this case is unsupervised and the training set used here is without tagging. The two concepts used in this case are one sense per discourse and one sense per collocation. The former means one word always reveals one meaning if used in a particular context and the latter means words that are neighbors of the target word, provides the information regarding the recognition of it. Words can exist in more than one collocation, so the advantage of this feature of word has been used to sense disambiguity. The researchers first took only a subset of disambiguate words and then made them learn to differentiate so that a word can be used in an intended situation. The knowledge obtained from a subset of words is applied to the whole sample. In this case, conflicts are fixed by using only one evidence rather than using integration of multiple evidences.

Andrew R. Golding [6] introduced Tribayes, which is based on trigram, and Bayes to correct the context-sensitive spellings errors. Trigram is based on parts-of-speech of words and Bayes is based on features. Tribayes has used the best of both methods to deal with the problem of real-word errors. Bayes is used in case of same POS tagging of confused words and Trigram is used for different POS tagging. They used brown corpus as their training set and commonly confused words as their data set. The commonly confused words are repeating words from the brown corpus. After applying Tribayes method, the probability is calculated and appropriate word is substituted.

Andrew R. Golding [7] implemented Bayesian hybrid to resolve the problem of context-sensitive spelling errors. In this, Bayesian classifier method puts forward decision lists to make best use of both context texts and collocations. To solve the problem of context-sensitive error, the collected evidences are transformed into a single piece of information. When it was applied to the real-word spelling errors correction, it performed much better than the component methods. Lidia Mangu [8] illustrated new way to correct the real-word spelling problem. In this, the newly proposed approach learned the linguistic knowledge automatically to correct the context sensitive spelling errors. Acquiring information in small set of rules is one of the important characteristic of this approach and is easily understandable. Rather than emphasis on large set of features and weight, it focused more on small set of rules. With the help of given technique, the machine can automatically understand and learn the rules. The learning based algorithm that is used to make the machine learn and understand the rules is called Transformation-Based learning.

Andrew R. Golding [9] proposed method called WinSpell to identify and correct context-sensitive spelling mistakes. It is one of the most efficient algorithms till date for correcting real-word or context-sensitive spelling mistakes or errors. In Winspell, the features are not pruned like Bayspell. The features during the training of Winspell are extracted and their weights are calculated and further assigned to them. In the same way, list of active features is created from the given sentence during the testing of Winspell after learning from the set of learned features. The connection between classifier and active features is created to distinguish one word from other words in the confusion set. The classifiers utilized the variants of Winspell algorithm, applied algorithm called weighted majority, which stored different values. The appropriate connection is created with the help of training, and furthermore their respective weights are learnt. It utilizes information from multiple classifiers (features) rather than using single classifier to decide on the substitution of intended word. One of the best characteristics of Winspell is that it is trained and tested using different corpora and still outperformed other methods that have utilized the same corpus for both purposes. It has used supervised learning for training and unsupervised leaning for testing.

Davide Fossati [10] used mixed trigram model to correct the real-word errors. In this, the POS tagging is performed in order to tag the sentences using the Stanford tagger. The tagged sentences having the confusion words are compared with the HMM (hidden markov model) labeled tags. If difference is detected while comparing tags it means there is a misspelling in the sentence and then mixed trigram is applied to correct it. A new empirical grounded technique is used to create the dataset of confused words. A corpus with large dataset of misspelled words is used and its probability is calculated. A test set is generated artificially by arbitrarily replacing words, as there is unavailability of appropriate test set to test the real-word errors. The precision of context sensitive spell checker is increased. Therefore the outcome of results have exhibited increase in coverage of spell checker using the mixed trigram model.

Ya Zhou [11] proposed a method known as RCW (real-word correction) for the real-word spelling errors based on tribayes. Due to inadequate training set, there is exclusion of essential features. In this, Word Net is used to extract the pruned features and the problem of pruned features is solved to a certain degree. Trigram performs well in case of different tagging of words in data set and Bayes performs well in case of same tagging of word in data set. RCW[11] has used the complementary of both to get the best results. The weight of context words are calculated

based on their contextual information, and considered as the determining feature for the correction of real-word spelling mistakes or errors. Furthermore, synonyms from Word Net are used in place of effective features that are pruned in order to improve the accuracy.

The limitation of using currently available corpus is that the contextual information is limited to only the text available in the corpus. Corpus contained contextual information of only limited text but does not cover all the contextual information of the language. Therefore, the scope of correcting mistakes remained limited to only the contextual information available in the corpus. Thus, the real-word error correction is limited to only small data set and its accuracy is also reduced.

It is being concluded after extensive analysis of literature that the identification and correction of real-word spelling errors or mistakes can be performed efficiently with trigram and Bayesian technique. Both techniques work well for real-word spelling correction.

3. Proposed framework

Spelling correction is an application used to identify and correct the spelling mistakes in the text written by the user. Conventional spell checker fix only non-word errors and the real-word errors that gives valid words but are not intended by the user goes undetected. Correcting this kind of problem requires a totally different approach from those used in the conventional spell checker. Considering this problem, SMC method is proposed, which is based on trigram, and Bayesian approaches but used both in different ways unlike used in other algorithms. This method is able to solve the problem to a certain extent by using all the features of the sentences unlike other methods that uses only 2 or 3 features. The approach also aims at retrieving the synonyms of the words, which is not available in the corpus by extracting synonyms from the dictionary of their corresponding words.

1.1 Training feature

Brown corpus [13] is used as a training set in this proposed method.

1.2 Trigram

Trigram approach takes full benefit of the data that is present in the surroundings of the target word i.e. collocation features. Trigram calculates the probability of all words in a sentence and adds all the calculated probabilities of a sentence. The probabilities of all the ambiguous words in the confusion set are calculated by substituting them one by one in a sentence. The target word having the highest probability is substituted in the final outcome and is considered as correct word.

$$p(w3|w1,w2) = \frac{f(w1,w2,w3)}{f(w1,w2)} \quad (1)$$

f(w1,w2,w3)→count of w3 is seen following w2 and w1in brown corpus
f(w1,w2)→count of w2 is seen following w1.

1.3 Bayesian approach

Bayesian approach takes full benefit of the data that is present in the surroundings of the target word i.e. context words. It extracts all the words surrounding the target word and names it as features. From the training corpus that is containing correct articles, Bayesian approach learns about the contextual information surrounding the target word. The probabilities of context words are calculated, which is based on corpus i.e. by calculating the frequency of occurrences of features individually and the frequency of occurrences of features along with the target word. If the feature is not found in the corpus then synonym of that particular feature is extracted from the dictionary and its probability is calculated. The synonym having the highest probability is substituted in place of its corresponding feature. The ambiguous word having the highest score is substituted in the final outcome and is considered as correct word.

$$\textbf{Value(f)}_{[10]} = log\left(\frac{p(c,a)}{p(c)*p(a)}\right) \quad (2)$$

p(w, a) → the joint probability between p(c) and p(a)
p(c) → probability of feature of the target word
p(a) → probability of target word

$$\textbf{Sum}(w_a) = \sum_{ci \in C} value(fi) + \sum_{sj \in S} \max(sj) \quad (3)$$

function max(s_j) → highest value of all synonyms of the feature s_j
function value(f_i) is used to calculate the value of feature c_i
c_a → ambiguous word

Bayesian approach is used when the POS tagging of ambiguous words are same else, in case of different tagging, its performance will degrade.

The following procedure summarizes the algorithm:
Input: Sentence T =w_1, w_2, w_3..,w_i,...,w_n ∈

w_j → input word
X → {w_i, $w_i{}^c$} is confusion set

Output: Corrected w_i **If** w_i ∈ X.
If w_j ∈ C
then tag the whole sentence
Goto Step 2
Else
print " word not found in data set"
For i= 0,1,2....,n where i ∈ X do
If w_j ∈ w_i, having different POS brown corpus tagging
then
trigram is applied
Extract the collocation(T) ∈ A

A → training set

Find fr(col)

fr(col) → frequencies of the collocation of sentences

Combine corresponding collocations and frequencies.
Calculate p(w_i)

p(w_i) → probability of w_a in the C

$$\mathbf{p(w3|w1,w2)} = \frac{f(w1,w2,w3)}{f(w1,w2)}$$

f(w1,w2,w3) → count of w3 is seen following w2 and w1in brown corpus
f(w1,w2) → count of w2 is seen following w1.

Print the word with highest probability
Else If w_j ∈ w_i, having same POS brown corpus tagging
then
bayes is applied
Extract the context words ∈ A

A → training set

Find the fr(c)

fr(c) → frequencies of the context words

Combine corresponding features and frequencies.

If fr(c) = 0

then

extract S

$S \rightarrow$ *Synonyms of* w_i

Calculate the sum of w_i.

$$\textbf{Value(f)} = log\left(\frac{p(c,a)}{p(c)*p(a)}\right)$$

p(w, a) → *the joint probability between p(c) and p(a)*
p(c) → *probability of feature of the target word*
p(a) → *probability of target word*

$$\textbf{Sum}(w_a) = \sum_{ci \in C} value(fi) + \sum_{sj \in S} \max(sj)$$

function $max(s_j)$ → *highest value of all synonyms of the feature* s_j
function $value(f_i)$ *is used to calculate the value of feature* c_i
c_a → *ambiguous word*

Print the word with highest probability

In above algorithm, different corpora are used for training set and testing set. Supervised learning approach is used for the training corpus, which is manually enhanced brown corpus, and unsupervised learning approach is used for testing and the test-set is a manually created set of incorrect sentences. We have supposed that the text in training and testing sets contains no spelling mistakes. Frequently occurring words in Brown corpus [13] is selected as the confusion sets. Test set is unsupervised as nobody indicates whether the spelling of the word it checks is correct or incorrect. SMC is better than RCW[11] to adapt because of the utilization of supervised and unsupervised strategy, and used Bayesian and trigram approaches in a different way by calculation probabilities of all the words surrounding the target word. We found that, using this strategy, the performance of SMC is able to improve on an unfamiliar test set.

4. Experimental results

In this empirical study, two widely used and publicly available datasets i.e. brown[13] corpus and set of confused words are used to evaluate our proposed system. Brown corpus contains 1,014,312 words sampled from 15 text categories and set of 30 confused words are used. When the POS tagging for ambiguous words are different then trigram is used and when it is same, then bayes method is used. Trigram uses the context information in the form of collocation and bayes uses context information in the form of features. The probabilities for ambiguous words are calculated in both cases of trigram and bayes and word having highest probability is selected. The calculation of probabilities for set of 30 confused words is shown in Table 1.

1. Input: Tom sat on the curb.

Table 1: The information of the feature and value of ambiguous words- (curb, kerb)

Collocation features			Total
Tom sat on	sat on the	on the curb	
0.43859646	0.4385965	0.20	1.07719296
Tom sat on	sat on the	on the kerb	
0.2385965	0.1385965	0.50	0.877193

5.2 Cumulative Empirical Analysis of Complete Confusion Set

The experimental results of SMC and RCW[11] is shown in Table 2

Table 2: The result of real-word error correction in SMC and RCW

S.No.	Confusion items	No. of test cases	SMC		RCW	
			No. of correct	Accuracy	No. of correct	Accuracy
1.	accept, except	20	19	95%	16	80.00%
2.	capital, capitol	20	19	95%	17	85.00%
3.	among, between	20	18	90%	17	85.00%
4.	brake, break	20	19	95%	17	85.00%
5.	farther, further	20	18	90%	18	90.00%
6.	formerly, formally	20	17	85%	15	75.00%
7.	hear, here	20	18	90%	18	90.00%
8.	instance, instant	20	19	95%	15	75.00%
9.	passed, past	20	18	90%	17	85.00%
10.	peace, piece	20	15	75%	18	90.00%
11.	principal, principle	20	19	95%	17	85.00%
12.	raise, rise	20	19	95%	16	80.00%
13.	sea, see	20	17	85%	20	100.00%
14.	stationary, stationery	20	17	85%	15	75.00%
15.	waist, waste	20	19	95%	17	85.00%
16.	weak, week	20	18	90%	19	95.00%
17.	than, then	20	18	90%	19	95.00%
18.	adverse, averse	20	18	90%	15	75.00%
19.	altar ,alter	20	16	80%	16	80.00%
20.	appraise ,apprise	20	18	90%	18	90.00%
21.	loose, lose	20	18	90%	18	90.00%
22.	pour, pore	20	18	90%	18	90.00%
23.	bare, bear	20	18	90%	17	85.00%
24.	censure, censor	20	19	95%	18	90.00%
25.	curb, kerb	20	16	80%	17	85.00%
26.	currant, current	20	18	90%	18	90.00%
27.	duel, dual	20	17	85%	18	90.00%
28.	storey, story	20	19	95%	19	95.00%
	Average	/	/	89.83%	/	86.16%

The results of comparison are shown in above table. Number of test cases implies number of times confused words occurred in test corpus. Number of correct implies the number of cases SMC method corrected. The accuracy achieved for SMC is 89.83% according to the results obtained as compared to RCW[11], which is 86.16%. Increase in accuracy of SMC for the set of test cases taken corroborates the fact that incorporating all the features of the sentence and using synonyms of the words not found leads towards better results in typo correction.

5. Conclusion

Misspelled words, which are present in articles created by human, are a common phenomenon and these misspelled words can be classified as either non-word errors or real-word errors. In this work, we proposed SMC system for automatically identifying and correcting real-word spelling mistakes. Considering real-word errors, the ambiguous words in the confusion set are identified by context information consisting of collocation and context words. To deal with the problem of real-word errors, an algorithm using trigram and Bayes methods is proposed. Supervised and unsupervised learning strategy is used in this work. Supervised learning is used for training and unsupervised is used for testing. Unsupervised learning is used so that this work can be applied to any context of the text. Brown corpus [13] is used as a training set and manually created sentences are used as a test set because of unavailability of test sets. The algorithm is run on the data of 30 confused sets that is extracted from “Words Commonly Confused” [14]. The algorithm also takes an advantage of the synonyms of the context words, which are not found in the brown corpus. We empirically evaluated and compared SMC with RCW[11] and achieved an error correction accuracy of

89.83% for real-word errors. Our work showed significantly higher performance for real-word errors when compared with RCW[11].

6. Future Scope

Although the SMC system developed in this research has gained some course of success in identifying and correcting the real-word spelling mistakes, it has also suggested several issues that needs to be addressed in the future development. Implementation and comparison with latest proposed model such as MS Word 2007 is done. Identification and correction of real-word errors is limited to small confusion sets. Therefore, the scalability of correction needs to be improved so that large set of words could be corrected. A large corpus is required that includes text of all possible contextual information. The number of real-word correction per sentence at a time is one; hence, algorithm has to be modified to do multiple corrections per sentence. Integration of proposed approach with the conventional spell checker needs to be done.

References

1. Yinghao Huang, Yi Lu Murphey and Yao Ge, “Automotive diagnosis typo correction using domain knowledge and machine learning.” IEEE Symposium Series on Computational Intelligence, pp 267-274 (2013).
2. Kukich K. Technique for Automatically Correcting Words in Text. *ACM Computing Surveys (CSUR)* ,vol. 24, pp. 377-439 (1992).
3. Peterson, James L. "A note on undetected typing errors." Communications of the ACM 29.7, pp 633-637, Year 1986.
4. Mays, Eric, Fred J. Damerau, and Robert L. Mercer. "Context based spelling correction." Information Processing & Management 27.5, pp 517-522 (1991).
5. Yarowsky, David. "Unsupervised word sense disambiguation rivaling supervised methods." Proceedings of the 33rd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics,pp 189-196 (1995).
6. Golding, Andrew R., and Yves Schabes. "Combining trigram-based and feature-based methods for context-sensitive spelling correction." Proceedings of the 34th annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp 71-78 (1996).
7. Golding, Andrew R. "A Bayesian hybrid method for context-sensitive spelling correction." arXiv preprint cmp-lg/9606001, pp 1-15 (1996).
8. Mangu, Lidia, and Eric Brill. "Automatic rule acquisition for spelling correction."ICML, Vol. 97, pp 187-194 (1997).
9. Golding, Andrew R., and Dan Roth. "A winnow-based approach to context-sensitive spelling correction." Machine learning 34.1-3, pp 107-130 (1999).
10. Fossati, Davide, and Barbara Di Eugenio. "I saw TREE trees in the park: How to Correct Real-Word Spelling Mistakes." LREC, pp 896-901 (2008).
11. Zhou, Ya, et al. "A Correcting Model Based on Tribayes for Real-Word Errors in English Essays." Computational Intelligence and Design (ISCID), 2012 Fifth International Symposium on IEEE, Vol. 1, pp 407-410 (2012).
12. Wilcox-O’Hearn, Amber, Graeme Hirst, and Alexander Budanitsky. "Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model." Computational Linguistics and Intelligent Text Processing. Springer Berlin Heidelberg, pp 605-616 (2008).
13. Francis, W. Nelson, and Henry Kucera. "Brown corpus manual." Brown University (1979).
14. Oxford University Press, “Oxford American Large Print Dictionary", Year 2008.