Building a Real Word Spell Checker based on Power Links

Mahmoud Rokaya
Taif University College of Computer & Information Technology Taif, Saudi Arabia
Permanent Address: Tanta, University Faculty of Science, Egypt

Sultan Aljahdali
Taif University College of Computer & Information Technology Taif, Saudi Arabia

ABSTRACT

A context-based spelling error is a spelling or typing error that turns an intended word into another word of language. Most of the methods that tried to solve this problem were depended on the confusion sets. Confusion set are collection of words where each word in the confusion set is ambiguous with the other words in the same set. the machine learning and statistical methods depend on predefined confusion sets. In this paper, the presented method by Rokaya to define the confusion sets depending on the content and external dictionaries is adopted. A merging between this method and WinSpell to develop a refined automatic context spell checker. This method joins between the advantages of statistical and machine learning method and the re-source based methods.

General Terms

Artificial Intelligence and Natural Language Processing

Keywords

Field Association Terms, Power Link, Context Spelling Checkers, WinSpell, WinFat.

1. INTRODUCTION

Since the first work by Glantz, [1] , a great deal of researches has taken place on the subject of spelling verification and correction. [5]

An approximate word matching algorithm is required to identify errors in queries where little or no contextual information is available and using some measure of similarity and recommend words that are most similar to each misspelled word(Hodge and Austin, 2003)

The problem of creating or developing algorithms for automatically catching and correcting spelling errors has become a primary challenge for researchers in the last few decades. Kukich divided the spelling errors into three types, non-word errors, isolated word errors and real word errors. In this paper, the real word errors are considered. This is the class of real-word errors in which one correctly spelled word is substituted for another. Some of these errors result from simple typos (e.g., from + form, form + farm) or cognitive or phonetic lapses (e.g., there + their, ingenious + ingenuous); some are syntactic or grammatical mistakes, including the use of the wrong inflected form (e.g., arrives ~ arrive, was + were) or the wrong function word (e.g., for + of, his ~ her); others are semantic anomalies (e.g., in five minuets, lave a message); and still others are due to insertions or deletions of whole words (e.g., the system has been operating system for almost three years, at absolutely extra cost) or improper spacing, including both splits and run-ons (e.g., myself ~ my self, ad here - adhere). These errors all seem to require information from the surrounding context for both detection and correction. Contextual information would be helpful also for improving correction accuracy for detectable nonword errors. [4].

The methods which tried to solve this problem fall in two classes: the first class is those methods that based on human made lexical, the other class is those methods that based on statistics or machine language.

An example of the first class is the method of Hirst and Budanitsky, [8]. They presented a method for correcting real-word spelling errors by restoring lexical cohesion. This method detects and corrects real word spelling errors by identifying tokens that are semantically unrelated to their context and are spelling variations of words that would be related to the context. Relatedness to context is determined by a measure of semantic distance initially proposed by Jiang and Conrath, [6].

An example of the second class is the method of Wilcox et. al., [10]. They presented a statistical method based on trigrams for correcting real-word spelling correction. In this method, they made a reconsideration of the trigram-based noisy-channel model of real-word spelling-error correction that was presented by Mays et. al., which has never been adequately evaluated or compared with other methods. They analyzed the advantages and limitations of the method, and presented a new evaluation that enables a meaningful comparison with the WordNet-based method of Hirst and Budanitsky. [2]

Typically, the machine learning and statistical approaches rely on pre-defined confusion sets, which are sets (usually pairs) of commonly confounded words, such as {their, there, they're} and {principle, principal}. The methods learn the characteristics of typical context for each member of the set and detect situations in which one member occurs in context that is more typical of another. Such methods, therefore, are inherently limited to a set of common, predefined errors, but such errors can include both content and function words. [10].

By contrast, the resource-based methods are not limited in this way, and can potentially detect a confounding of any two words listed in the resource that are spelling variations of one another, but these methods can operate only on errors in which both the error and the intended word are content words. [10]

Dix et. al., described briefly three systems: onCue a desktop internet-access toolbar, Snippet a web-based bookmarking application and ontoPIM an ontology-based personal task-management system. These embody context issues to differing degrees, and they used them to exemplify more general issues concerning the use of contextual information in intelligent interfaces.[9]

Rokaya and Atlam, [11], proposed the concept of power link. The power link algorithm was suggested to measure how tow terms tend to appear together in a given corps. If the value of the power link between two terms was high then the chance that one of the terms is substituted using the other term is low. This means that those two terms cannot be confused.

Rokaya et. al., [14], proposed a method to extract and refine the confusion sets based on power link algorithm. This method joins between the advantages of statistical and machine learning method and the re-source based methods. The values of precision, recall and F indicated that the proposed algorithm can produce in average 90%, 70% and 78% respectively which means that the algorithm tends to produce a low percentage of false negative errors. The value of F indicates the strong of the algorithm.

This work will try to recover the limitation of pre-defined errors by presenting an algorithm which is capable of detecting the errors. This means that the algorithm will start by checking every token in a given document and it will determine the candidates that can replace a given token depending on the automatic construction confusion sets algorithm [14]. The number of the alternative can by as much as the algorithm can guess. If the number of the alternative exceeds three the power algorithm will be used to decide what terms should be removed from the confusion set.

The new method combines between the advantages of WinSpell algorithm and the power of the automatic construction of confusion sets. Also some vital modifications of WinSpell are proposed and tested.

The remaining sections of this paper are organized as follows. Section 2 review the concept of the power link and confusion sets method construction. Section3 presents the details of the proposed method. Finally, section 4 provides the experiments and its results. Section 4 also discusses the results and provide the impact of each proposed factor in the algorithm.

2. The power link and confusion sets construction method

The term power link was proposed by Rokaya and Atlam, [11], as a method of building dynamic field association terms dictionary. Power link algorithm presented new rules to improve the quality of filed association terms (FATs) dictionary in English [13] .

The origin of this concept comes from the co-word analysis researches. Co-word analysis considers the dynamics of science as a result of actor strategies. Changes in the content of a subject area are the combined effect of a large number of individual strategies. This technique should allow us in principle to identity the actors and explain the global dynamic (Callon et al., 1991).

If any two terms t_1 and t_2 belongs to a document D it will be said that there is a link between t_1 and t_2. The power of this link is measured by the function $LT(t_1, t_2)$ where:

$$LT(t_1,t_2) = \frac{|D| \times cr(t_1,t_2)}{\underset{i,j}{averageL}(t_{1i},t_{2j})} \qquad (1),$$

where $|D|$ is the number of different terms in the document D, $cr(t_1,t_2)$ is the co-occurrence frequency of t_1 and t_2 in the document D and $\underset{i,j}{averageL}(t_{1i},t_{2j})$ represents the average distance between any instants t_{1i} and t_{2j} of the terms t_1 and t_2 in the document D. For more details see [11].

To estimate the power like between two terms t_1 and t_2 over a given corps, the function $LCORPS(t_1,t_2)$ is defined. This function can be defined as:

$$LCORPS(t_1,t_2) = \underset{D \in corps}{average}(LT(t_1,t_2)) \qquad (2)$$

This function states that the terms t_1 and t_2 will tend to appear nearer together if the value of this function reasonably high. To give a threshold many values were experimentally has been tried around the mean value for the power link. This means that the threshold is not unique and it is dynamic. In fact it is dependent on the given corps. For the experiments the mean value was 31.5 so the algorithm is activated with values around this mean to cover the interval (mean-STD, mean+STD), where STD is the standard deviation of the mean value.

Confusion set is defined that specifies a list of confusable words, e.g., {their, there}

or {cite, site, sight} [12]

In the following part, an algorithm to extract and refine confusion sets depending on the power link concepts.

Depending on these concepts Rokaya et., al. [14] presented a method to produce the confusion sets Both the training corps and the refined confusion sets represents the input data for the context spelling checking algorithm (WinFat Algorithm).

3. The Context Spelling Checking Algorithm (WinFAT Algorithm)

Two types of features are used. Context feature that test for the presence of a word that has a power link greater than Θ within ±k words of the target word. Using the power link impose a type of pre pruning and reduce the probability to produce a rare extracted features. Collocations features test for a pattern of up to l contiguous words and/or part of speech tags around the target word [7].

The feature extractor is used to convert a given sentence to the corresponding active feature list. The extractor has a preprocessing phase in which it learns a set of features corresponding to a given task. When the extractor get a sentence the extractor will convert the sentence into a list of active features through matching the sentence against the set of learned features. Instead of depending on an imposed condition for pruning or a complicated process the algorithm implements the co-occurrence information for the pruning process give a sense in terms of utilizing the strong relations among words within the given corps.

There are many methods for using a learning algorithm. Hidden Markov models are a powerful technique to model and classify temporal sequences, such as in speech and gesture recognition. However, defining these models is still an art: the designer has to establish by trial and error the number of hidden states, the relevant observations,.. ,etc. [3].

This section introduces the proposed algorithm. The algorithm adopts the winnow method with the following modifications:

1- The automatic generating of confusion sets, section 2, which provide a dynamic procedures to get the confusion sets, errors, based on the languages dictionaries and the given training corps.

2- The pruning process uses a different approach, section 4, This approach depend on the power link algorithm to refine the extracted features.

3- Bernard and Sapir, 2003, mentioned that the expert rule is far more likely to be optimal than the majority rule. But they mentioned also that combining all the experts in the final decision makes the majority rule is the loser and gives the expert rule better chance to be optimal. They suggested to use a restricted majority rule or balanced expert rule. For this reason the weighted majority is replaced by the modified restricted weighted majority approach.

Let n be the number of classifiers Cj. For each Cj, , j= 1,2,3,..., n let Pj be the probability of correctness. The values of logarithmic expertise levels is $f(p_1), f(p_2),..., f(p_n)$ where

$$f(p_i) = \ln\frac{p_j}{1-p_j}, \; j = 1,2,...,n \tag{3}$$

Definition1: The restricted majority rule of order k = 2s+1 (where 1 ≤ s ≤ n/2) is characterized by assigning equal weights to the k most competent group members and zero weights to the remaining members. The restricted rule of order k = 2s+1 is optimal if and only if

$$\sum_{j=n-2s}^{n-s} w_j \geq \sum_{j=1}^{n-2s-1} w_j + \sum_{j=n-s+1}^{n} w_j \tag{4}$$

where $w_1, w_2,..., w_n$ are the ordered values of $f(p_1), f(p_2),..., f(p_n)$. (Bernard and Sapir, 2003).

The probability of success $p_j, j = 1,2,...,n$ is defined by:

$$p_j = \frac{\text{Total number of correct decision}}{\text{Total number of decision}} \tag{5}$$

The restricted majority rule will be used to combine the results of the classifiers $C_j, j = 1,2,...,n$. The restricted majority rule takes the form

$$\frac{\sum_{j=k}^{n}\gamma^{m_j}C_j}{\sum_{j=k}^{n}\gamma^{m_j}} \tag{6}$$

where k is the value that guarantees the optimality of the restricted weighted majority rule. Also the value of γ is chosen to guarantees the optimality of the restricted weighted majority rule.

According to Bernard and Sapir, 2003, the value of γ that guarantees the optimality of the restricted weighted majority rule is given by:

$$\gamma(n,s) = \frac{n}{2^n}\cdot\frac{\binom{n-1}{s,s,n-2s-1}}{(2s+1)(s+1)^{n-s-1}s^s} \tag{7}$$

This method to combine the classifiers results is expected to be optimal. Also this method has a mathematical proof for its optimality over any other combining methods.

4. Evaluation

To get a chance for positive comparison and fair results. I followed Hirst and Budanitsky, [8], in using the 1987–89 Wall Street Journal corpus (approximately 30 million words), which is presumed to be essentially free of errors. 500 articles are reserved (approximately 300,000 words) to create test data. Also the standard tokenization, and the Good–Turing smoothing and Katz back off techniques of the toolkit are adopted. [10]

To create a refined test sets, real word errors are automatically inserted in the reserved set of testing. Instead of using a fixed density distribution a varied density distribution is used. The properties of the power link are used to produce a balanced real errors e according to the following rule. Let ρ be the average of power link contained in a single document D, then the number of artificial errors added to this document is

$$e = \frac{N \times p}{TN} \tag{8}$$

where, N is the number of unique words in D and TN is the total number of tokens in D. If the resulting number of errors e is greater than $\max\rho$ (max power link value in the document D) then, e is set to equal $\max\rho$. Also if the resulting I number of errors e is smaller than $\min\rho$ (min power link value in the document D) then, e is set to equal $\min\rho$. This balanced inserting of real errors prevent to harm the natural power link distribution and guarantee an accepted density of real errors in each document. Note that the number of errors is proportional to the number of unique tokens in the document. This follows a simple rule that when many different words are written the probability for writing some errors is increased. Also putting the maximum and minimum of the power link as a boundaries for the number of errors guarantees that the inserted errors will not affect the distribution of the power link between terms in a single document. A spelling variation is defined to be a singlecharacter insertion, deletion, or replacement. [8]. This method, for insertion errors, is called FATI [14]

In this evaluation four experiments are designed to test each modification that was added to the original algorithm.

Evaluate the three modifications that were added to the winnow algorithm.

Evaluate the overall improvement resulting from applying the three modification.

In the results, the path of Wilcox et. al., [8], is adopted.

They created three test sets, each containing 15,555 sentences, which varied according to which words were candidates for replacement and for substitution:

T20: Any word in the 20,000-word vocabulary of the trigram model could be replaced by a spelling variation from the same vocabulary; this replicates MDM's style of test set.

T62: Any word in the 62,000 most frequent words in the corpus could be replaced by a spelling variation from the same vocabulary; this reflects real typing errors much better than T20.

Table 1: results of applying WinSpell without modification

	Detection			Correction		
α	**P**	**R**	**F**	**P**	**R**	**F**
Test set T20						
0.9	0.334	0.647	0.441	0.327	0.618	0.428
0.99	0.474	0.668	0.555	0.467	0.547	0.504
0.995	0.546	0.636	0.588	0.539	0.616	0.575
0.999	0.594	0.658	0.624	0.690	0.543	0.608
FATI	0.529	0.559	0.544	0.607	0.440	0.510
Test set T62						
0.9	0.235	0.437	0.306	0.229	0.419	0.296
0.99	0.347	0.478	0.402	0.341	0.366	0.353
0.995	0.423	0.460	0.441	0.417	0.350	0.381
0.999	0.593	0.400	0.478	0.590	0.395	0.473
FATI	0.599	0.396	0.477	0.667	0.438	0.529
Test set Mal						
0.9	0.145	0.367	0.208	0.140	0.352	0.200
0.99	0.306	0.320	0.313	0.299	0.310	0.304
0.995	0.371	0.304	0.334	0.365	0.296	0.327
0.999	0.446	0.261	0.329	0.443	0.257	0.325
FATI	0.379	0.313	0.343	0.421	0.239	0.305
Test set MFATC						
0.9	0.112	0.496	0.183	0.105	0.471	0.172
0.99	0.298	0.436	0.354	0.29	0.419	0.343
0.995	0.359	0.41	0.383	0.353	0.397	0.374
0.999	0.52	0.344	0.414	0.516	0.336	0.407
FATI	0.588	0.378	0.460	0.573	0.380	0.457

Mal: Any content word listed as a noun in Word-Net (but regardless of whether it was used as a noun in the text; there was no syntactic analysis) could be replaced by any spelling variation found in the lexicon of the ispell spelling checker; this replicates Hirst and Budanitsky's "malapropism" data.

MFATC: Every confusion set was tested to classify it according to each of the three classes. For confusion sets that does not belong to any of these classes are placed in a fourth class. The proposed approach is applied to this class and is used as a parameter to test the approach independently. In the results, this class is called MFATC.

Table 2: WinFat algorithm using the first modification, pruning based on FAT properties

	Detection			Correction		
α	**P**	**R**	**F**	**P**	**R**	**F**
Test set T20:						
0.9	0.218	0.728	0.336	0.299	0.791	0.434
0.99	0.532	0.642	0.582	0.495	0.811	0.615
0.995	0.572	0.708	0.633	0.601	0.688	0.642
0.999	0.738	0.527	0.615	0.711	0.599	0.650
FATI	0.834	0.585	0.688	0.775	0.665	0.716
Test set T62:						
0.9	0.319	0.839	0.462	0.322	0.82	0.462
0.99	0.54	0.785	0.640	0.541	0.758	0.631
0.995	0.608	0.716	0.658	0.595	0.736	0.658
0.999	0.756	0.666	0.708	0.761	0.667	0.711
FATI	0.854	0.753	0.800	0.837	0.754	0.793
Test set Mal:						
0.9	0.212	0.596	0.313	0.205	0.571	0.302
0.99	0.398	0.536	0.457	0.39	0.519	0.445
0.995	0.459	0.51	0.483	0.453	0.497	0.474
0.999	0.62	0.444	0.517	0.616	0.436	0.511
FATI	0.701	0.488	0.576	0.684	0.493	0.573
Test set MFATC						
0.9	0.212	0.696	0.325	0.205	0.647	0.311
0.99	0.498	0.636	0.559	0.429	0.641	0.514
0.995	0.559	0.61	0.583	0.543	0.539	0.542
0.999	0.72	0.544	0.620	0.751	0.533	0.624
FATI	0.814	0.598	0.690	0.834	0.603	0.700

WinFat approach is applied to each of the classes. First set of experiments were applied to test the WinSpell. Table 1 indicates that the performance of WinSpell is lower than the performance of Wilcox et. al., [10]. This fact is the main reason for which the proposed approach was suggested. Since two modifications were proposed for pruning and voting, three experiments were designed to test each modification separately and to test the overall improvement resulting from applying the two modifications at once

The results in Table 1 for precision and recall insures the poor performance for the WinSpell. There is no significance difference in the performance of the algorithm among different test groups.

Table 2 reflects the slight improvements in performance of the WinFat algorithm after considering the new pruning method depending on the power link properties. The last row in each testing set insures that the top improvement came with applying the FATI insertion method. Table 3 shows results of applying the new rule for voting, power link voting.

Table 3: Results of WinFat considering the modified voting rule

Detection				Correction		
α	**P**	**R**	**F**	**P**	**R**	**F**
Test set T20:						
0.9	0.283	0.887	0.429	0.263	0.838	0.400
0.99	0.505	0.778	0.613	0.505	0.768	0.609
0.995	0.576	0.776	0.661	0.576	0.747	0.650
0.999	0.747	0.668	0.705	0.737	0.667	0.700
FATI	0.845	0.755	0.797	0.833	0.753	0.791
Test set T62:						
0.9	0.118	0.661	0.201	0.118	0.589	0.196
0.99	0.337	0.554	0.419	0.337	0.553	0.418
0.995	0.444	0.552	0.492	0.443	0.520	0.479
0.999	0.661	0.445	0.532	0.601	0.444	0.511
FATI	0.747	0.503	0.601	0.679	0.502	0.577
Test set Mal:						
0.9	0.278	0.856	0.420	0.288	0.844	0.429
0.99	0.501	0.778	0.610	0.511	0.776	0.616
0.995	0.578	0.746	0.651	0.588	0.744	0.657

Results in Table 3 reflects a slight improvement in the values of the precision, the values of F reflects that the performance is accepted. Also the row of FATI inserting error method still gives the highest rate of performance. Comparing results of Tables 6 and 7 there is no significance proof whether to apply any of these modifications and leave the other. The question now suggests itself. What about applying the two modifications at once. Table 4 shows results of applying the modified rules of pruning and voting together.

Table 4 insures the benifits of combining the tow modifications together at once. There are significance improvements in values of precision P and F that the results of applying the modifications together supports each other to improve the performance. Also the row of FATI insertion of errors gives the best results at all levels of modifications. Place

5. Conclusion

In this work, based on our method that was proposed for automatic construction of confusion sets (errors) for a given dictionary of terms and corresponding corps. and two modifications, of the WinSpell algorithm, the WinFat algorithm is presented. This method is proposed for context spelling checking. The effect of each new modification was tested individually. Each of them gave a significance improvement for the WinSpell performance. Applying both modifications at once gave the best performance. Finally a new method of inserting artificial errors was proposed to be more realistic. The results of the WinFat performance insures that the proposed modifications has an effective impact on performance..

Table 4 Results of applying the modified rules of pruning and voting together.

Detection				Correction		
α	**P**	**R**	**F**	**P**	**R**	**F**
Test set T20:						
0.9	0.429	0.786	0.555	0.429	0.783	0.554
0.99	0.752	0.678	0.713	0.752	0.676	0.712
0.995	0.759	0.676	0.715	0.759	0.674	0.714
0.999	0.875	0.568	0.689	0.874	0.567	0.688
FATI	0.962	0.619	0.753	0.944	0.606	0.739
Test set T62:						
0.9	0.319	0.756	0.449	0.319	0.658	0.429
0.99	0.538	0.654	0.591	0.538	0.652	0.589
0.995	0.545	0.651	0.593	0.644	0.650	0.647
0.999	0.761	0.645	0.698	0.761	0.544	0.634
FATI	0.807	0.696	0.748	0.829	0.598	0.695
Test set Mal:						
0.9	0.433	0.785	0.479	0.533	0.782	0.467
0.99	0.657	0.677	0.657	0.657	0.775	0.645
0.995	0.765	0.674	0.688	0.764	0.772	0.675
0.999	0.879	0.566	0.720	0.779	0.643	0.709
FATI	0.868	0.672	0.795	0.787	0.576	0.780
Test set MFATC						
0.9	0.424	0.654	0.327	0.423	0.652	0.318
0.99	0.545	0.548	0.462	0.644	0.547	0.453
0.995	0.652	0.546	0.489	0.752	0.645	0.481
0.999	0.769	0.640	0.507	0.769	0.440	0.502
FATI	0.878	0.546	0.576	0.780	0.544	0.561

6. REFERENCES

[1] Glantz, 1957, On The Recognition Of Information With A Digital Computer, J. Acm, Vol. 4, No. 2, pp. 178-188.

[2] Mays E., Damerau J., And Mercer L., 1991, Context Based Spelling Correction. Information Processing And Management, 23(5): pp. 517-522.

[3] Mart´Inez M. And Sucar L., 1991, Learning Dynamic Naive Bayesian Classifiers, Proceedings Of The Twenty-First International Flairs Conference, pp. 655-659.

[4] Kukich K., 1992, Techniques For Automatically Correcting Words In Text. Acm Comput. Surv. 24(4): pp. 377-439.

[5] Boubaker M., 1994, Logic Compression Of Dictionaries For Multilingual Spelling Checkers, August Coling '94: Proceedings Of The 15th Conference On Computational Linguistics - Volume 1, pp. 293-296.

[6] Jiang J. And Conrath W., 1997, Semantic Similarity Based On Corpus Statistics And Lexical Taxonomy. Proceedings Of 10th Interna-Tional Conference On Research In Computational Linguistics.

[7] Golding A., And Roth D., 1999, A Winnow-Based Approach To Context-Sensitive Spelling Correction, Machine Learning 34, pp. 107–130.

[8] Hirst G. And Budanitsky A., 2005, Correcting Real-Word Spelling Errors By Restoring Lexical Cohesion, Natural Language Engineering, 11(1), pp. 87--111, March.

[9] Dix, A., Katifori, A., Lepouras, G., Vassilakis, C., Shabir, N., 2010, Spreading Activation Over Ontologies: From Personal Context To Web Scale Reasoning, International Journal of Semantic Computing, Special issue on Web Scale Reasoning, Volume: 4, Issue: 1(2010) pp. 59-102.

[10] Wilcox A.; Hirst, G.; And Budanitsky A., 2008, Real-Word Spelling Correction With Trigrams: A Reconsideration Of The Mays, Damerau, And Mercer Model. Proceedings, 9th International Conference On Intelligent Text Processing And Computational Linguistics (Cicling-2008), Haifa, pp. 605–616.

[11] Rokaya M., And Atlam S., 2010, Building Of Field Association Terms Based On Links, Int. J. Computer Applications In Technology, Vol. 38, No. 4, pp. 298–305.

[12] Rozovskaya A., and Roth D., 2010, Generating confusion sets for context-sensitive error correction, EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

[13] Rokaya M., And Nahla A., 2011, Building A Multi-Lingual Field Association Terms Dictionary, International Journal Of Computer Science And Network Security, Vol 11, No. 3, pp. 208-213.

[14] Rokaya M., Nahla A., and Aljahdali S., 2012, Context-Sensitive Spell Checking Based on Field Association Terms. IJCSNS International Journal Of Computer Science And Network Security. 12(5), pp 64-68.