Spelling Checker using Algorithm Damerau Levenshtein Distance and Cosine Similarity

Nur Hamidah [a,1], Novi Yusliani [b,2], Desty Rodiah [c,3]

[a, b, c] Teknik Informatika, Universitas Sriwijaya, Palembang, Indonesia
[1] hamidahnur964@gmail.com; [2] novi.yusliani@gmail.com; [3] destyrodiah@gmail.com

ARTICLE INFO

Article history
Received
Revised

Keywords
Dictionary Lookup
Damerau Levenshtein Disstance
Spelling Checker
Cosine Similarity
Mean Reciprocal Rank

ABSTRACT

Writing is an embodiment of the author's ideas that are to be conveyed to others. A writer often experiences typos in typing the script, so that it can influence the meaning of the text. Therefore, a system is needed to detect word errors. In this study, checking is done by using the Dictionary Lookup method and giving the candidate words using the Damerau Levenshtein Distance algorithm. Candidates will then determine the ranking by breaking the word into Bigram form and calculating the similarity value using the Cosine Similarity algorithm. The test results based on the data used yield different Mean Reciprocal Rank (MRR) values for each type of error. The type of error deletion produces an MRR value of 88.89%, the type of insertion error produces an MRR value of 97.78%, the type of substitution error produces an MRR value of 88.89%, the type of transposition error produces an MRR value of 89%.

1. Introduction

Writing is a manifestation of the writer's ideas that are intended to be conveyed to others. Writing skills are important skills in life, both in educational and community life. Writing activity is one of the most recent manifestations of language abilities and skills mastered by language users after listening, reading, and speaking [1]. A writer often experiences typos in writing his script, so that it can affect the meaning of the writing. Correcting it manually can take a long time because it is done repeatedly to get results that are completely free of typing errors (typographical errors).

2. Method

In this study, we used the dictionary lookup method to search for wrong words, the Damerau Levenshtein Distance algorithm to determine the distance between words in the dictionary and wrong words to produce word candidates, then the Cosine Similarity algorithm to sort the word candidates.

In this section is a brief explanation of the algorithm in this research :

a. Spelling Checker

Error detection in words can be done with computer-based applications that are used to detect and handle errors in words called spelling checkers. The spelling checker looks for all types of errors contained in the document which then warns the writer about the mistakes made and gives some suggestions to correct the errors. There are two main methods used to build a spelling checker application, namely identification (error detection) and correction (error correction) [2].

b. Typographical Error

Typographical error is an error that occurs during the process of typing text and can change the meaning of a word and even the meaning of a sentence. This term includes errors due to mechanical failure or slip of the hand or finger, and also arises due to the ignorance of the author such as spelling mistakes. Typographical errors can be caused by, for example, fingers pressing two adjacent keyboard keys simultaneously. Spelling errors of words consist of two forms namely [3]:

1. Non-word errors are errors that focus on words formed generally by typos. These non-spelling errors produce words that are not meaningful.

2. Real word error is a mistake that emphasizes handling the placement of words in a sentence. Error words that are actually in accordance with the rules of language but not according to the meaning in the sentence.

There are 4 types of typographical errors:

- Deletion (Letter deletion)
- Insertion (Addition of letters)
- Substitution (Substitution of letters)
- Transposition (Exchange letter position)

c. Preprocessing

Preprocessing is an initial process of managing data that aims to prepare text into data that will undergo further processing. The steps taken include [4] :

1) Case Folding is a process of equating letters in a document, from uppercase to lowercase letters. So, only the letters 'a' to 'z' will be accepted. Characters other than letters are omitted and are considered delimiter.
2) Tokenizing is the process of breaking a sentence into the smallest units (words / tokens).

d. Damerau Levenshtein Distance

Damerau Levenshtein Distance determines the minimum number of operations needed to convert one string into another string, where the operation used is the same as Levenshtein Distance, namely insertion, deletion, substitution but with the addition of transposition operations between two characters [4].

e. N-Gram

N-Gram is a series of substrings along the n characters of a string [4]. N-Gram is a method implemented for word or character generation. With the N-Gram method, a word or sentence is cut into pieces of character letters with a number of n [4].

f. Term Frequency-Inverse Document Frequency (TF-IDF)

The tf-idf method is a method used to determine how far the word / term is related to a document by weighting each word [5]. The tf-idf formula in (1) follows:

$$W_{ij} = tf_{ij} \text{ x } idf$$
$$W_{ij} = tf_{ij} \text{ x } log\frac{N}{n} \quad (1)$$

Note : w_ij is term weight for documents, tf _ij is number of occurrences of the word, idf is number of documents containing term appears, N is the sum of all documents compared, amd n is number of documents containing term.

g. Cosine Similarity

Similarity is a function used to measure the degree of similarity between two vectors. In the text, this function is used as a measure of the similarity between the query and every document in the database. From this calculation, the level of similarity generated in the document in accordance with the query entered [6].

Cosine similarity is a formula used to calculate similarity by determining the angle between a document vector and a query vector in the V dimension in the Euclidean plane. The result of cosine similarity has a value between 0 and 1. The value 0 is the value obtained if the document is not related to the query, while the value of 1 means the document has a high connection with the query [7]. The cosine similarity formula in (2) is as follows:

$$Cos\ \alpha = \frac{\mathbf{A} \bullet \mathbf{B}}{|\mathbf{A}||\mathbf{B}|} = \frac{\sum_{i=1}^{n} \mathrm{A_i} \times \mathrm{B_i}}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (2)$$

Note : A is Vector A, which will be compared similarity, B is Vector B, which will be compared similarity, A • B is dot product between vector A and vector B, | A | is length of vector A, | B | is length of vector B, and | A || B | is cross product between | A | and | B |.

h. Mean Reciprocal Rank (MRR)

Reciprocal Rank (RR) is the inverse or inverse of the target rank in the list of suggested words displayed. Mean Reciprocal Rank (MRR) is the average word rank rating correct of all queries. MRR is a statistical measure that is suitable for evaluating question search rankings [8]. The results of the MRR have values between 0 and 1. A value of 0 is obtained if there is not a single target included in the word candidate results displayed, while a value of 1 is obtained if all targets are ranked 1 of all the candidate results displayed. The MRR formula in (3) is as follows:

$$\text{RR} = \frac{1}{r_{ji}}$$

$$\text{MRR} = \frac{1}{q} \sum_{j=1}^{q} \frac{1}{Q_j} \sum_{i=1}^{Q_j} RR_{ji} \quad (3)$$

3. Results and Discussion

Software testing is carried out using abstract data of the final project of Informatics Engineering Faculty of Computer Science Sriwijaya University. The data consists of 4 documents, there is 1 document consists of 30 deletion error type words, 1 document consists of 30 insertion error type words, 1 document consists of 30 substitution error type words, and 1 document consists of 30 transposition error type words. Graph of Mean Reciprocal Rank (MRR) values can be seen in Figure 1.

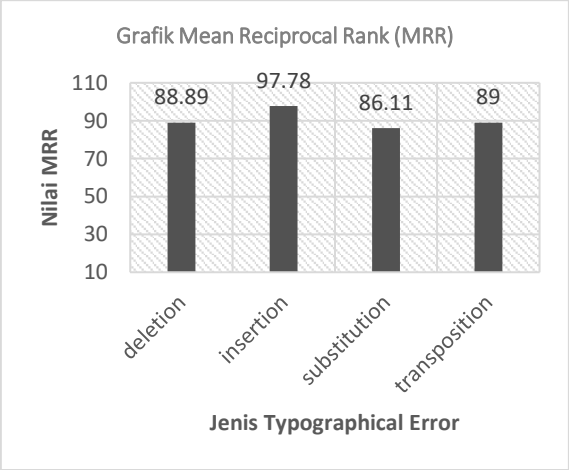

Fig. 1. MRR Value Chart

Figure 1 shows the results of the Mean Reciprocal Rank (MRR) value for each type of error. This type of error deletion produces MRR value of 88.89%, this is because there are several word candidates ranked 2nd and 3rd and are outside the top 5 ranking. Type of insertion error produces MRR value of 97%, this is because the average word candidate is ranked 1. Types of substitution errors produce MRR value of 86.11%, this is because there are several word candidates ranked in rank 2 and 3 and are outside Top 5 ranking. Types of transposition errors produce MRR value of 89%, this is because there are several candidate words that are ranked 2nd and some are ranked 5. Types of errors Insertion produces the largest MRR value because in terms of word structure have the most bigram similarity with candidates, so average candidate words for each query are ranked 1. Damerau Levenshtein Distance's algorithm can produce actual word candidates based on misspelled words. But when sorting based on similarity values using the Cosine Similarity algorithm, the said candidate is not included in the top 5 candidates. This is because the N-grams produced produce a little bigram. In the process of sorting word candidates, if there are candidate words that have the same Cosine Similarity value, they will be sorted according to where the words are first in the dictionary.

4. Conclusion

Based on the results of the test, it can be concluded that the Damerau Levenshtein Distance algorithm can produce suitable word candidates based on four types of word errors, namely insertion, deletion, substitution, and transposition. Cosine Similarity algorithm can calculate the value of similarity between the wrong word with the word candidate.

For further similar studies, the system can be developed in the process of checking words using a more complete Indonesian dictionary. And the linkages of words in one sentence can also be considered in subsequent studies, because in this study only in the form of identification of words based on words in the dictionary or referred to as non-word errors. In addition, the system can be developed by providing word recommendations for words that were previously selected.

References

[1] Luqman, "Web-based indonesian spelling correction application program," pp.13-19, 2009.

[2] F. J. Damerau, "A technique for computer detection and correction of spelling errors," Commun. ACM, vol. 7, no. 3, pp. 171–176, 1964, doi: 10.1145/363958.363994.

[3] M. Y. Soleh and A. Purwarianti, "A non word error spell checker for Indonesian using morphologically analyzer and HMM," Proc. 2011 Int. Conf. Electr. Eng. Informatics, ICEEI 2011, no. July 2019, 2011, doi: 10.1109/ICEEI.2011.6021514.

[4] J. Jupin, J. Y. Shi, and Z. Obradovic, "Understanding cloud data using approximate string matching and edit distance," Proc. - 2012 SC Companion High Perform. Comput. Netw. Storage Anal. SCC 2012, no. May 2016, pp. 1234–1243, 2012, doi: 10.1109/SC.Companion.2012.149

[5] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity," Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 1, pp. 306–312, 2018.

[6] A. I. Fahma, "Identifikasi Kesalahan Penulisan Kata (Typographical Error) pada Dokumen Berbahasa Indonesia Menggunakan Metode N-gram dan Levenshtein Distance," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 2, no. 1, pp. 53–62, 2018.

[7] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," Proc. 2016 4th Int. Conf. Cyber IT Serv. Manag. CITSM 2016, 2016, doi: 10.1109/CITSM.2016.7577578.

[8] S. Plansangket and J. Q. Gan, "A query suggestion method combining TF-IDF and Jaccard Coefficient for interactive web search," Artif. Intell. Res., 2015, doi: 10.5430/air.v4n2p119.

[9] A. A. Abdillah and M. I. B, "Implementasi Vector Space Model untuk Pencarian Dokumen," in Seminar Nasional Matematika dan Pendidikan Matematika, 2013, pp. 0–7.

[10] A. Kornain, F. Yansen, and T. Tinaliah, "Penerapan Algoritma Jaro-Winkler Distance Untuk Sistem Pendeteksi Plagiarisme Pada Dokumen Teks Berbahasa Indonesia," Progr. Stud. Tek. Inform. STMIK GI MDP, pp. 1–10, 2014.

[11] Y. Rochmawati and R. Kusumaningrum, "Studi Perbandingan Algoritma Pencarian String dalam Metode Approximate String Matching untuk Identifikasi Kesalahan Pengetikan Teks," J. Buana Inform., 2016, doi: 10.24002/jbi.v7i2.491.

[12] T. Maghfira, I. Cholissodin, and A. Widodo, "Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance," J. Pengemb. Teknol. Inf. dan Ilmu Komput., vol. 1, no. 6, pp. 498–506, 2017.