Scientific Journal of Informatics
Vol. 8, No. 1, May 2021
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Identifying The Common Type of Spelling Error by Leveraging Levenshtein Distance and N-gram

Margareta Hardiyanti[1*]
[1]Information System Department, Faculty of Industrial Engineering,
Universitas Telkom, Indonesia

Abstract
Purpose: This study aims to identify the common type of spelling error and it uses the list of common misspelling words submitted by Wikipedia contributors.
Methods: Levenshtein and N-gram distance are utilized to predict the correct word of misspelling from English dictionary. Then, the result of both algorithms is observed and evaluated using recall metrics to determine which technique works more effectively.
Result: The result of this study shows that Levenshtein works well to correct substitution single letter and transposition two sequenced letters, while N-gram operates effectively to fix the word with letter omission. The overall result is then evaluated by recall measurement to see which technique that works well on correcting the misspellings. Since the recall of Levenshtein is higher than N-gram, it is concluded that the frequency of misspelling words that are correctly fixed by Levenshtein occurs more often.
Novelty: This is the first study that compares two spelling correction algorithms on identifying the common type of spelling error.

Keywords: Spelling Error Type, Levenshtein, N-gram
Received February 2021 / **Revised** March 2021 / **Accepted** March 2021

INTRODUCTION

During the process of document writing, a typewriter tends to produce a spelling error. According to Grudin [1], skilled typists have an average of 1% error rate and 3.2% for novice typists. A study conducted by Lunsford and Lunsford shows that 6.5% of spelling errors are found in a US national sample of college composition essays [2]. The spelling errors also frequently happen to English learners [3] and authors who lack of knowledge [4]. Apart from this, even though the typewriters are familiar with the word, they can make a mistake in typing because of slips movement in the motor coordination [5]. Lee et al. also address that the typographical error is caused by pressing the wrong key in the keyboard [6]. Based on a previous study, more than 80 percent of spelling errors are categorized into one of these following types of error, including [5]: (1) Insertion, this error appears in a word with an addition of one extra letter, such as: "bannana" instead of banana, and "smaert" for smart; (2) Deletion, this type of error occurs when one letter is missing from the word, such as: "goverment" for government, and "durin" for during; (3) Substitution, this error happens when a letter in a word is replaced by another letter such as "analize" instead of analyze, and "wvent" for event; (4)Transposition, this type of error emerges when the position of two adjacent letters in a word is switched, such as "hygeine" for hygiene, and "tets" instead of test.

Fahma, et al. conducted a study to identify the typographical error which occurs in Indonesian language documents by using Levenshtein and N-gram model [7]. While Hameed analyzed the spelling error of beginner English learners by computing the type of error manually [8]. A study conducted by Madi Murdilan, et al. used the techniques implemented in this study, Levenshtein and N-gram, to correct typographical errors made by elementary students in their Indonesian essays [9]. Okuda et al. proposed new techniques based on Levenshtein to correct distorted messages [10], while Abdulkudhur et al. leveraged an improved Levenshtein to reduce the process of comparing incorrect words to the words in dictionary list [11]. There is also an improvement of Levenshtein and N-gram method to develop a language-independent spell checker and automatic spelling correction [12][13]. Xie, et al. built a Chinese

* Corresponding author.
Email anddresses: margareta@telkomuniversity.ac.id (Hardiyanti)
DOI:10.15294/sji.v8i1.29273

spelling checking system by combining bi-gram and tri-gram which are based on N-gram as well to get a better result of correction [14].

However, most of the previous studies focused on fixing the misspelled words. There has been limited research which concerned with determining the most common type of spelling error itself. Therefore, this study intends to determine the common type of spelling errors by utilizing the spelling correction methods. The list of commonly misspelled English words that is compiled by Wikipedia contributors is chosen as the dataset of this study [15]. Accordingly, the list of misspelled words is converted into a dataset of tokens.

METHODS

Research Methods

The dataset used in this study is taken from the list of common spelling mistakes, which contains 4453 English misspelling tokens that are contributed by Wikipedia authors [15]. Along with the misspelling token, the dataset also provides the correct words for each token that are used to correct the typographical errors in Wikipedia page. For the study purpose, the correct words are used to process the computation of recall in the evaluation stage of this study.

This study performs four stages of activities which consist of:

1) Identify the Best Match(Es) of the Misspelling Tokens Using The Spelling Correction Algorithms
   In order to identify the best match(es) of the misspelling token, this study implements two programs which leverage a spelling correction method for each: Levenshtein distance and N-gram model. Those algorithm yields a distance value between words [13][16]. Hence, the token in the misspelling dataset is compared to each word in English dictionary by using those algorithms to obtain the distance result. The best match(es) of the token is a word/words from the pairs which produce the least distance value. The process of taking the token as an input to obtain the closest match(es) from the dictionary is called Dictionary Lookup [17][18].
2) Observe the Detail Result of Each Algorithm
   As Levenshtein distance and N-gram have quite a different approach, the misspelled words which are able to be corrected by each algorithm may have different nature. Therefore, the result of each spelling correction method is observed thoroughly in this stage.
3) Determine the Evaluation Metric
   In this stage, both algorithms are evaluated with particular metrics to obtain the performance of each algorithm on correcting the misspelled words. As the algorithms can produce more than one prediction for each token, accuracy is not relevant to the context of this study. Therefore, recall is considered as the appropriate measurement technique as it is capable to show the effectiveness of system or model on retrieving the relevant items effectively [19]. The relevant item is perceived as the closest match of the misspelling word.
4) Draw the conclusion based on the evaluation metric
   The result of recall aids this study to gain some insight into the algorithm comparison, so that the conclusion about the common type of spelling error can be drawn.

Levenshtein Distance

Levenshtein distance is a technique which computes the differences between two words by calculating the minimum operations required, which are insertions, deletions, or substitutions the letters to transform one word into another word [16]. Levenshtein distance has a parameter in the matrix [Match, Insert, Delete, Substitution] = [0, +1, +1, +1] [11].

One of the example tokens from the dataset is “youseff”. This token needs one insertion of ‘r’ and one substitution of ‘f’ to ‘l’ to be transformed into “yourself”, the intended word. When applying the Levenshtein distance matrix parameter, the words “yourself” and “youseff” has two operations distance.

N-gram

N-gram technique measures the distance between two strings by using the combination of subset sequence characters with the length of N for each string [12]. As an example, given a parameter N = 2, the token “yoru” and word “you” produce these substrings:

1) 2-grams of “yoru” = ‘#y’, ‘yo’, ‘or’,’ru’, ‘u#’
2) 2-grams of “your” = ‘#y’, ‘yo’, ‘ou’, ‘’ur’, r#’

The more the pairs have similar substrings, the more similar those words are. In this case, "yoru" and "your" have two similar substrings.

RESULT AND DISCUSSION

This study provides the first analysis on determining the common type of spelling error written by Wikipedia authors by leveraging the algorithm of Levenshtein distance and N-gram. Those algorithms are executed toward the dataset of token to be compared with a given dictionary so it could return the best matches word for each input token. Based on the output result, it can be observed that each method behaves differently to correct different kinds of spelling errors.

Levenshtein Distance Result

Table 1 shows some examples of spelling correction output using Levenshtein distance method. From the result that is shown in Table 1, it is observed that Levenshtein distance could effectively correct a spelling error with a single substitution:

1) "abandon" (substitution of 'o' to 'a')
2) "circuit" (substitution of 'u' to 'i')
3) "effords" (substitution of 'e' to 'a')

In addition, this method could also fix a type of typographical error with transposition of two sequenced characters. The Levenshtein distance simulated the transposition by operating one substitution and one deletion [20]:

1) "guarantees" (transposition of 'a' and 'u')
2) "personality" (transposition of 'o' and 's')
3) "typical" (transposition of 'i' and 'c')

However, this method is not really useful to correct some misspelled word with multiple characters missing:

1) "sophicated" (missing 'sti')
2) "honory" (missing 'ra')

Table 1. The example output of best string match(es) using Levenshtein Distance

Token	Correct Word	Prediction Words	Right /Wrong
abondon	abandon	abansdon bondon	√
beteen	between	bedeen beseen beteem between	√
curcuit	circuit	circuit	√
effords	affords	affords efford efforts	√
gaurantees	guarantees	grantees guarantees warrantees	√
honory	honorary	honor honora honors	×
perosnality	personality	personality	√
sophicated	sophisticated	sonicated spicated	×
typcial	typical	thecial typal typhia typical trucial	√

N-gram Result

Table 2 shows some examples of spelling correction output using N-gram method. Unlike Levenshtein distance, the method of N-gram is capable to fix the omission error even though the number of missing characters is quite significant. The method implemented in this study is bigram, as it has better correction hit rate compared to unigram or trigram [21]. As long as the subsets of the words are similar, the distance will be less. One of the examples is "foundland" is correctly predicted as "newfoundland".

Table 2. The Example Output of Best String Match(Es) Using N-gram

Token	Correct	Predictions	Right /Wrong
adres	addres	address	√
ahppen	happen	alpen arpen aspen	×
bewteen	been	been	×
chasr	chase	chaser	×
comited	committed	committed	√
derived	derived	dived	×
foundland	newfoundland	newfoundland	√
inaugure	inaugurates	inaugurates	√
remaing	remaining	remailing remaining	√
secceeded	seceded	seceded	√

However, this method is not effective to fix the transposition and substitution error. A further instance of this can be seen in Table 3. It is quite prominent that the subsets of token "between" is more similar to the subsets of "been" rather than the subsets of the correct word "between". Accordingly, the N-gram algorithm predicts "been" as the correct word instead of "between".

Table 3. The example Subset of N-gram

Word	Subsets	Category
bewteen	#b, be, ew, wt, te, ee, en, n#	Misspelled
between	#b, be, et, tw, we, ee, en, n#	Correct
been	#b, be, ee, en, n#	Prediction

Evaluation Metric

Based on the output of word predictions by Levenshtein distance and N-gram, the evaluation metrics results can be seen in Table 4.

Table 4. The recall of Levenshtein and N-gram

Method	Recall
Levenshtein	79 %
N-gram (N = 2)	65%

Table 4 indicates that the recall of Levenshtein is higher than the recall of N-gram. Levenshtein distance can predict 3520 correct words out of 4453, while N-gram only predicts 2916 correct words. It means that the presence of spelling error in the dataset that can be corrected by Levenshtein is greater than the typical spelling error which N-gram can fix effectively; transposition and single substitution. Thus, this study also contributes to the growing evidence that the single edit error holds the large percentage of misspellings that occurs in a document [5][22].

CONCLUSION

The output results show that the recall value of Levenshtein distance is higher than N-gram. As Levenshtein distance indicates the effectiveness of fixing spelling errors on substitution single character and transposition of two sequenced letters, it can be concluded that transposition and substitution errors are more commonly present in a document typed by authors.

REFERENCES

[1] J. T. Grudin, "Error Patterns in Novice and Skilled Transcription Typing," *Cogn. Asp. Ski. Typewriting*, pp. 121–143, 1983.

[2] A. A. Lunsford and K. J. Lunsford, "'Mistakes are a fact of life': A national comparative study," *Coll. Compos. Commun.*, vol. 59, no. 4, pp. 781–806, 2008.

[3] R. Nagata, H. Takamura, and G. Neubig, "Adaptive Spelling Error Correction Models for Learner English," *Procedia Comput. Sci.*, vol. 112, pp. 474–483, 2017.

[4] K. Kukich, "Techniques for Automatically Correcting Words in Text," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 377–439, 1992.

[5] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, pp. 171–176, 1964.

[6] J. H. Lee, M. Kim, and H. C. Kwon, "Deep Learning-Based Context-Sensitive Spelling Typing Error Correction," *IEEE Access*, vol. 8, pp. 152565–152578, 2020.

[7] 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.

[8] P. F. M. Hameed, "A Study of the Spelling Errors committed by Students of English in Saudi Arabia: Exploration and Remedial Measures," *Adv. Lang. Lit. Stud.*, vol. 7, no. 1, 2015.

[9] M. A. Murdilan, A. Abdiansyah, and N. Yusliani, "Sistem Koreksi Kesalahan Tulisan Pada Teks Karangan Siswa Sekolah Dasar Mengunakan Metode N-Gram Dan Levenshtein Distance," Universitas Sriwijaya, 2020.

[10] T. Okuda, E. Tanaka, and T. Kasai, "A Method for the Correction of Garbled Words Based on the Levenshtein Metric," *IEEE Trans. Comput.*, vol. C–25, no. 2, pp. 172–178, 1976.

[11] H. N. Abdulkhudhur, I. Q. Habeeb, Y. Yusof, and S. A. M. Yusof, "Implementation of improved Levenshtein algorithm for spelling correction word candidate list generation," *J. Theor. Appl. Inf. Technol.*, vol. 88, no. 3, pp. 449–455, 2016.

[12] F. Ahmed, E. W. De Luca, and A. Nürnberger, "Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness," *Polibits*, vol. 40, no. 40, pp. 39–48, 2009.

[13] J. R. Ullmann, "A Binary n-Gram Technique for Automatic Correction of Substitution, Deletion, Insertion and Reversal Errors in Words," *Comput. J.*, 1975.

[14] Q. Huang *et al.*, "Chinese Spelling Check System Based on Tri-gram Model," in *Proc. Eighth SIGHAN Workshop Chin. Lang. Process.*, 2015, pp. 128–136.

[15] "Wikipedia:Lists of common misspellings/For Machines," 2020. [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines.

[16] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet physics doklady*, vol. 10, no. 8. pp. 707–710, 1966.

[17] R. Haldar and D. Mukhopadhyay, "Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach," *ArXiv*, 2011.

[18] J. L. Peterson, "A note on undetected typing errors," *Commun. ACM*, vol. 29, no. 7, pp. 633–637, 1986.

[19] M. Buckland and F. Gey, "The Relationship between Recall and Precision," *J. Am. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, 1994.

[20] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, 2001.

[21] V. Christanti Mawardi, N. Susanto, and D. Santun Naga, "Spelling Correction for Text Documents in Bahasa Indonesia Using Finite State Automata and Levinshtein Distance Method," *MATEC Web Conf.*, vol. 164, 2018.

[22] W. J. Wilbur, W. Kim, and N. Xie, "Spelling correction in the PubMed search engine," *Inf. Retr. Boston.*, vol. 9, no. 5, pp. 543–564, 2006.