Spelling Checker Algorithm Methods for Many Languages

Novan Zukarnain
Information Systems Department, School of Information Systems, Computer Science Department, BINUS Graduate Program-Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
novan@binus.ac.id

Bahtiar Saleh Abbas
Computer Science Department, BINUS Graduate Program-Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
bahtiars@binus.edu

Suparta Wayan
Department of Informatics, Faculty of Design and Technology
Pembangunan Jaya University
South Tangerang, Indonesia
wayan.suparta@upj.ac.id

Agung Trisetyarso
Computer Science Department, BINUS Graduate Program-Doctor of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
atrisetyarso@binus.edu

Chul Ho Kang
Department of Electronics and Communication Engineering
Kwangwoon University
Seoul, South Korea
chkang5136@kw.ac.ar

***Abstract*— Spell checking plays an important role in improving document quality by identifying misspelled words in the document. The spelling check method aims to verify and correct misspelled words through a series of suggested words that are closer to the wrong word. Currently, Spell checkers for English language are well established. This article analyses several studies conducted in various types of languages other than English. Like Africa, Arabia, China, Indonesia, India, Japan, Malaysia and Thailand. We use the systematic literature Review approach to paper published 2008 - 2018. there are 23 papers to represent each language and view methods used. The result show that each language has a different Spelling Check Method. The Damerau-Levenshtein algorithm method is most often used for spelling checkers.**

Keywords—spelling checker, systematic literature review, algorithm method.

I. INTRODUCTION

A common mistake that occurs when we use a computer system is spelling errors. Entering data correctly and quickly into a computer system is a very important factor in improving the performance and workloads of individual in the company [1]. Typing errors will be fatal, especially in the medical field. where in this field there are many scientific words that are difficult to spell even though using English [2]. In research writing, high language skills are needed. this is a major difficulty for foreigners who are not native in a country [3]. Spelling errors on social media occur a lot. It can make something viral if not handled immediately it will create a cyber bullying on social media [4]. A quality of documents can be assessed from the use of the words. Therefore, correct spelling is the main key for that quality [5].

Statistically, the use of spelling justification technology is increasing in every country [6]. In this study, we will discuss the problems in spelling error and how the solution that has been researching using a different spelling checker method.

Japanese-English users do when they write in English. A Research describe an experimental result of improved spell checking for Japanese-English users that takes into account their linguistic idiosyncrasy. Using grammar, sentence analysis and Pattern matching for create an educational program that eliminate some syntactic irregularities error [3].

In Africa, the spelling checker is used to detect native African languages. It is said that the use of South African native language worsens in each individual. To overcome this problem, a Windows-based software was created called eSpellingPro sa Sesotho sa Leboa [7]. Different in West Africa, checking spelling errors are solved by creating software using trie data structure [8].

For Arabic, because of the many languages used, a standard language is used, namely Modern Standard Arabic (MSA). This is made in the form of a Web, so that it can be accessed easily. MSA uses Deep Learning to check Arabic spelling [9]. In addition, for foreigners who cannot speak Arabic, there are studies that use context words and n-gram language models as the solution. Where the Arabic language was originally converted into English and then changed into Arabic [10]. In other studies, the Levenshtein algorithm is used, which proposes vocabulary management and is then searched in the lexicon file containing the standard English database [11]. Punjabi is the most challenging language among other types of Arabic. In this study two approaches were used, namely Gurmukhi Manuscripts and Perso-Arabic Scripts. Because spelling is different Directions from left to right and vice versa [12].

The world largest computer program, Google. has been well-known worldwide for its options. One of the most concern in giving a pleasant user expertise is the spell checker. There are a unit 2 main issues in spell checker application, thats it, speed and accuracy. Googles spell checker might perform quickly with a rather smart accuracy in its correctness. Yet, the associate degree algorithmic program itself could not provide a hundred accuracy as user’s expectation. However, theres perpetually a trade-off between higher accuracy and quicker speed in spell checker. We need

978-1-7281-3333-1/19/$31.00 ©2019 IEEE

to form an associate degree algorithmic program with a suitable rate of period whereas maintaining a suitable rate of accuracy. We will implement many analyses with mathematician theorem and possibilities so as to boost the accuracy of a spell checker. Also, we are going to integrate this implementation with the Damerau-Levenshtein algorithmic program [13].

In Chinese, spell checking is very difficult. This is because there is no standard and broad Chinese. For this reason, the solution used by researchers is to use Basic Learning Rule [14]. Other paper, an approach is used different from the previous approach. For the words examined, the mapping from the desired word to sentence is proposed. The proposed sentences are evaluated using the data set provided by SigHan 7 bakeoff database [15].

Another investigation into Indian language is the existence of a large number of compound words formed by Euphony and Assimilation. Compound word problems are also treated with caution. For example, in Bangla, there is two stages. First stage is finding the similarity and second stage is take care the error other than similarity [16]. Other Indian language Malayalam, the solution is using a Deep Learning base for spell checker. Which are, identify the misspelled words and the position where the error has occurred based neural network [17].

Research in Indonesia combines spelling mistakes and grammar in Bahasa language. To deal with these two things, statistical rules and methods are used. The spell checker module examines each word using Trie's dictionary to find the misspellings and Damerau-Levenshtein as a correction candidate [18]. Other, use a complete glossary as a reference that consists of five main elements, in particular the pre-processing of documents, the detection of errors in words, the correction of errors in words, the classification of the words of the candidates and the user feedback. HMM is used to choose the most effective candidate for correct words. The experimental results achieved an accuracy of ninety three percent [19].

Malay uses a workbook that contains a combination of words that are usually incorrect and that are written correctly in police work and correct the wrong words. The search creates a separate document retrieval system that can automatically detect and correct misspelled words in Malay without user interaction. The projected approach replaces mechanically. They are classified using the Levenshtein distance to choose the most probable word for the wrong word. Written words that have the best classification will be chosen instead of the wrong word. The designed approach was effective in police work and mechanically corrected the wrong Malay word [20].

In Thailand, the 3-gram methods is selected to be used as a spelling checker for the Tilandese documents. The ArnThai software is OCR software, which is provided by the proposed technical version. The goal of subsequent OCR processing is to correct the OCR result error. It is important to use a spell check tool to correct error mistake and misspelled words [21].

All languages have a way of handling spelling. Among all spelling checking methods, which are the best and are often used in creating applications or systems that can help users, so that they can improve document quality and performance in their respective fields?

II. Methodology

This research conducted a comprehensive review of the research literature on the use of the spelling checker. This process is divided into several parts, which are: determine the sources of research, define the keyword model for the search process, start inclusion and exclusion criteria, extract data and analyze the result to answer a research question.

A. Search Process

The initial process in this research is to determine the literature source to find the right article / journal. In Figure1, The documents in this study only use Google Scholar and Science Direct for search engines. In the keyword column we use Boolean operators (AND and OR) to filter the search data. If there are keywords with two words in them then use AND, but if only one word in the keyword then we use OR. The combinations of the keywords are" Spelling' OR 'Sp ell' OR 'Checker' OR 'Spelling Checker' OR 'Spell Checker'.

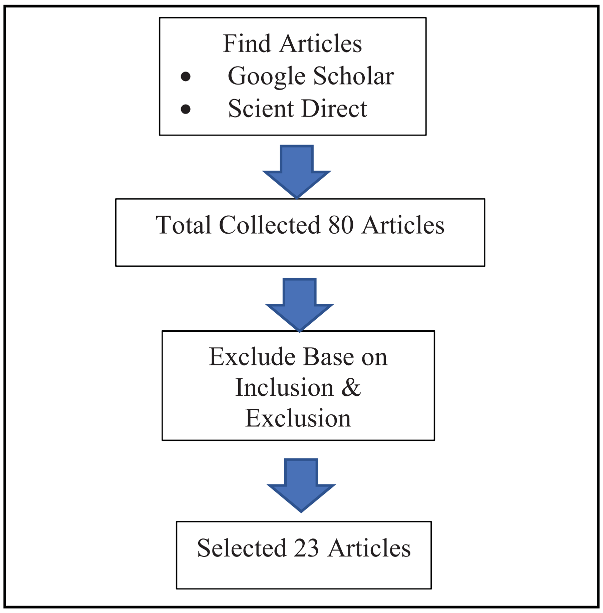
Find Articles
Google Scholar
Scient Direct
Total Collected 80 Articles
Exclude Base on Inclusion & Exclusion
Selected 23 Articles

Fig. 1. Systematic review process.

The second step is to carry out the inclusion criteria from the search mechanism which consists of three stages of the process. The first is the process of searching for articles / journals. All documents that we find are tailored to our original goals, where Google Scholar and Science Direct are different in the search process. Search is adjusted until it reaches the expected number of articles / journals (no more than 100). The paper or article that we collect are from 2000 to 2018. We exclude a duplicate paper and collect 80 paper and article.After that, the next step we include the article by filtering it based on the title and abstract. Additional titles and properties will be defined in the policy definition. For example, we look for the language and country keywords in it. After that, if appropriate, we will document it in the" Candidate Study" status. and the final step, we filter all candidate documents by reading the abstract and conclusions. We use matrix to define the country, language and the methodology that they use, and then save it for selected articles folder for this paper.

B. Data Extractions

The study literature was reviewed 202 articles of all resources and criteria. Of the 202 documents examined, there are 80 articles that are candidate studies based on related titles and summary of the research question. After studying more, there are only 23 articles that can be used in this research (Table 1).

TABLE I. DATA EXTRACTION AND INCLUSION CRITERIA

Language	Paper	*Total*
Africa	[22], [8], [7]	3
Arabic	[12], [9], [11], [10]	4
Chinese	[15], [14]	2
English	[13], [23], [24]	3
India	[25], [17], [26], [16]	4
Indonesia	[19], [18]	2
Japanese	[3], [27]	2
Malaysia	[20], [28]	2
Thailand	[21]	1
	Total	**23**

III. RESULT AND DISCUSSION

This research has intended to investigate about the problem, understanding and solution of Spelling Checker Technology in so many languages. After a search, this paper takes at least one paper that represents each language. Then it is categorized based on the language and technology solutions used. The solutions are summarized in Table 1.

Among the results obtained, research in Arabic and India is the most studied language, there are 4 studies for each. This is because these two languages have many types of languages. The most difficult language for spell checking algorithms is Chinese. First of all, it must be translated into English and therefore the meaning is similar.

A web-based system created to verify spelling. The system is a penguin system. This system is designed to use a database on a website as a tool to help computer users with language tools such as idioms, everyday expressions, names and slang expressions. One interesting thing is the spelling system that is carried out on social media datasets. Which serves to help detect cyberbullying as in the table II below [4]. The system is carried out thoroughly. by applying several methods such as:

1. Adding Character Approach (ACA)
2. Anagram Approach (AA)
3. Dividing Approach (DA)
4. Plural to Singular Approach (PSA)
5. Removing End Characters Approach (RECA)
6. Reordering Approach (RA)
7. Removing Middle Characters Approach (RM|CA)
8. Removing Extra Character Approach (RExCA)
9. Similar Sound Approach (SSoA)
10. Similar Shape Approach (SShA)

The algorithm with Levenshtein is most commonly used in spelling checking. Where Levenshtein transforms a string into another string, where an operation involves addition, deletion, and / or replacement. Frederick Damerau had improved the Levenshtein Algorithm that added the new operation that can verify the distance between the string, ie a transposition of two or more characters. He stated that Levenshtein Algorithm corresponded to over 80% of human spelling. In America, this method is mostly used in the health sector. Levenshtein Algorithm has also use in biology for measure the variation between DNA. Lots of words that are difficult to spell and even difficult to pronounce even though using English. Errors in this field are very fatal, and can even cause death.

TABLE II. MISSPELLING PATTERN IN DATASET [4]

Misspelled Words	Corrected Word
uqly	ugly
Trashhy	Trashy
bitchezz	bitches
f**kin	f**king
siht	shit
slutz	slut
ihateyou	i hate you
p**syyy	p**sy

TABLE III. SPELLING CHECKER METHODS FOR MANY LANGUAGE

Language	Subhead
Africa	1. Word-length algorithm [22] 2. Trie Data Structure [8] 3. eSpellingPro Software [7]
Arabic	1. Gurmukhi & Perso-Arabic Script [12] 2. Recurrent Neural Network (RNN) [9] 3. Levenshtein Algorithm [11] 4. Context words and N-gram [10]
Chinese	1. The Inverted Index List [15] 2. Statistical Machine Translation [14]
English	1. Damerau-Levenshtein Algorithm [13] 2. Intelligent Spelling Correction System SMC [23] 3. Penguin Prototype [24]
India	1. Rule Based approach [25] 2. Deep Learning [17] 3. Convergence to English [26] 4. Phonetic Similarity Error [16]
Indonesia	1. Dictionary, forward reverse & Lexicon resource [19] 2. Statistical and Rule-Based Spelling [18]
Japanese	1. The Idiosyncratic errors [3] 2. Reading Pair Database (RPD) [27]
Malaysia	1. Respelled Word dictionary [20] 2. Malay Language Sentence System [28]
Thailand	1. Statistical Method [21]
Africa	4. Word-length algorithm [22] 5. Trie Data Structure [8] 6. eSpellingPro Software [7]

IV. Implication And Conclusion

This study has two main advantages for theory and practice. As a theory, the results can be a reference for research into the technique of the spelling test method. For practice, these results can be used to identify which methods can be used in a particular language. It is known that every language has a different method of spell checking. The Damerau-Levenshtein algorithm method is most frequently used for spell checkers.Authors and Affiliations

To that end, we believe that many other methods still need to be checked to be used in spelling checking. In addition, there are still many languages in the world so that it becomes an interesting opportunity for IS scholars to further investigate the appropriate methods in each language. In conclusion, the findings of this review study provide preliminary insights into the current trends of the spell-checking method

Based on the spelling checker method identified, there is still much to be added for future research. Journal search engines, though added, such as IEEE, ACM, Springer and others. But for the time being the number of journals is sufficient, and fully represents the facts. Therefore, empirical testing is needed which extensively uses formal statistics to validate these methods.

References

[1] P. Knierim, V. Schwind, A. M. Feit, F. Nieuwenhuizen, and N. Henze, “Physical Keyboards in Virtual Reality: Analysis of Typing Performance and Effects of Avatar Hands,” in *Proceedings of the 2018 CHI Confer- ence on Human Factors in Computing Systems - CHI '18*, 2018.

[2] A. B. Fleischer, “Increasing incidence within PubMed of the use of the misspelling pruritis (sic) instead of pruritus for itch,” *Acta Dermato-Venereologica*, vol. 96, no. 6, pp. 826–827, 2016.

[3] T. Furugori, “Improving Spelling Checkers for Japanese Users of English,” *IEEE Transactions on Professional Communication*, vol. 33, no. 3, pp. 138–142, 1990.

[4] Z. Z. Wint, T. Ducros, and M. Aritsugi, “Spell corrector to social media datasets in message filtering systems,” in *2017 12th International Conference on Digital Information Management, ICDIM 2017*, vol. 2018-Janua, no. Icdim, 2018, pp. 209–215.

[5] M. Octaviano and A. Borra, “A spell checker for a low-resourced and morphologically rich language,” in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2017-Decem, 2017, pp. 1853– 1856.

[6] S. W. Low, “Investment performance analysis of managerial expertise: Evidence from Malaysian-based international equity unit trust funds,” *Jurnal Pengurusan*, vol. 38, no. July 2014, pp. 41–51, 2013.

[7] L. a. Grobbelaar and J. D. M. Kinyua, “A spell checker and corrector for the native South African language, South Sotho,” *Proceedings of the 2009 Annual Conference of the Southern African Computer Lecturers' Association on - SACLA '09*, pp. 50–59, 2009. [Online]. Available: http://doi.acm.org/10.1145/1562741.1562747

[8] H. Naroua and L. Salifou, “On the Computerization of African Languages,” *American Journal of Applied Sciences*, vol. 13, no. 11, pp. 1228–1234, 2016.

[9] N. Madi and H. S. Al-Khalifa, “A Proposed Arabic Grammatical Error Detection Tool Based on Deep Learning,” *Procedia Computer Science*, vol. 142, pp. 352–355, 2018

[10] M. M. Al-Jefri and S. A. Mahmoud, “Context-Sensitive Arabic Spell Checker Using Context Words and N-Gram Language Models,” in Proceedings - 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, NOORIC 2013, 2015, pp. 258–263.

[11] N. Mohammed and Y. Abdellah, “The vocabulary and the morphology in spell checker,” *Procedia Computer Science*, vol. 127, pp. 76–81, 2018.

[12] K. S. Dhanju, G. S. Lehal, T. S. Saini, and A. Kaur, “Design and implementation of shahmukhi spell checker,” *Indian Journal of Science and Technology*, vol. 8, no. 27, 2015.

[13] I. Setiadi, “Damerau-Levenshtein Algorithm and Bayes Theorem for Spell Checker Optimization,” *Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung*, no. December 2013.

[14] H.-w. Chiu, J.-c. Wu, and J. S.Chang, “Chinese Spelling Checker Based on Statistical Machine Translation,” no. October, pp. 49–53, 2013.

[15] W.-Y. Yeh, J.-F, Chen and M.-C. Su, “Chinese Spelling Checker Based on an Inverted Index List with a Rescoring Mechanism spelling checker based on an inverted index list with a rescoring mechanism,” *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, vol. 14, no. 4, 2015.

[16] B. B. Chaudhuri, “Towards Indian language spell-checker design,” in *Proceedings - Language Engineering Conference, LEC 2002*, 2002, pp. 139–146.

[17] S. Sooraj, K. Manjusha, M. Anand Kumar, and K. P. Soman, “Deep learning-based spell checker for Malayalam language,” *Journal of Intelligent and Fuzzy Systems*, vol. 34, no. 3, pp. 1427–1434, 2018.

[18] A. Fahda and A. Purwarianti, “A statistical and rule-based spelling and grammar checker for Indonesian text,” in *Proceedings of 2017 International Conference on Data and Software Engineering, ICoDSE 2017*, 2018.

[19] R. N. Aqsath, M. Kamayani, R. Reinanda, S. Simbolon, M. Y. Soleh, and A. Purwarianti, “Application of document spelling checker for Bahasa Indonesia,” in *International Conference on Advanced Computer Science and Information System (ICACSIS)*, 2011, pp. 249–252.

[20] R. Alfred, S. B. Basri, J. H. Obit, and Z. I. B. A. Ismail, “Improved automatic spell checker for malay blog,” *Advanced Science Letters*, vol. 21, no. 10, pp. 3342–3345, 2015.

[21] S. Watcharabutsarakham, “Spell checker for thai document,” in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2007, 2007, pp. 5–8.

[22] A. Selamat and N. Akosu, “Word-length algorithm for language identification of under-resourced languages,” *Journal of King Saud University - Computer and Information Sciences*, vol. 28, no. 4, pp. 457–469, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jksuci.2014.12.004

[23] S. Sharma and S. Gupta, “A Correction Model for Real-word Errors”, *Procedia Computer Science*, vol. 70, pp. 99–106, 2015.

[24] D. Fallman, “The penguin,” in *CHI '02 extended abstracts on Human factors in computing systems - CHI '02*, 2002, p. 616.

[25] Z. Bhatti, I. A. Ismaili, A. A. Shaikh, and W. Javaid, “Spelling Error Trends and Patterns in Sindhi,” *Journal of Emerging Trends in Computing and Information Sciences*, vol. 3, no. 10, pp. 1435–1439, 2012.

[26] S. Khan, “Convergence in spelling, and spell-checker for Romanized Bangla in computers and mobile phones,” in *2014 International Conference on Informatics, Electronics and Vision, ICIEV 2014*, 2014, pp. 1–5.

[27] P. P. Cai and P. H. Halstead, “Consistency checker for documents containing japanese text,” Jan. 16, 2001, uS Patent 6,175,834.

[28] R. Kasbon, N. Amran, E. Mazlan, and S. Mahamad, “Malay language sentence checker,” *World Appl. Sci. J. (Special Issue on Computer Applications and Knowledge Management)*, vol. 12, pp. 19–25, 2011.