Spell Checker for Somali Language Using Knuth-Morris-Pratt String Matching Algorithm

Ali Olow Jimale[1], Wan Mohd Nazmee Wan Zainon[2(✉)], and Lul Farah Abdullahi[3]

[1] Faculty of Computing, SIMAD University, Mogadishu, Somalia
colowyare2@simad.edu.so
[2] School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia
nazmee@usm.my
[3] Center for Research and Development, SIMAD University, Mogadishu, Somalia
luulf2l@simad.edu.so

Abstract. The Somali language is the mother tongue and official language of Somali people in Somalia, a national language in Djibouti. The Somali language is officially written with the Latin alphabet. Before 1990, the Somali language was the main instruction of the educational institutions in Somalia. However, after the collapse of the Somali central government, foreign languages such as Arabic and English have been taking over as the languages of instruction in Somali educational institutions such as schools and universities. It is feared that the language may die out in about 50 years. One reason for this is a lack of computer tools such as word processors with spell checkers which may promote the use of Somali language. The absence of a spell checker in Somali may lead to spelling mistakes in Somali written documents, including social media posts and friendly chats. The focus of this paper is to propose a spell checker for Somali Language using Knuth-Morris-Pratt String Matching Algorithm with Corpus. A spell checker for the Somali language will provide a word processing interface for writing documents that identifies a misspelled word, underlines it, and suggests the proper spelling of the typed word, if any.

Keywords: Somali language · Spell checker · Knuth-Morris-Pratt algorithm Corpus

1 Introduction

The Somali language, the only language spoken throughout Somalia [1], is the official language of the Federal Republic of Somalia. In 1972, the written system of Somali language was implemented. In the golden period of the Somali language, it became the medium of education, replacing English, Italian and Arabic. In 1980s, Somali based textbooks were made available for schools, with some faculties at Somali National University using Somali as a medium of instruction [2].

F. Saeed et al. (Eds.): IRICT 2018, AISC 843, pp. 249–256, 2019.
https://doi.org/10.1007/978-3-319-99007-1_24

However, starting from 1991 onwards, foreign languages such as Arabic and English took over Somali educational institutions such as schools and universities. [3] stated that the use of the Somali language is in danger of extinction. The language will likely die soon if the Somali people do not participate to save and preserve it from possible of extinction.

There are many reasons for the potential death of Somali language. One of the main reasons is the lack of computer tools that can encourage the use of the language, such as a spell checker.

Spell checking is the process of detecting the misspelled words and giving suggestion candidates to correct misspelled words using computer applications commonly known as spell checkers [4]. A spell checker is considered one of the most important Natural Language Processing applications used to increase the success rate of natural language processing applications, including machine translation, word segmentation, information retrieval and natural language understanding [5]. Most spell checker applications are imbedded with text editors. A prominent example of a spell checker is the Microsoft Word processor spell check. Figure 1 shows an example of Microsoft spell checker. This spell checker allows Microsoft Word processing users to check automatically that their documents are free from spelling and grammar errors.

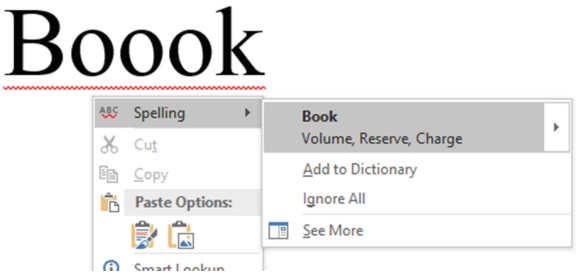

Fig. 1. Example of Microsoft Word spell checking

Spell checkers can be used as a way to promote and maintain the existence of a language. It can help writers to prepare a provisional document in any existing language.

The main advantages of spell checkers are:

1. Checking the accuracy and professionalism of the document.
2. Saving time required for the human to check over the document manually for misspellings.

2 Related Work

This section discusses the literature related to this study. It emphasizes current studies related to spell checking carried by other researchers. Many researchers have tried to solve spelling errors in specific language using different methods and techniques.

An example of spell checking is the work of Mandal and Hossain [6]. These researchers have proposed clustering-based spell-checking technique for Bangla language that can handle both typographical and phonetic errors using Bangla dictionary with clustering algorithm. Their approach was able to reduce both search space and search time. Figure 2 shows the spell correction process of their proposed approach.

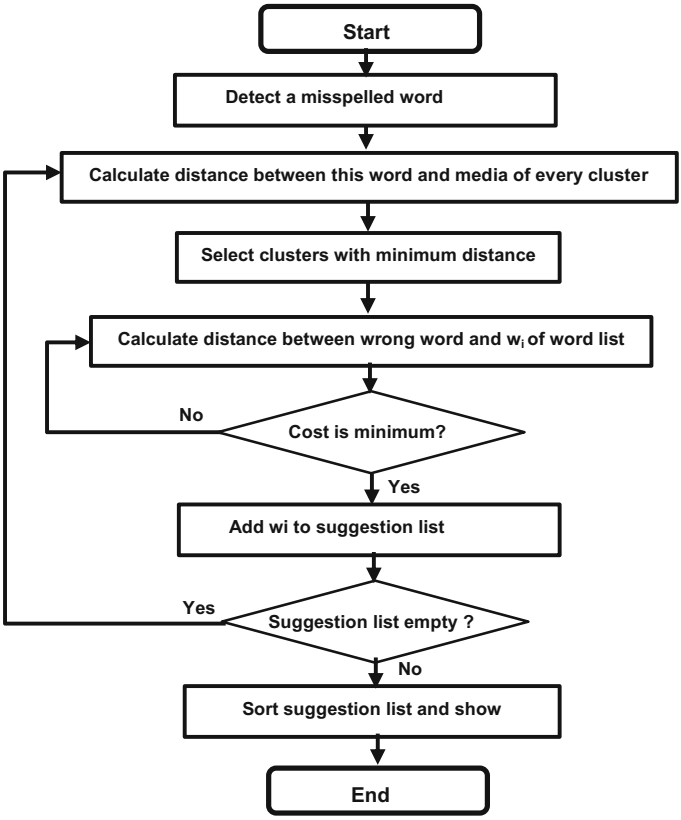

Fig. 2. Clustering-based spell correction process for Bangla language [6]

Another example of a study for solving spelling errors is the work of Basri et al. [7]. They proposed an approach to detect and automatically correct misspelled words in the Malay language without any interaction from the user using lexicon Malay dictionary that contains a list of Malay words. Figure 3 shows their proposed approach.

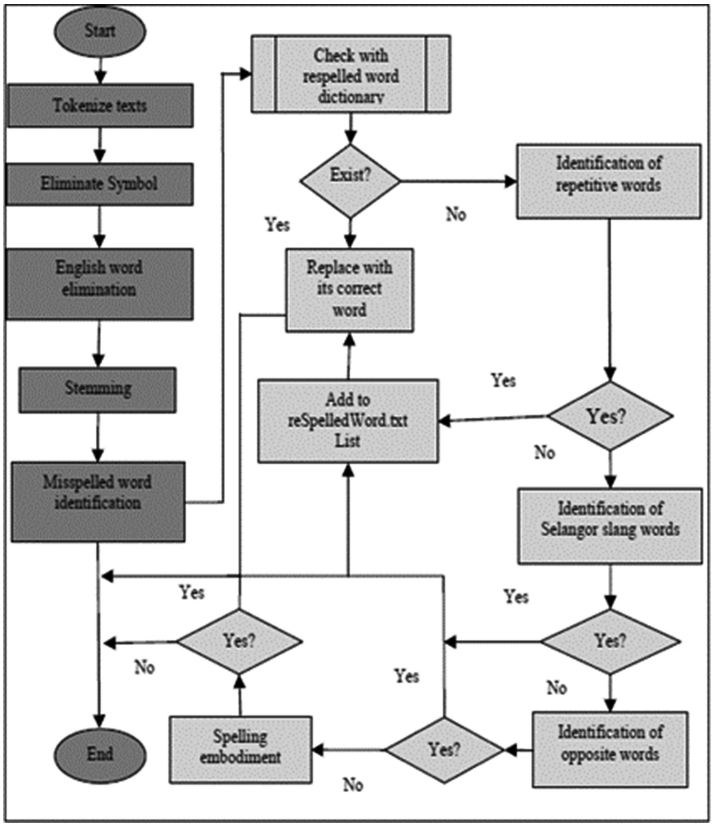

Fig. 3. Malay spell checker proposed approach [7]

In addition to that, a Document Spelling Checker for Bahasa Indonesia was proposed by Kamayani et al. [8]. These authors provide a solution to spelling errors using dictionary. They focused on word error problems. Figure 4 presents their approach.

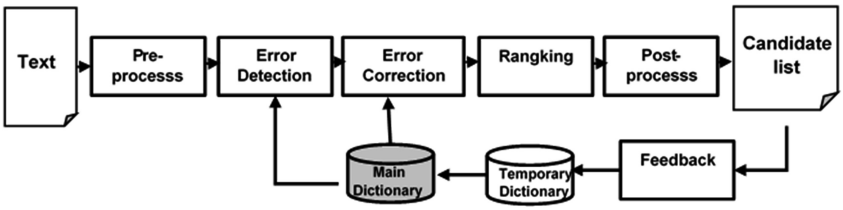

Fig. 4. Document Spelling Checker for Bahasa Indonesia [8]

For Somali language, there have been limited efforts to develop spell checker application. The literature shows the absence of proper word processing application with a spell checker that can detect the spelling errors of Somali language.

3 Proposed Approach

This section discusses the overall techniques used for spell checking and covers in detail the major components of the proposed approach and how these components collectively collaborate each other to detect word errors. Additionally, these techniques offer suggestions to fix the word errors.

The aim of this research is to develop a spell-checking technique that identifies misspelled words of Somali language and gives a proper spelling suggestion. To achieve this objective, a spell-checking approach that can automatically detect misspelled words and give suggestion to do the correction has been proposed. Both the detection and correction of misspelled word can be performed using Knuth-Morris-Pratt String Matching Algorithm with a corpus of Somali words.

The proposed approach of this research takes Somali language as an input text, processes the typed text by the user using Knuth-Morris-Pratt String Matching Algorithm with corpus of Somali words, and generates the detected misspelled word and its correction suggestion as an output.

This consists of three phases: pre-processing, processing, and post processing components. Figure 5 shows the proposed approach. The overall phases of the proposed approach are explained after Fig. 5.

The pre-processing phase is the process of typing words into the proposed approach. The typed words can be misspelled or not. This phase prepares the raw data (Somali language words) to be processed to detect misspelled words and suggest corrections.

Processing phase contains the operations performed to detect and give correction suggestion of the misspelled word. It consists of two sub parts: Knuth-Morris-Pratt String Matching Algorithm and corpus of Somali words. Both subparts collaborate to detect and correct misspelled word.

3.1 Corpus of Somali Words

This component is a document that contains 30,000 Somali words collected from an online dictionary. The dictionary was selected as the corpus of the proposed approach because it is up to date and contains more words than other available Somali dictionaries. The corpus words are used to check the spellings and the occurrence of pattern in the text string. Table 1 shows some of these words.

3.2 Knuth-Morris-Pratt (KMP) Algorithm

KMP is a String Matching algorithm that searches for occurrences of a pattern in a text string. KMP uses information from partial matching of pattern to skip over portions of mismatch and avoid checking characters in text string that we already know matches a

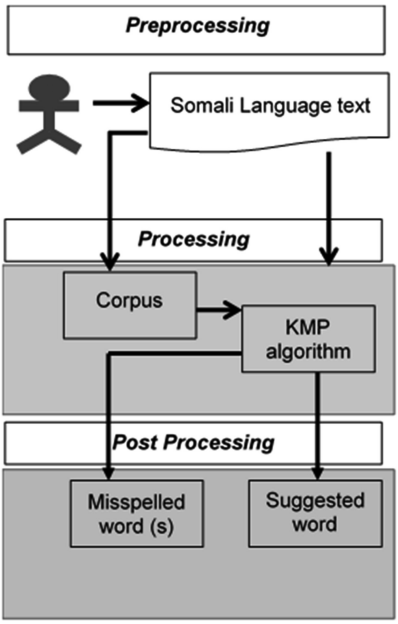

Fig. 5. Proposed approach

Table 1. Example of Somali words

Number	Some of Somali words
1.	*aqoonyahan*
2.	*baafin*
3.	*cabayo*
4.	*daabacaad*
5.	*eeddo*
6.	*falanqayn*
7.	*guuto*
8.	*hoosgunti*
9.	*isir*
10.	*jawaan*
11.	*kalkaaliye*
12.	*laaq*
13.	*macaane*
14.	*nadiifsanaan*
15.	*ogaal*
16.	*qurjujubso*
17.	*raaci*
18.	*shabaq*

prefix of pattern [9]. Assume that our text string, say T, equals 0 0 1 0 0 1 0 0 2 0 0 0 1 0 0 2 0 1 2 2 0 0 and the pattern that we are looking for, say p, is 0 0 1 0 0 2 0 1. Figure 6 shows how the algorithm finds the match of pattern P in the text string T.

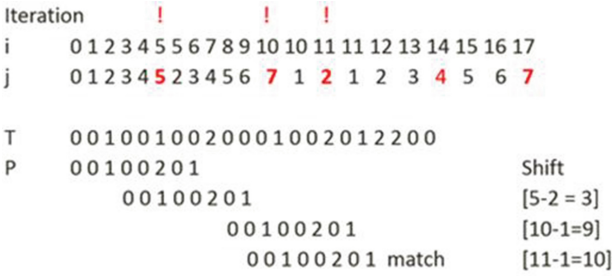

Fig. 6. KMP example

This algorithm is applied in the proposed approach to generate suggestion lists for misspelled Somali words. KMP was used to speed up the pattern matching with Somali words stored in the Corpus to improve the speed of the proposed approach.

4 Spell Checker for Somali Language

The Spell Checker for Somali Language (HINGAAD HUBIYAHA AF-SOOMAALIGA) is an application designed to detect spelling errors in Somali written document. Figure 7 shows an example the interface of this tool. It contains an editor window which allows the users to type or browse a document. After the user types word(s), the application parses the text and checks if the typed word(s) are correctly spelled or not. The checking is carried out

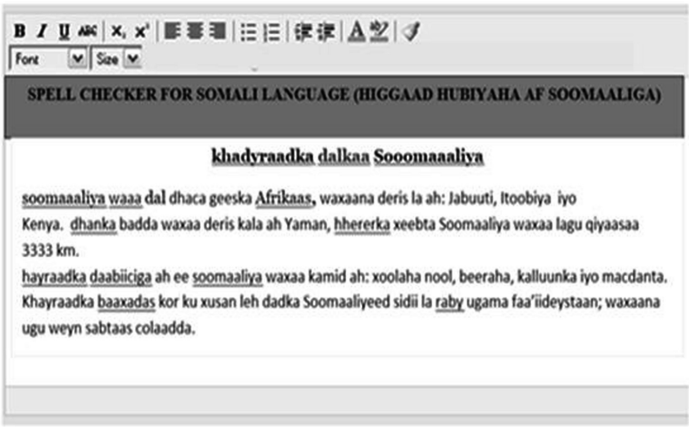

Fig. 7. Spell checker for Somali language

by comparing the words against corpus. If there is no match, it underlines the spelling error and gives suggestion which are similar in structure to the misspelled word.

5 Conclusion and Future Work

In this paper, a spell checker word processing approach for the Somali language was proposed. This method could detect misspelled word(s) and provide suggestion candidates to correct the misspelled words. This study is in the early stage, but the goal is to implement this approach in a functioning text editor computer application to automatically detect misspelled words and suggest the correct words. The proposed application can help the Somali people to ensure the accuracy and professionalism of their typed documents in their native language. This will minimize human spellings errors and reduce the time required to check for spelling errors in a given document.

References

1. Orwin, M.: Somalia: language situation A2. In: Brown, K. (ed.) Encyclopedia of Language and Linguistics, 2nd edn, pp. 509–510. Elsevier, Oxford (2006)
2. Hashi, A.: Developing a model corpus for endangered languages. Graduate thesis, University of Calgary (2014)
3. Patrick, D.: Language endangerment, language rights and indigeneity. Bilingualism: A Social Approach, pp. 111–134. Springer, Berlin (2007)
4. Soleh, M.Y., Purwarianti, A.: A non-word error spell checker for Indonesian using morphologically analyzer and HMM. In: International Conference on Electrical Engineering and Informatics (ICEEI) (2011)
5. Mon, A.M.: Spell checker for Myanmar language. In: International Conference on Information Retrieval and Knowledge Management (CAMP) (2012)
6. Mandal, P., Hossain, B.M.: Clustering-based Bangla spell checker. In: IEEE International Conference on Imaging, Vision and Pattern Recognition (icIVPR) (2017)
7. Basri, S.B., Alfred, R., On, C.K.: Automatic spell checker for Malay blog. In: IEEE International Conference on Control System, Computing and Engineering (ICCSCE) (2012)
8. Kamayani, M., Renainda, R., Simbolon, S.: Application of document spelling checker for Bahasa Indonesia. In: International Conference on Advanced Computer Science and Information System (ICACSIS) (2011)
9. Nonghuloo, M.S., Krishnamurthi, K.: Spell checker for Khasi language. Int. J. Softw. Eng. **7** (1), 1–12 (2017)