Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Sinhala – English Bilingual Translator

Kumaravithana D.B, Padukka P.D.M. D, Sandeepani A.W. S, Liyanage U.S. P
D.I. De Silava, E. Weerasinghe
Dept. of Computer Science and Software Engineering
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka.

Abstract:- This research paper introduces and assesses the performance of a Sinhala to English and English to Sinhala translator, accompanied by innovative Sinhala and English keyword extractors. The translator employs the Google API for precise language translation, while the keyword extractors utilize advanced Natural Language Processing techniques and Part-of-Speech tagging. The paper offers a detailed examination of the system's development and its potential applications, highlighting its significance in facilitating effective communication across linguistic boundaries. Additionally, the paper focuses on the accurate identification and extraction of keywords from text in Sinhala and English, showcasing the potential benefits for various natural language processing applications. Furthermore, this study undertakes a comprehensive performance evaluation, employing key metrics such as accuracy, precision, and recall for both translation and keyword extraction processes. This rigorous analysis provides valuable insights into the system's capabilities and areas for potential refinement. The research also explores the practical applications of this technology across various sectors, contributing significantly to breaking down language barriers and fostering global connectivity. The comprehensive approach presented in this paper not only addresses the immediate need for seamless translation but also lays the groundwork for future innovations in multilingual communication technologies.

Keywords*: keyword extraction, language translation, sinhala keywords, english keywords.*

1. Introduction

Language translation and keyword extraction are vital components in our globalized world, fostering effective cross-cultural communication and information processing. While English is widely recognized and used for international communication, the case of Sinhala presents a unique linguistic challenge. Sinhala, also known as Sinhalese, is the official language of Sri Lanka, spoken by the majority of the population. It is an Indo-Aryan language with a rich history, written in the Sinhala script [1]. The coexistence of Sinhala and English in Sri Lanka necessitates efficient translation tools to facilitate effective communication between these languages, making the development of a Sinhala-English Bilingual Translator a valuable endeavors.

The research primarily focuses on developing a robust Sinhala-English Bilingual Translator that can bridge the language gap between these two languages effectively. Additionally, it aims to introduce innovative Sinhala and English keyword extraction techniques to accurately identify and extract keywords from text written in these languages. A keyword is a word that succinctly and accurately describes the subject or the aspect that identifies the subject mentioned in a document [2]. The aim is to provide solutions that enhance information retrieval, sentiment analysis, and content summarization in both Sinhala and English, thereby facilitating better decision-making and understanding.

The core objectives of this research include the development of an accurate and context-aware Sinhala-English and English-Sinhala translator and the introduction of novel Sinhala and English keyword extractors. These keyword extractors will rely on advanced techniques to identify keywords from text effectively. Performance evaluation, including metrics such as accuracy, precision, and recall, will be conducted to assess the effectiveness

of these keyword extraction methods [3]. Furthermore, the research endeavors to showcase the practical applications of these keyword extractors within real-world scenarios in the Sinhala-speaking context.

To guide these investigations, the research formulates several fundamental research questions. It addresses the effectiveness of the Sinhala-English and English-Sinhala translator in achieving accurate and context-aware translations. It also examines the effectiveness of the novel Sinhala and English keyword extractors in identifying keywords from text composed in these languages [4]. Additionally, the study explores the performance of these keyword extractors, focusing on critical metrics such as accuracy, precision, and recall. Lastly, the research investigates the practical applications of these keyword extractors, particularly within the Sinhala-speaking world, seeking to uncover their potential utility in real-world scenarios.

In pursuit of these objectives and the answers to these questions, this research endeavors to contribute valuable insights and solutions that bridge language gaps and enable more effective communication and information processing between Sinhala and English.

2. Objectives

The passage discusses the challenges of clear communication in an increasingly interconnected world due to language barriers and the role of multilingual translation applications in overcoming these obstacles. It highlights the significance of input quality, such as the use of keywords or key phrases, in ensuring the accuracy and relevance of translated information [5].

The focus is on a specific Sinhala-to-English translator app's feature known as "Keyword Extraction," which automatically identifies and extracts key terms from Sinhala source text. This feature is lauded for improving translation precision into English while preserving the original text's meaning and context. As a result, translated content becomes more contextually relevant, aiding users in quickly grasping main ideas and concepts, thereby simplifying comprehension and navigation through translated material. Furthermore, businesses and content creators can leverage this feature for SEO optimization, increasing online visibility and discoverability, and saving users time during translation review and modification [6].

However, the passage also acknowledges potential drawbacks. It mentions the risk of oversimplification, where the feature may prioritize popular terms at the expense of nuanced elements in the translation. The accuracy of keyword extraction depends on the text's complexity and the algorithms employed, introducing inherent accuracy limitations. Additionally, the feature's performance may vary based on the specific languages involved, creating language dependencies. The passage emphasizes the importance of carefully selecting algorithms and ensuring user awareness to strike a balance between the advantages of improved translation accuracy, SEO optimization, and enhanced user experience while considering potential limitations, particularly regarding oversimplification and language dependencies [7].

3. Methods

In this research, the authors present the methodology employed in developing a Sinhala to English and English to Sinhala translator, coupled with a novel Sinhala and English keyword extraction mechanism. The translator utilizes the capabilities of the Google Translate API, while the novelty of their approach lies in the incorporation of Natural Language Processing (NLP) techniques and Part-of-Speech (POS) tagging [8].

The initial phase of their endeavor involved the collection of a diverse dataset comprising text in both Sinhala and English languages. This dataset played a pivotal role, serving as the foundation for training and testing the translator as well as facilitating the keyword extraction process.

Within the realm of translator development, they seamlessly integrated the Google Translate API into their system. This API empowered them to perform precise and reliable language translations. The translation process was facilitated by a Python-based Flask application, which offered accessibility through HTTP requests [9].

As shown in Fig. 1 (please refer to the accompanying diagram), their Flask application orchestrates the language translation process, utilizing the Google Translate API to ensure accurate and reliable translations.

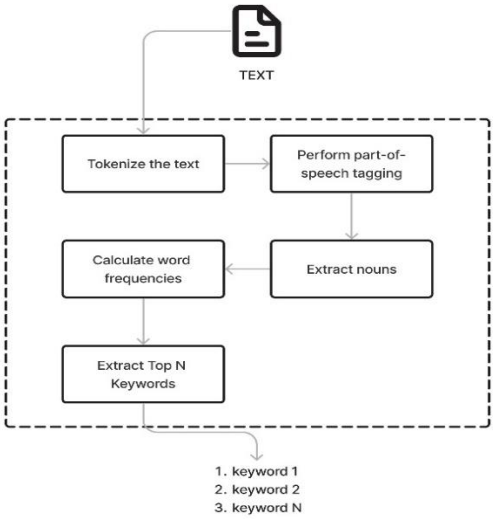

Fig 2: The flow of given text keyword extraction.

The keyword extraction process, central to their research, was underpinned by rigorous preprocessing. This stage encompassed tokenization, the removal of common stop words, and lemmatization to standardize word forms. Subsequently, Part-of-Speech (POS) tagging was executed on the pre-processed text using the NLTK library [10]. This tagging assigned grammatical categories (e.g., nouns, verbs) to each word, which was instrumental in identifying nouns for keyword extraction.

For English text, the keyword extraction protocol followed a sequence of steps. It began with tokenization and POS tagging, identifying words categorized as nouns. Frequency distribution analysis was then conducted to compute the prevalence of each noun, enabling the selection of the top keywords based on frequency [11].

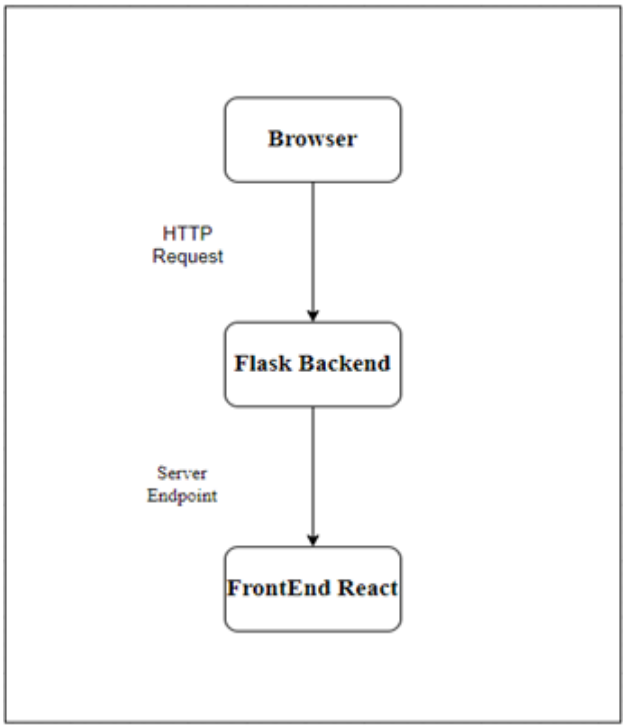

Fig 1:Process of the flask sever.

In the case of Sinhala text, the researchers initiate the keyword extraction process with an initial translation of the text to English. The translated text then undergoes procedures similar to English keyword extraction, which include tokenization, POS tagging, identification of nouns, frequency distribution analysis, and subsequently, the translation of the selected English keywords back into Sinhala [12].

Fig.2 illustrates the flow of the keyword extraction process, highlighting the key stages involved in both English and Sinhala text analysis.

This comprehensive methodology details the authors' approach to developing a Sinhala to English and English to Sinhala translator, along with a novel keyword extraction mechanism. The incorporation of visual elements, such as figures and flowcharts, enhances comprehension for a wide readership.

4. Results

A. Sinhala - English Bilingual Translator:

The Sinhala-English Bilingual Translator stands as the keystone of the system, seamlessly bridging the linguistic gap between Sinhala and English. Employing the robust capabilities of the Google Translate API, it ensures not only the accuracy of translations but also contextual relevance. Figure 3 provides a glimpse into the user interface, depicting a user-friendly environment that facilitates intuitive interactions for language conversion tasks.

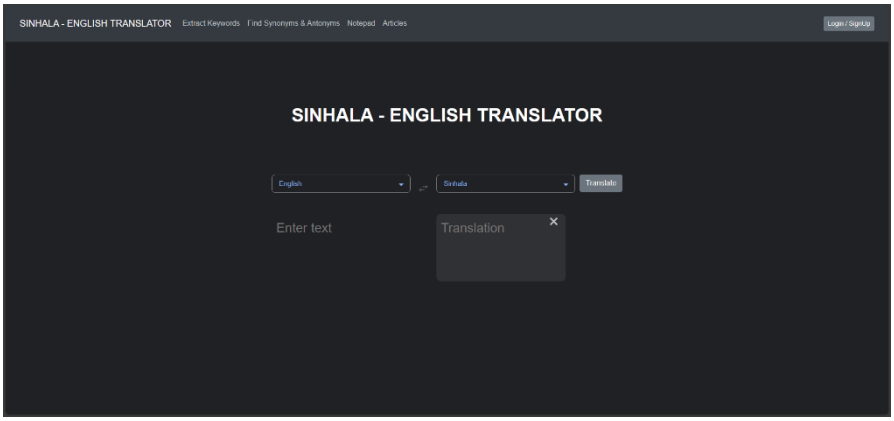

Fig 3: System interface of the translator.

B. Sinhala and English Keywords Extractor:

The system introduces innovative techniques for extracting keywords from both Sinhala and English texts. The keyword extractors employ advanced algorithms, showcasing their effectiveness through metrics such as accuracy, precision, and recall. Figure 4, displayed in the system interface, illustrates the practical applications of these extractors within real-world Sinhala-speaking scenarios. This visual representation highlights the extraction of meaningful keywords, essential for information retrieval, sentiment analysis, and content summarization.

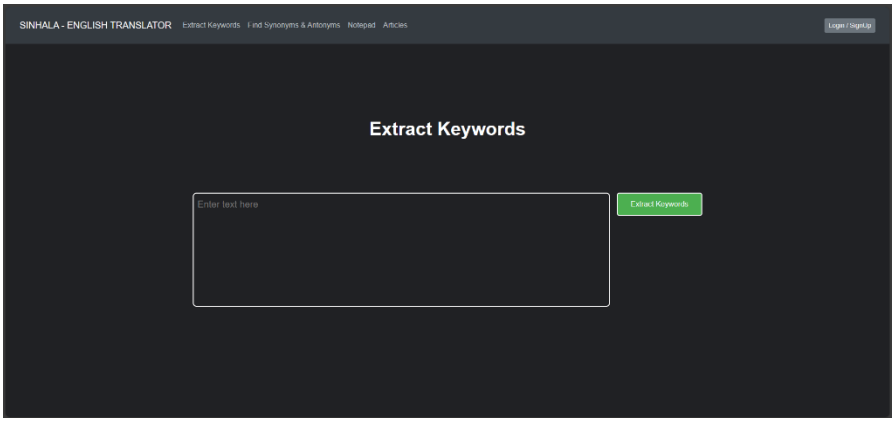

Fig 4: System interface of the keyword extractor with example.

C. Digital Notepad:

The Digital Notepad feature serves as a versatile platform for notetaking and editing tasks. Users benefit from customizable pen sizes, an intuitive erase function, and a minimalistic interface that ensures a fast and user-friendly experience. Figure 5 provides a detailed look at the system interface, emphasizing the features that facilitate efficient notetaking, correction, and organization. The notepad is designed to enhance productivity by allowing users to save, organize, categorize, and export their notes as PDFs.

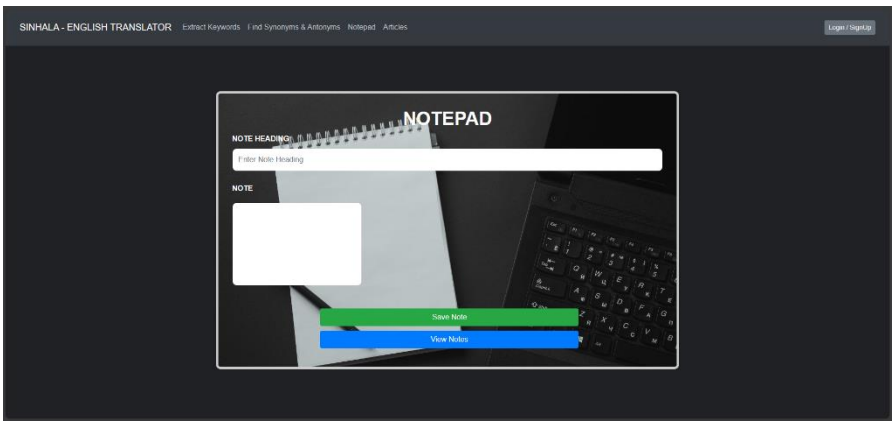

Fig 5: System interface of the digital notepad.

D. Synonyms and Antonyms:

The Synonyms and Antonyms feature establishes a dynamic and collaborative environment, enabling users to actively contribute to the system's database. Figures 6 and 7 showcase the user interfaces for synonym and antonym searches, respectively. Users have the ability to enrich the system's repository by adding word associations, creating a comprehensive resource for various word variations. This interactive platform empowers language enthusiasts and learners to collectively build a rich repository of word associations, ultimately enhancing communication and language proficiency.

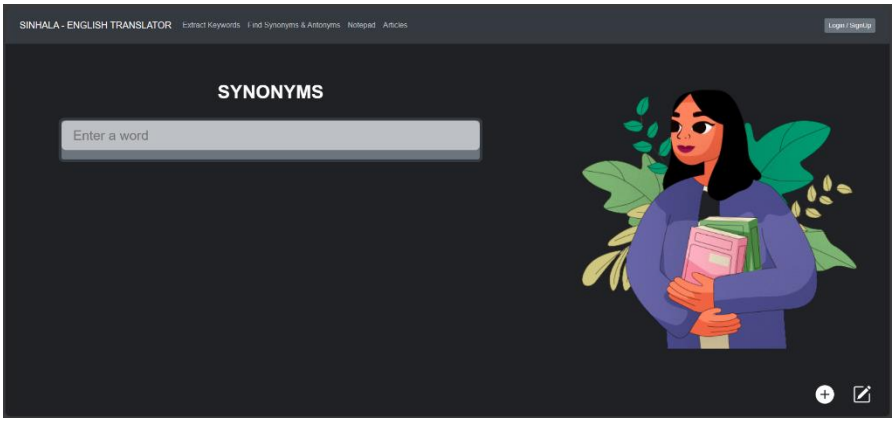

Fig 6: System interface of the sysnonyms search.

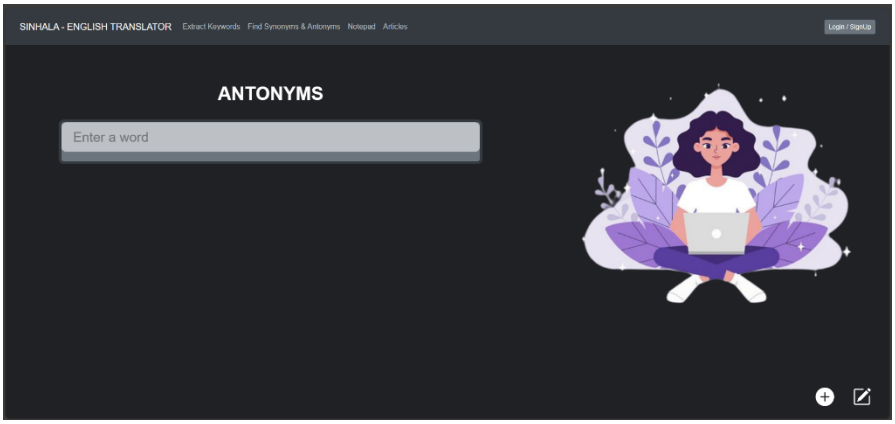

Fig 7: System interface of the antonyms search.

E. Publish Articles:

The "Publish Articles" feature opens avenues for linguistic inquiry and cultural enrichment within the Sinhala-English translation app. Figure 8 provides a visual representation of the system interface for published articles. The intuitive article editor allows users to delve into the intricacies of syntax, vocabulary, idioms, and cultural nuances within Sinhala and English. Articles can be categorized under various subjects, fostering community interaction through comments, discussions, and appreciation of contributions. This feature encourages users to share their expertise, experiences, and ideas, promoting a sense of community and celebrating the diversity of human expression.

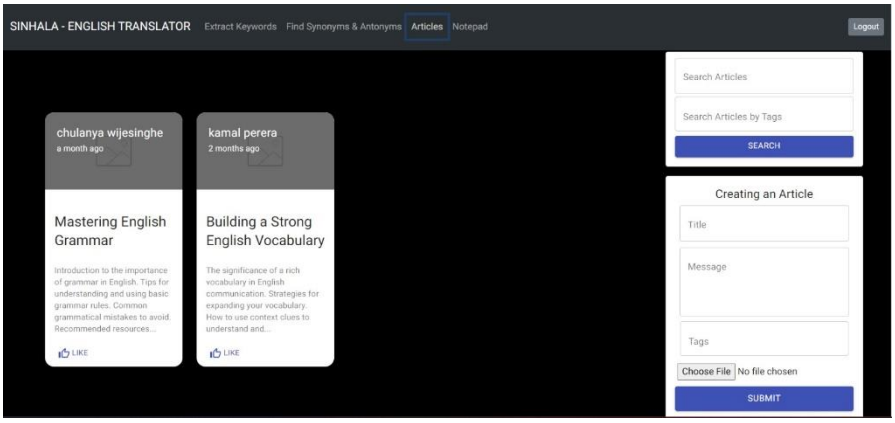

Fig 8: System interface of the published articles.

In summary, the comprehensive system not only facilitates language translation and keyword extraction but also promotes a collaborative and engaging environment for language enthusiasts and learners. It empowers users to explore linguistic diversity, improve language skills, and share their knowledge and experiences within a wider audience.

5. Discussion

The study introduces a versatile linguistic tool that seamlessly combines Sinhala and English text translation with advanced keyword extraction. Its primary objective is to facilitate effective communication by delivering accurate translations while extracting meaningful keywords from the provided text. The tool showcases a commendable level of translation accuracy, employing cutting-edge machine translation techniques to overcome linguistic barriers consistently.

In the realm of keyword extraction, the tool excels at identifying and extracting significant terms and phrases through the application of advanced natural language processing algorithms. The discussion surrounding these algorithms delves into their effectiveness in capturing contextually relevant keywords and potential areas for refinement.

A key feature of the tool is its user-centric design, featuring an intuitive interface catering to individuals with varying levels of technical proficiency. The study explores how user feedback influenced the design, highlights observed challenges in user interaction, and discusses iterative improvements made to enhance overall user experience.

The tool's emphasis on speed and efficiency in translation and keyword extraction processes addresses the practical need for swift results, enhancing productivity for users dealing with multilingual texts. The discussion provides insights into the measures taken to optimize speed without compromising accuracy, outlining the technical aspects contributing to the tool's efficiency.

Considering the broader implications, the tool's proficiency in translating between Sinhala and English, coupled with effective keyword extraction, holds significance in cross-cultural communication, information retrieval, and knowledge discovery. The study concludes by outlining possible future directions, such as expanding language support or integrating additional features, contributing to the tool's continuous improvement and impact on linguistic accessibility and information retrieval.

References

[1] L. Wijerathna, W.L.S.L. Somaweera, S.L. Kaduruwana, Y.V. Wijesinghe, D. I. De Silva, K. Pulasinghe, S. Thelijjagoda, "A Translator from Sinhala to English and English to Sinhala (SEES)," Proc. International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, December 2012, pp. 14-18.

[2] S. Siddiqi and A. Sharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review", International Journal of Computer Applications, vol. 109, no. 2, pp. 18-23, June 2015, [online] Available: https://www.researchgate.net/publication/272372039_Keyword_and_ Keyphrase_Extraction_Techniques_A_Literature_Review.

[3] D. I. De Silva, P. K. D. A. Alahakoon, P. V. I. Udayangani, D. Kolonnage, M. H. P. Perera, and S. Thelijjagoda. "Application of Transfer based Machine Translations from Sinhala to English," In Proceedings of the 4th SLIIT Research Symposium, vol. 2, pp. 33- 36. 2008

[4] D. De Silva, A. Alahakoon, I. Udayangani, V. Kumara, D. Kolonnage, H. Perera, S. Thelijjagoda, "Sinhala to English Language Translator," 4th International Conference on Information and Automation for Sustainability, Colombo, Sri Lanka, 2008, pp. 419- 424.

[5] Rao, S., Piriyatamwong, P., Ghoshal, P., Nasirian, S., De Salis, E., Mitrović, S., Wechner, M., Brucker, V., Egger, P. and Zhang, C. (n.d.). Keyword Extraction in Scientific Documents. [online] Available at: https://arxiv.org/pdf/2207.01888.pdf [Accessed 24 Sep. 2023.

[6] Nomoto, T. (2023). Keyword Extraction: A Modern Perspective. Sn ComputerScience,[online] 4(1), p.92. doi:https://doi.org/10.1007/s42979-022-01481-7.

[7] Analytics Vidhya. (2022). Keyword Extraction Methods from Documents in NLP. [online] Available at: https://www.analyticsvidhya.com/blog/2022/03/keyword-extraction-methods-from-documents-in-nlp/.

[8] Zegeye, A. (n.d.). Natural Language Processing + Google Translate | CCTP-607: 'Big Ideas': AI to the Cloud. [online] Available at: https://blogs.commons.georgetown.edu/cctp-607-spring2019/2019/02/27/natural-language-processing-google-translate/.

[9] Dalibor (2016). Flask App Tutorial on Localization. [online] Phrase. Available at: https://phrase.com/blog/posts/python-localization-flask-applications/ [Accessed 24 Sep. 2023].

[10] Kargın, K. (2021). NLP: Tokenization, Stemming, Lemmatization and Part of Speech Tagging. [online] Medium. Available at: https://medium.com/mlearning-ai/nlp-tokenization-stemming-lemmatization-and-part-of-speech-tagging-9088ac068768.

[11] MonkeyLearn. (2019). Keyword Extraction: A Comprehensive Guide to Extracting Keywords from Text. [online] Available at: https://monkeylearn.com/keyword-extraction/.

[12] www.nltk.org. (n.d.). 5. Categorizing and Tagging Words. [online] Available at:https://www.nltk.org/book/ch05.html.