A Review of Natural Language Processing Tools for Sinhala Language

Conference Paper · November 2024

CITATIONS
0

READS
46

1 author:

Piyumi Weerarathna
University of Kelaniya
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

A Review of Natural Language Processing Tools for Sinhala Language

Piyumi Weerarathna(CS/2020/024), Imesha Kularathna(CS/2020/046)
Department of Computer Systems Engineering, University of Kelaniya, Sri Lanka.
nadeesh-cs20024@stu.kln.ac.lk
kularat-cs20046@stu.kln.ac.lk

***Abstract*—Tools for natural language processing (NLP) are essential for processing languages with limited resources, such as Tamil and Sinhala. The current state of NLP tools, their features, and their drawbacks are examined in this research. This review finds important resources including machine translation models and text categorization methods, part-of-speech (POS) taggers and lexical resources by examining several papers. Potential areas for future improvement are highlighted, along with obstacles including the lack of annotated datasets and computational resources. This study emphasizes how researchers must work to enhance NLP for Sinhala.**

Keywords - NLP, Sinhala, Wordnet, Text Classification, Language Corpus, POS Tagging, SAFS3 Algorithm, Hidden Markov Models, Sinmin Corpus, Annotated Corpora.

1. INTRODUCTION

One of the most important areas of AI for enhancing human-computer interaction is natural language processing or NLP. Because of their intricate morphological patterns and scarcity of resources, Sinhala languages provide particular difficulties [1] for NLP development. This evaluation of the literature assesses the NLP methods and tools now in use for these languages, pointing out any shortcomings and recommending future lines of inquiry. Fostering digital inclusion and protecting linguistic heritage require cooperative and scalable solutions.

2. METHODOLOGY

Using scholarly publications and datasets from sources such as IEEE Xplore and Google Scholar, resources for Sinhalal. Performance measures, usability and adaptability were used to evaluate important this review methodically assesses NLP tools and resources, including the SAFS3 Algorithm, Wordnet, the HMM-based POS Tagger and the Sinmin Corpus. Studies that addressed issues like managing complicated morphology and the lack of annotated datasets were chosen for their applicability and contributions to NLP. A fair evaluation that identified these instruments' advantages, disadvantages and potential future paths was guaranteed by the use of both quantitative and qualitative methodologies.

3. LITERAURE REVIEW

3.1 WordNets

For the advancement of NLP tasks including information retrieval, machine translation and word sense disambiguation (WSD), a Sinhala WordNet is essential. With their own contributions and difficulties, the three major studies by Arukgoda et al.[2], Welgama et al.[3] and Wijesiri et al[4]. take distinct ways to creating this resource.

Arukgoda et al.[2] use a Sinhala-specific version of the Lesk algorithm to concentrate on WSD. With a 63% accuracy rate, their rule-based approach resolves semantic ambiguity by utilizing context and gloss overlaps. Not with standing its achievements the study emphasizes the necessity of sense-annotated corpora and a thorough morphological analyzer. Welgama et al.'s work, which takes a data-driven approach and uses a frequency-based selection of synsets from the UCSC Sinhala corpus, echoes this problem. Although their staged approach stresses manual sense identification, it has trouble managing intricate morphological features and cultural quirks.

Wijesiri et al.[4], on the other hand, use a NoSQL database for scalability and employ a crowdsourced technique to construct a Sinhala WordNet. By involving the community and enabling contributors to add synsets and relationships via a web-based interface their method tackles the problem of resource scarcity. Although encouraging, it is still difficult to maintain quality control in crowdsourced data.

Three studies emphasize the need to adapt WordNet structures to Sinhala's linguistic characteristics, including morphological inflections and spoken/written forms. A unified approach combining rule-based WSD, data-driven synset selection, and community-driven lexicon expansion could create a robust Sinhala WordNet.

3.2 Sinhala Text Classification Using SAFS3 Algorithm.

The study "Sinhala Text Classification Using SAFS3 Algorithm" investigates novel approaches to improve text classification for the morphologically complex but resource-poor Sinhala language. The work, carried out by academics at the University of Moratuwa, presents the SAFS3 Algorithm, which enhances feature representation and classification performance by combining WordNet-based semantic similarity measures with TF-IDF (Term Frequency-Inverse Document Frequency)[1].

The SAFS3 algorithm tackles the problem of generating significant feature vectors for Sinhala text by fusing semantic enrichment offered by WordNet, specifically using the Wu-Palmer similarity metric, with statistical techniques such as TF-IDF, which gives weights to terms according to their significance. Because of its intricate morphology and rich inflections Sinhala benefits greatly from this hybrid method, which allows for a greater comprehension of the links between words [1].

The SAFS3 algorithm outperformed English models in Sinhala sentence classification with an article classification accuracy of 99.57% and validation accuracy of 75.51%, despite reliance on multilingual dictionaries and lack of contextual understanding resources like Sinhala WordNet. Future research should focus on creating annotated corpora for informal and code-mixed text [1].

3.3. Corpus based Sinhala Lexicon

The creation of an extensive Sinhala lexicon is presented in the publication "Corpus based Sinhala Lexicon" by Ruvan Weerasinghe, Dulip Herath and Viraj Welgama from the Language Technology Research Laboratory, University of Colombo[5]. This lexicon has 35,000 terms and is based on a corpus of 10 million words from a variety of genres, including news reporting, technical writing and creative writing. Notable examples include the use of phonetic transcriptions for voice processing and the usage of the ISO-compliant Lexical Markup Framework . In order to overcome the shortcomings of conventional grammatical techniques, the research also introduced a novel classification for parts of speech, improving the lexicon's usefulness for NLP tasks including machine translation, grammar building and POS tagging [5].

On uncontrolled text, the lexicon showed an average coverage of 80.9%; for cleaned data, this coverage was higher in genres such as creative writing (86.71%). Its versatility is noteworthy since it may be used for tasks ranging from speech synthesis to semantic role labeling in natural language processing. Nonetheless, a lack of thorough verb coverage is highlighted by the very small number of verbs (only 3% of the lexicon). Furthermore, the addition of named entities and functional words improves its usability for a range of applications, including information retrieval and dependency parsing [5].

Future research could concentrate on enhancing verb coverage and adding compound verbs to the lexicon. The lexicon's strength would be increased by adding a morphological analyzer and generator, which would allow it to directly generate every word form. Further work could improve error handling for typographical and spelling errors and handle genre-specific constraints, such as correct noun recognition in news reporting [5].

3.4 Hidden Markov Model Based Part Of Speech Tagger For Sinhala Language

This study used a stochastic model technique based on Hidden Markov Models (HMM) to construct a Part-of-Speech (POS) tagger for the Sinhala language. Important issues were addressed, including the lack of vocabulary resources and the morphological complexity of Sinhala. We used a 90,551-word, 2,754-sentence Sinhala text corpus from newspaper articles to compute observation likelihoods and transition probabilities. To maximize tagging accuracy, the Viterbi algorithm was used. The efficiency of the tagger in handling trained vocabulary was demonstrated by the above 90% accuracy achieved for known words. however, performance on unknown words (out-of-vocabulary terms) remained a significant restriction. Addressing linguistic difficulties and laying the groundwork for future developments in Sinhala NLP resulted in significant contributions. The use of an extensive annotated corpus and a thorough approach that confirmed the viability of stochastic models for morphologically rich languages are two of the study's strong points. However, shortcomings including the incapacity to manage unfamiliar words and intricate linguistic elements, like compound nouns and particles, were noted as areas in need of development. All things considered, the study emphasizes the value of bigger, more varied corpora and the investigation of hybrid strategies to improve tagging precision and flexibility in subsequent research. [6]

3.5 Analysis of Sinhala Using Natural Language Processing Techniques

The paper "Analysis of Sinhala Using Natural Language Processing Techniques" examines basic

NLP experiments meant to provide a computer knowledge of Sinhala. This was accomplished by gathering and preprocessing a sizable corpus of Sinhala text which included approximately 2.2 million characters and 681,233 word tokens. The use of Maximum Likelihood Estimates (MLE) for character frequency analysis Naïve Bayes for language recognition (which achieved 100% accuracy) and an investigation of Zipf's law behavior which showed that stop words were more common were among the key experiments. Support Vector Machines (SVM) displayed over 90% accuracy in topic categorization while n-gram language models showed decreasing perplexity as model complexity grew indicating potential uses in language modeling. The construction of a sizable Sinhala corpus and the application of various NLP algorithms which resulted in excellent accuracy in tasks like language identification and subject classification are the paper's strong points. Nevertheless drawbacks were noted such as possible overfitting in n-gram models and a lack of investigation into intricate linguistic elements like compound words. Notwithstanding these difficulties the study provides a solid basis for Sinhala natural language processing (NLP) highlighting the necessity of more extensive and varied corpora as well as sophisticated tools for applications like speech recognition and machine translation. It also identifies areas where future research can investigate deep learning methods and hybrid approaches.[7]

4. Discussion and Comparison

The discipline of Sinhala Natural Language Processing (NLP) has made great strides in tackling the morphological complexity and resource constraints of the language. With a hybrid strategy that combines statistical techniques and semantic enrichment, the SAFS3 algorithm obtained an article classification accuracy of 99.57%[1]. However, because there isn't a specific Sinhala WordNet, it has restrictions. There is little verb representation in the corpus-based Sinhala Lexicon project[5], a lexicon with 80.9% coverage. Although it achieves over 90% accuracy for known words, the Hidden Markov Model-based POS tagger has trouble with compound linguistic structures and unknown terms[6]. A dynamic resource that facilitates query-based analysis, the Sinmin Sinhala Corpus may provide scalability issues. In order to fill these deficiencies and enable more extensive applications in speech recognition, machine translation and semantic analysis, future research should concentrate on creating scalable resources and models.

5. Conclusion

With an emphasis on text categorization, lexical resource generation, part-of-speech tagging and foundational analysis, the literature review demonstrates advancements in Sinhala natural language processing techniques. But issues like resource limitations and morphological complexity require constant improvement. Although resources like the Sinmin Corpus and the corpus-based Sinhala Lexicon offer a strong basis, their shortcomings draw attention to the need for more study. The absence of annotated corpora, a Sinhala WordNet and sophisticated morphological analyzers are common deficiencies in Sinhala natural language processing. Future work should concentrate on cooperation, integrating linguistic rules and providing scalable resources.

6. References

[1] N. de Silva, "Sinhala Text Classification: Observations from the Perspective of a Resource Poor Language".

[2] J. Arukgoda, V. Bandara, S. Bashani, V. Gamage, and D. Wimalasuriya, "A Word Sense Disambiguation Technique for Sinhala," in *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, Kota Kinabalu, Malaysia: IEEE, Dec. 2014, pp. 207–211. doi: 10.1109/ICAIET.2014.42.

[3] V. Welgama, D. L. Herath, C. Liyanage, N. Udalamatta, R. Weerasinghe, and T. Jayawardhane, "Towards a Sinhala Wordnet".

[4] I. Wijesiri *et al.*, "Building a WordNet for Sinhala".

[5] R. Weerasinghe, D. Herath, and V. Welgama, "Corpus-based Sinhala lexicon," in *Proceedings of the 7th Workshop on Asian Language Resources - ALR7*, Suntec, Singapore: Association for Computational Linguistics, 2009, pp. 17–23. doi: 10.3115/1690299.1690302.

[6] A. J. P. M. P. Jayaweera and N. G. J. Dias, "Hidden Markov Model Based Part of Speech Tagger for Sinhala Language," *Int. J. Nat. Lang. Comput.*, vol. 3, no. 3, pp. 9–23, Jun. 2014, doi: 10.5121/ijnlc.2014.3302.

[7] S. Gallege, "NLP Analysis of Sinhala".