Available online at www.jilile.in ISSN: 3048-4952

JOURNAL OF INDIAN LANGUAGES
AND INDIAN LITERATURE IN ENGLISH

Journal of Indian languages and Indian literature in English, 03(02), 341-348; 2025

STEMMING IN LANGUAGE COMPUTING: ADDRESSING MALAYALAM-ENGLISH MACHINE TRANSLATION CHALLENGES

Lizbeth James,
Postgraduate student,
Department of Linguistics,
University of Kerala,
Thiruvananthapuram, India,
lizbethjames3@gmail.com,
9961995126.

Dr. Darwin L,
Head of the Department,
Department of Linguistics,
University of Kerala,
Thiruvananthapuram, India,
darwin@keralauniversity.ac.in,
9446904488

APA Citation:
James, L., & Darwin, L. (2025). Stemming in Language Computing: Addressing Malayalam-English Machine Translation Challenges, Journal of Indian Languages and Indian literature in English, 03(02,341-348; 2025

Submission Date: 19.03.2025
Acceptance Date: 24.03.2025

ABSTRACT:

Stemming, a fundamental process in language computing, plays a critical role in text processing, search optimization, and machine translation by reducing words to their root forms. However, in morphologically rich languages like Malayalam, stemming presents unique challenges, particularly in adjectival compound words where meaning is highly dependent on collocation and co-occurrence patterns. For instance, pacha meen (‘fresh fish’) and pacha manushyan (‘naive man’) demonstrate how the adjective pacha shifts its semantic role based on the noun it modifies. Similarly, pacha vellam (‘normal water’) is not ‘green water,’ and pacha kuthira (‘grasshopper’) does not translate as ‘green horse.’ These variations illustrate the limitations of direct word-for-word translation and emphasize the necessity for context-sensitive processing in machine translation. Eugene Nida’s sixth morphological principle—Distribution—highlights that morphemes adhere to specific combinatory patterns to form meaningful words, making it crucial for translation systems to recognize fixed patterns in compound structures. Traditional stemming methods that indiscriminately strip affixes often fail in this regard, leading to semantic distortions. To address this, advanced stemming techniques incorporating syntactic and semantic analysis must be employed, ensuring accurate preservation of intended meanings in Malayalam-English translation. By integrating rule-based morphological analysis with deep learning models, translation accuracy can be significantly enhanced, minimizing errors caused by rigid stemming approaches. The study adopts a morphological stemming process as its methodology, focusing on refining machine translation models to handle complex compound structures

Keywords- Co-occurrence, Collocation, Stemming, Compound forms, Adjective compounds.

1.Introduction

Stemming is a crucial process in natural language processing (NLP), aiding in text processing, search optimization, and machine translation by reducing words to their base forms. However, in morphologically rich languages like Malayalam, stemming presents unique challenges, particularly in adjectival compound words where meaning is highly dependent on collocation and co-occurrence patterns. For example, in Malayalam, the adjective pacha (പച്ച) [pʌʧʃːɐ] can mean "green," "fresh," "true," or "normal," depending on the noun it modifies. While pacha meen (പച്ചമീൻ) [pʌʧʃːɐ miːn] means "fresh fish," pacha manushyan (പച്ചമനുഷ്യൻ)[pʌʧʃːɐ mɐɳ uʂːɪjɐn] translates to "naive person," not "green man." Such variations illustrate the limitations

of direct word-for-word translation and emphasize the necessity for context- sensitive processing in machine translation. This paper explores these challenges by integrating rule-based morphological analysis with deep learning models to refine machine translation accuracy. It reviews existing literature on morphological processing and machine translation challenges in Indian languages while addressing key research questions.

1.1 Literature Review

Sreelekha & Bhattacharyya (2018) explore how morphology plays a crucial role in enhancing statistical machine translation (SMT) between English and Malayalam. The research highlights that direct stemming often results in the loss of semantic information, making it difficult to retain the correct meaning of words. To address this, the authors propose a morphology-sensitive approach that considers Malayalam's rich inflectional and derivational morphology, improving translation accuracy.

Patel (2017) discusses the challenges faced in machine translation (MT) for Indian languages, with a specific focus on their complex morphological structures. The study identifies key issues such as polysemy, agglutination, and non-standardized linguistic resources that hinder accurate translation. To overcome these obstacles, the authors suggest incorporating deep learning models with rule-based morphology processing to enhance translation quality.

Nida's Principle of Distribution (1949) states that morphemes acquire meaning based on their arrangement and context within a sentence. This principle is especially relevant in Malayalam, where adjectives and compound words often shift meaning depending on their syntactic position. Traditional stemming techniques, which treat morphemes as static units, fail to account for these contextual variations, leading to inaccurate translations in machine translation systems.

1.2 Research Questions

This study seeks to address the following research question:

a) How do Malayalam adjectival compound words challenge traditional stemming techniques in machine translation?
b) What role does context-sensitive processing play in improving Malayalam-English machine translation accuracy?

c) How can rule-based morphological analysis be integrated with deep learning models to enhance machine translation quality?

1. Stemming Challenges in Malayalam Adjectival Compounds

Malayalam adjectival compounds pose significant challenges for machine translation because their meanings often change depending on the words they modify. These challenges fall into three main categories: context-dependent meaning shifts, non-literal translations, and noun-based compounds with altered semantics.

1.1. Context-Dependent Meaning Shifts in Adjectival Compounds

Example 1: പച്ചവെള്ളം (pacha vellam) [pʌʧʃːʁveɭːʁm]

- Incorrect translation: Green water
- Correct meaning: Normal water
- Glossing:
- [pʌʧʃːʁ] – green
- [veɭːʁm] – water

The adjective *pacha* usually means "green," but in this context, it refers to "normal," highlighting how Malayalam adjectives change meaning based on context. Direct stemming methods fail to capture this distinction, leading to incorrect translations.

Example 2: പച്ചമനുഷ്യൻ (pacha manushyan) [pʌʧʃːʁ mʁn̪ uʂːɪjʁn]

- Incorrect translation: Green man
- Correct meaning: True person
- Glossing:
- [pʌʧʃːʁ] – green
- [mʁn̪uʂːɪjʁn] – human

Here, *pacha* does not refer to color but rather signifies "true" or "honest," illustrating the problem of literal translations in MT systems. Without contextual awareness, translation algorithms fail to recognize such semantic shifts

Example 3: പച്ചമീൻ (pacha meen) [pʌʧʃːʁ miːn]

- Incorrect translation: Green fish
- Correct meaning: Fresh fish

- Glossing:
- [pʌʧʃːɐ] – green
- [miːn] – fish

In Malayalam, *pacha* can also mean "fresh," showing how adjectives carry multiple meanings. A rule-based morphology system integrated with machine learning could help differentiate between such variations.

1.2 Compound Words with Non-Literal Translations

Example 4: ചിന്നവീട് (chinna veedu) [ʧʃiɳ ɳavi:ɖi]

- Incorrect translation: Small house
- Correct meaning: Illegitimate family
- Glossing:
- [ʧʃiɳ ːɐ] – small
- [ʋi:ɖi] – house

This phrase does not refer to an actual house but is a cultural metaphor for an illegitimate family, demonstrating the difficulty in translating idiomatic expressions.

Example 5: വെചിതിന്നുക (chevi thinnuka) [ʧʃevi ʈiɳ ːukɐ]

- Incorrect translation: Eat the ear
- Correct meaning: To whisper excessively
- Glossing:
- [ʧʃeʋi] – ear
- [ʈiɳ ːukɐ] – to eat

The literal translation makes no sense in English, emphasizing the need for context-aware processing in MT systems.

Example 6: ഇടിവെട്ട് (idivett) [iɖiʋeʈːi]

- Incorrect translation: Thunderbolt
- Correct meaning: Eye-catching
- Glossing:
- [iɖi] – thunder
- [ʋeʈːi] – bolt

This word functions as a descriptive adjective rather than a direct reference to a thunderbolt.

1.3 Noun-Based Compounds with Altered Semantics

Example 7: ചേനത്തണ്ടൻ (chenathandan) [ʧeːn̪ ɐt̪ɐn̪ɖɐn]

- Incorrect translation: Sweet potato
- Correct meaning: Name of a snake
- Glossing:
- [ʧeːn̪ ɐ] – elephant foot yam
- [t̪ɐn̪ɖɐn] – stem

The phrase does not describe a vegetable but instead refers to a snake species, showing how noun compounds gain unique meanings.

Example 8: പഴംതുണി (pazham thuni) [pɐɻɐm t̪uɳi]

- Incorrect translation: Fruit cloth
- Correct meaning: Old clothes
- Glossing:
- [pɐɻɐm] – fruit
- [t̪uɳi] – cloth

In this context, *pazham* signifies "old," proving that direct stemming cannot always capture semantic shifts.

Example 9: ചെമ്മാനം (chemmaman) [ʧemːaːn̪ i]

- Incorrect translation: Red prawn
- Correct meaning: Red sky
- Glossing:
- [ʧem] – red
- [mːaːn̪i] – sky

This phrase exemplifies how Malayalam compounds require contextual interpretation.

Example 10: മുച്ചുണ്ട് (muchund) [muˀʧuɳɖi]

- Incorrect translation: There are three
- Correct meaning: Cleft lips
- Glossing:
- [muɾiɲɲɐ] – cut

- [ʧʃuɳɖɨ] – lips

Direct translation does not convey the intended meaning, highlighting the importance of morphological analysis in machine translation.

2 Conclusion

To conclude, the study highlights the challenges Malayalam adjectival compounds pose to traditional stemming techniques in machine translation. It finds that context-dependent meaning shifts, non-literal translations, and noun-based compounds significantly affect translation accuracy, as direct stemming methods fail to capture the semantic variations inherent in Malayalam morphology. The initial research question demonstrates that traditional stemming techniques struggle with adjectival compounds, leading to incorrect word-for-word translations. Regarding the second research question, findings confirm that context-sensitive processing is essential in improving translation accuracy, as words like *pacha* (‘green’) change meaning based on collocation. The study also answers the third research question by proposing the integration of rule-based morphological analysis with deep learning models to enhance machine translation quality. By incorporating syntactic and semantic analysis, translation systems can recognize fixed patterns in compound structures, reducing errors caused by rigid stemming. The study suggests that hybrid translation models leveraging morphological rules and AI-driven contextual learning can effectively preserve intended meanings in Malayalam-English translation. Future research should explore fine-tuned deep learning architectures trained on morphologically rich corpora to refine translation models further. In machine translation of Malayalam compound words, when the compound word possesses a meaning entirely different from the combined meanings of its constituent words, it should be treated as a single stem to prevent inaccuracies during the stemming process. Additionally, training the machine with a large corpus may help it detect the accurate meanings of such compound words in Malayalam and various morphologically similar languages.

3. References

1. Sreelekha, S., & Bhattacharyya, P. (2018). Morphology injection for English-Malayalam statistical machine translation. In Proceedings of LREC 2018. ELRA.
2. Patel, R. N., Pimpale, P. B., & Sasikumar, M. (2017). Machine translation in Indian languages: Challenges and resolution. arXiv preprint arXiv:1708.07950.
3. Nida, E. A. (1949). Morphology: The descriptive analysis of words. University of Michigan Press.

4. Asher, R. E., & Kumari, T. C. (1997). *Malayalam*. Routledge.
5. Nida, E. A. (1976). *Morphology: The descriptive analysis of words*. University of Michigan Press.